UNIVERSITY OF SOUTHAMPTON
Faculty of Engineering and Physical Sciences
School of Electronics and Computer Science

DATA MATTERS

Towards a Data-centric Theory of Generalisation

by Antonia Marcu

A thesis submitted for the degree of Doctor of Philosophy

August 22, 2023

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCES
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Thesis for the degree of Doctor of Philosophy

by Antonia Marcu

The ability of a learning machine to perform outside the training data is referred to as its generalisation performance. Despite being researched for many years, generalisation is one of the key unresolved puzzles in machine learning. In this thesis we start building the *understanding* needed to construct a new framework for reasoning about generalisation. We start with a theoretical perspective but conclude that *the field needs to build stronger intuitions* before being able to formalise generalisation in a meaningful way. Our theoretical exploration, however, highlights that *the data plays a much more central role than previously acknowledged.*

To better understand how the data can be incorporated in generalisation studies, we start exploring the practice of modifying images. The modifications we consider are mixed data augmentation, patch-shuffling, and patch-based occlusion. We find that there are a number of *incorrect implicit assumptions* in the literature regarding the side effects of data modification. These assumptions deem some distortion-based approaches to evaluating model attributes to be incorrect. In the case of modifying data to assess robustness to occlusion, we propose a solution that addresses the side effects.

The existence of these incorrect assumptions attests to the fact that the field has a poor understanding of data modification. Despite the field's limited understanding, data distortion has most recently been used to *empirically predict generalisation performance.* We focus on this practice and claim that data modification has been carelessly used in this case as well. We argue that it is the limited evaluation settings that caused the modification-based predictors to appear successful despite relying on poorly founded intuitions. We end by proposing the backbone for an extensive evaluation of empirical predictors of generalisation. We believe that such a practical approach to generalisation, when thoroughly designed, has the potential to provide the understanding needed to create a theoretical framework in future. Our proposed evaluation setting seeks to explore a variety of data-centric scenarios, highlighting the *central role played by the data in the generalisation puzzle.*

Contents

Declaration of Authorship

I declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- Parts of this work have been published as:

Antonia Marcu and Adam Prugel-Bennett. On the effects of artificial data modification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15050–15069. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/marcu22a.html`.

Antonia Marcu and Adam Prügel-Bennett. On data-centric myths. *Neural Information Processing Systems Workshop on Data-Centric AI*, 2021. URL `https://arxiv.org/abs/2111.11514`.

Antonia Marcu. On pitfalls of measuring occlusion robustness through data distortion. *Interneational Conference on Learnt Representations Workshop on Robust Machine Learning*, 2021. URL `https://arxiv.org/abs/2211.13734`.

Antonia Marcu and Adam Prügel-Bennett. Rethinking generalisation. *Neural Information Processing Systems Workshop on Machine Learning with Guarantees*, 2019. URL `https://arxiv.org/abs/1911.04301`.

Signature ..... Date .....

Foreword

This chapter gives an introduction to the universe of this thesis. It motivates the choice of topic, but also the way the subject is addressed and presented. It gives the perspective from which we encourage the reader to approach the thesis.

After the AI winter, machine learning has been everything from the great revived hope to buzzword generator. So what is machine learning really about? What makes machine learning – "machine learning"? We argue that in one way or another, the root of machine learning applications is learning a function that best captures the given data. This could be either learning to represent the data or simply finding a rule based on which the data can be understood or discriminated. Whether for predictive, generative, adversarial or other purposes, learning such a function of the data is key. In other words, there is no machine learning without learning from the data. Thus, determining the quality of what was learnt is arguably one of the most important endeavours of machine learning. Determining this quality is the subject of the present thesis.

Why is it challenging? When learning a function of the data, we only have access to a limited set of samples. What we are interested in is how suited the function is for describing the samples that are part of the true underlying data distribution yet which we do not have access to. But how can we determine how suitable the function is if we do not have access to those samples or to the true underlying distribution? Although this is a massively oversimplified view, in its simplicity it captures the essence of the problem. All the technical challenges we will introduce throughout this work essentially boil down to this.

How do we measure the function quality? So far we have talked about the quality of a learnt function without explaining how it is measured. This is because the answer is not straightforward and requires more context. For the moment, we will give an inexact response that will be detailed later on. In this thesis we focus on the task of prediction. In this context, we will consider the quality of the learning function to be how well the function can make predictions on unseen parts of the data distribution. This is referred to as the *generalisation performance* and it is the main object of study in this work.

Why is generalisation performance so important? The most recent advances in machine learning are largely characterised by great intuitions and happy accidents. It is very rarely the case that a newly proposed approach is driven by a theoretical finding. Capturing the quality of the learned function would implicitly lead to a solid understanding of the true underlying mechanism. From a practical point of view, capturing generalisation could lead to more principled advances. From a theoretical point of view, it can lead to formalising a relevant framework that can both fuel progress as well as warrant the reliability of current practices. The latter is particularly important since the field is hoping to apply machine learning to sensitive applications that would benefit from some form of theoretical validation in order to be fully trusted.

How can it be approached? From a reductionist perspective, there are two types of approaches to scientific research: the generalist and the specialist approach. As we will recount later in the thesis, the field of generalisation studies has seen over the recent years an increase in outlooks and directions, uncovering the true complexity of the problem. So much so that what we believed would be a specialist task, turned out to be an entire field worthy to be dedicated a life-long career. We see the present thesis as the initial search of a future "generalised specialist"; an expert in generalisation studies that is not committed to one single view of this extremely challenging problem.

This perspective is key to understanding the nature of the content of the thesis: a mixture of high-level discussions of the main directions in generalisation studies and in-depth, methodical analyses of various phenomena. This mixture represents a search for a unified view of generalisation; for understanding the bigger picture and at the same time understanding what are the details that hold it together. We will constantly remind the reader that this thesis is, above everything else, about *understanding*. This is the leitmotif of the work and its driving force.

From a structural point of view, the thesis follows the order in which the subjects have emerged. Our search started from theoretical studies. Slowly, the belief that meaningful theories need to be based on strong empirical intuitions was formed. As a result, we shifted to empirical analyses. We then took the first steps towards bridging theory and practice. Our hope is that the insights we provide in this thesis will one day help the community formulate a framework for reasoning about generalisation that is relevant in practice.

In summary, this thesis reflects, both in its content and structure, an exploration. Our exploration is motivated by the belief that good practical intuitions are required to build a relevant theoretical framework. This is the belief that we hope the reader will share by the end of this work.

Chapter 1

Directions in Generalisation: a Short Introduction

This chapter gives a high-level account of the main research directions in capturing the generalisation ability of learning machines. We refer to such studies as "generalisation studies". We use this chapter to propose an initial classification of the various research directions, which aim to either bound generalisation, or to estimate it. Based on this classification, we define the scope of the thesis and outline some of the main contributions. Lastly, we provide an overview of the following chapters.

In this chapter we introduce one of the main beliefs expressed in this thesis, namely that the machine learning field needs to rethink generalisation. To justify this claim, throughout the thesis we provide an overview of the main directions in generalisation studies. Here we start this overview by proposing a first high-level classification of approaches to reasoning about generalisation. In the Foreword we have informally described generalisation as the quality of the learnt function to perform on unseen data. There are a few important observations to be made.

Performance is Task-dependent

First, it must be noted that here we refer to functions learnt for the purpose of prediction, which is the setting we are concerned with in this thesis. For other applications, our definition might not apply exactly. For example, in the case of generative tasks, the quality of the learnt function is less straightforward to determine. In that case, notions of quality typically refer to how "realistic" and diverse the generated samples are (e.g. Salimans et al., 2016; Bińkowski et al., 2018; Heusel et al., 2017; Kynkäänniemi et al., 2019), but the existing evaluations are biased (Barratt and Sharma, 2018; Kynkäänniemi et al., 2022) and an objective quality is yet to be defined. Thus although the gist of

generalisation is the same across all tasks, it is only for prediction that our specific definition holds.

The Data Can Trick Us

We remind the reader that the greatest challenge is to determine the quality of the learnt function on a distribution that we do not have access to. In practice, we empirically estimate the generalisation performance by evaluating the function on a held-out set, the test set. This is typically selected out of the available data samples. Naturally, this method gives a very crude estimation that can give misleading results, especially if the data acquisition process is not carefully thought out.

One such example of misleading results caused by a fault in the data collection process was exposed by Rueckel et al. (2020). They looked at predicting pneumothorax based on chest scans. An intervention carried out *after* a patient is detected with pneumothorax is the insertion of thoracic tubes. Studying the images used to learn and evaluate functions of the data, Rueckel et al. (2020) noted that an overwhelming majority of the scans classified as pneumothorax-positive were taken from patients with thoracic tubes already inserted. Rueckel et al. then chose a function with good reported performance and investigated its predictions on a new set of radiographs for patients that were pneumothorax positive but who had not yet had thoracic tubes inserted. They found that pneumothorax was predicted only when marks of thoracic tubes were present. Thus, the function with good reported performance could not predict pneumothorax in patients which had not previously been diagnosed. The function had no real predictive power.

In the case presented above, the test performance, which is the empirical *estimate* of the generalisation performance, would not be reflective of the *true* generalisation performance. Oftentimes, such unintentional biases in the data exist. This is the first observation that needs to be considered when we equate model performance with generalisation performance and generalisation performance with performance on test data.

Generalisation Performance Is Not *All* We Care About

As mentioned in the Foreword, the field is changing rapidly, and with it, the notions of function quality. As applications evolved, practitioners started to account for other notions of quality apart from predictions themselves. For example, we can imagine a function that is created to assist medical staff in diagnosing patients. In such a case, we might not only be interested in the predicted disease but also the confidence associated with the prediction. This brings us to our second caveat. Apart from ideal cases where a perfect function exists, the generalisation performance cannot measure all the aspects of function quality that we care about. This is even less so the case in practical settings,

where we can only *estimate* the true generalisation performance. The estimation can be improved in clever ways (e.g. computing an average estimate via cross-validation), but we would still rely on heavy assumptions. Throughout the thesis we will expand on this caveat and its nuances.

Generalisation is Still Very Important but Needs Rethinking

We have pointed out a number of challenges when dealing with generalisation. These are abstracted away in generalisation studies. We will mostly follow prior art in this respect. Although by setting aside such details, the problem of capturing function quality is massively simplified, it still remains one of the most difficult and fundamental problems in machine learning. Despite its paramount importance, and the numerous attempts to understand and capture generalisation, we still lack to this day a comprehensive framework that gives real insight into when and how functions generalise. As we will argue throughout the thesis, we believe that one of the reasons for this is that the field still lacks a strong intuition grounded in well-defined and principled experimentation.

What are the numerous attempts to understand and capture generalisation? To the best of our knowledge, a comprehensive introduction to generalisation studies is missing from the literature. In this thesis we aim to incorporate an overview that highlights the main ideas in the field, while being accessible to a wide audience. The motivation for this is similar to that of Alquier (2021), who proposes an introduction to one of the directions in generalisation, namely PAC-Bayes bounds, which we will introduce in Chapter 4. Alquier argues that given the large corpora of papers on PAC-Bayes bounds, it is difficult for one to be aware of what has been proposed so far. We believe this observation extends to the entire field of generalisation.

Secondly, Alquier points out that most works assume familiarity with the field, making it difficult for non-specialists to engage with the content in the absence of a separate introductory work. This holds true for most papers in the field of generalisation, not only PAC-Bayes ones. For these reasons, we aim to present, classify, and contextualise the main ideas in the field and their evolution. We direct this overview at machine learners, assuming basic prior knowledge. Nonetheless, we will explicitly introduce the core terminology, especially where terms have been used with varied meanings in prior art.

Why does the thesis not subsequently focus on a single approach? This thesis has a broad and ambitious goal. One of the reasons for this, as mentioned in the Foreword, is the accelerated pace at which new directions are being proposed, turning this pursuit from a specialised task to a generalised one. But more importantly, we believe the field is still missing a key ingredient that *none* of the individual approaches addresses. Note that although we do not know what the missing ingredient is, we believe

it requires a holistic understanding. Therefore, we advocate for rethinking generalisation and in this thesis we work towards this. In our overview we highlight the strengths and limitations of the main schools of thought in generalisation and provide our view on what we believe are promising future directions. We hope our work will invite the community to think outside of the confines of established theories in an informed way, so as to find the missing piece of the generalisation puzzle.

What are the established theories? Until recently, generalisation frameworks were highly formal, rigorously calculating results. We refer to these as "classical approaches". In the following chapter we will discuss the irrelevance of classical results to practical settings which caused a change in the community's take on generalisation. As such, more empirical analyses started to be carried out, and most recently a new field has emerged: empirical theory of generalisation. This change in perspective leads to what is, in our view, the main branching in generalisation studies at the time of writing: whether the objective is to bound or estimate the generalisation performance. Below, we briefly introduce these two perspectives.

1.1 Bounding or Estimating Generalisation?

There is no clear answer to the dilemma of bounding versus estimating generalisation. In this section we look at how prior art has approached this question, and we emphasise the strengths and limitations of the chosen perspectives. We start by more precisely defining the terminology specific to generalisation while still maintaining an informal tone. This will allow us to make more in-depth arguments but keep the introduction lightweight.

So far we have talked about learning a function of the data. The term "function" is an improper one and was only used for illustrative purposes. In machine learning, such a "function" is a neural network commonly referred to as a model, and can be informally thought of as a particular instance of an architecture (i.e. a certain configuration of parameters). Note that what we are actually interested in when defining a learner is the functional effect on the representation space. Two different parameter configurations can partition the space in the same way and as a result they would be considered, in effect, the same learner. Therefore the correspondence between model instances and functions is not exactly one-to-one. However, to keep the introduction simple, we will use this analogy for the moment.

The generalisation performance is usually measured through the generalisation *error*, which can be informally thought of as the probability that the model will output an incorrect prediction. Until recently, the literature was mainly dedicated to *bounding* the generalisation error. The bounding approach is able to provide guarantees but has the downside of being disconnected from practical results. For this reason, increasing

efforts are being dedicated to empirically *estimating* generalisation. However, the most effective measures lack a good theoretical foundation and are still far from capturing the true mechanism. To build a relevant framework for reasoning about this important problem we believe the field needs to bridge the two approaches.

More precisely, we believe the ultimate goal of generalisation studies should be to provide *guarantees* that are relevant in *practical* settings. However, this has proved to be difficult to achieve with the field's current level of understanding of the true mechanism. Our take is that estimating generalisation could be a more approachable target. The hope is that by iteratively refining our estimators we could gain an intuition of what the mechanism behind generalisation is. Once such an intuition is constructed, we could compute more informative bounds.

The purpose of this thesis is to further the understanding of the community rather than solve this complex problem altogether. To set out the scene, this section briefly introduces the bounding and estimation approaches, with more detailed descriptions in Chapter 2 and Chapter 4 respectively. The bounding approaches started off as highly theoretical but recently incorporated empirical elements. On the other hand, estimators are entirely empirical. Naturally, the difference between theoretical and empirical directions is given by the number of elements the results abstract away from. Although, in essence, the empirical bounds bridge theoretical results with practical quantities, we argue that they are still uninformative because they typically do not pay sufficient attention to a core ingredient – the data. Thus, before presenting the *bounding* and *estimating* directions, we quickly present the main factors to be considered when creating a theory of generalisation. These elementary notions of generalisation studies will later help us differentiate between the theoretical and empirical approaches to bounding.

Theoretical versus Empirical Perspective – Elements to Consider

Typically, there are two approaches to reasoning about generalisation based on the object of interest, which can be either the choice of model class or the quality of the model instance. Based on these two de facto components for describing generalisation, we will next define what is known in the field as the generalisation error decomposition. With the emergence of studies that, like ours, advocate for the importance of accounting for the data, a number of works have started to consider the data as one of the core elements of study. Thus, we will differentiate prior art based on how they relate to these three elements.

In addition to the existing terminology in the literature, we propose a number of terms to more easily discriminate between methods throughout the thesis. Note that the exact view on the concepts we present varies between theories. We often adopt the deep

learning terminology but strive for generality such that consistency with the classical view is maintained. Where this is not possible, we spell out the differences.

- **Model Class Choice.** When faced with a learning problem, one must choose a model class. An instance from this class will be picked as a result of the learning process. In the more familiar terms of deep learning, the model class can be thought of as the choice of architecture, while the particular "learner", or instance, is given by a fixed choice of model parameters[1].

  For choosing the model class, we use prior knowledge about what structural biases exist in the data and the type of architecture that is fit for addressing them. Traditionally, the model class' ability to capture the data is reflected in the generalisation performance of the best instances of the chosen class. Note that in practice we cannot know this quantity since the model space is dauntingly vast. Instead, it is a theoretical quantity termed *approximation error* that helps us reason about generalisation.

- **Model Instance Quality.** Another source of error stems from choosing a suboptimal model instance. In this setting, we are interested in the model instance quality relative to the optimal model instance. This gives the *estimation error*. Together with the approximation error, it encompasses the generalisation error.

  In practice there are a number of factors at play, such as the training procedure or the optimisation process, which is itself a product of initialisation methods, hyperparameter choices, optimisation algorithm, etc. From a classical theory point of view, however, the learning process is abstracted away. This error is seen as stemming from the size of the training set and the *complexity* of the class. In some textbooks the trade-off between the approximation and estimation error appears as the "bias-complexity trade-off" (e.g. Shalev-Shwartz and Ben-David, 2014).

  Capacity and complexity are two terms associated with the model class. These terms are sometimes used interchangeably in the literature and refer to the level of *expressivity* of the model *class*. The term complexity is sometimes also used in relation to the model *instance*, creating much confusion in the field. In this chapter we focus on the expressivity of the model class alone. In the language of classical theory, the expressivity of a class can be thought of as the total number of *functionally different* learners belonging to that class. We will expand on this notion as well as the instance-centric definition of complexity in Chapters 2 and 4.

  For completeness, we must mention that there exists an alternative view of the generalisation error named the bias-variance decomposition. This view is concerned with the *expected* generalisation performance. In essence, it captures the same phenomenon as the approximation–estimation decomposition with a similar emphasis on complexity.

[1] Once again, the mapping of terminologies is inexact due to network symmetries.

When computing the expected generalisation performance, we take the expectation over the training set, which is sampled from the true data distribution. Note that although this is a theoretical setting, it can be estimated in practice by sampling different partitions of the test set. Instead of considering a single model instance, we pick a model instance for each sampled training set and average their predictions. As such, we obtain the *mean learner*. Akin to the estimation error, we take the difference between the best model instance and the learnt model, which in this case will be the mean learner. This gives the *bias* of the model class.

The other term in this alternative decomposition is given by the *variance* in the learned model instances. Once again, it is argued that too high a model class complexity *could* lead to overfitting to each specific training set, thus resulting in a high variance. Conversely, a too-low model complexity would lead to underfitting, which implies that the average learner would be far from the best model instance, hence leading to a large bias component. Both decompositions of the generalisation error are sometimes referred to as the complexity trade-off, reflecting the belief that generalisation is controlled by the class complexity.

- **Data distribution.** Generalisation studies can also be differentiated based on how they treat the data. Most generalisation theories assume train and test data are both drawn from the true data distribution. In the literature, this is referred to in short as "the i.i.d. (independent and identically distributed) assumption". From this assumption, two directions emerge. On the one hand we have the classical theories and those stemming from it, which do not take the data into account. We refer to these as *data-agnostic*. On the other hand, we have those which take the data into account. Some do so indirectly and aim to estimate performance before seeing the test data. In this thesis we refer to them as *a priori* methods. A different line of work argues that in practice the test data distribution can differ from training and for this reason it studies generalisation *a posteriori*, i.e. at inference time, after the test data is given. Note that *a priori* and *a posteriori* as defined here are not to be mistaken with the Bayesian terminology. We simply start from the larger philosophical notions of a priori and a posteriori and choose the test data as the central point of reference ("that which is experienced").

A Note on Statistical Learning Theory

Using the notions mentioned above, we can now differentiate between theoretically and empirically bounding generalisation. To the best of our knowledge, there is no clear classification of generalisation studies. The first seminal attempts to formalise generalisation were those of statistical learning theory. The field has much evolved since the first results of this framework were presented, with so many iterative changes that we would argue it is difficult to decide what bounding methods can be considered part of

statistical learning theory and which ones cannot. Based on the terms introduced earlier, we will refer to the original setting of statistical learning as "classical". The classical framework, which we will introduce in Chapter 2, completely abstracts away from the data and is concerned with the model class.

Later statistical approaches started incorporating details about the data distribution in various ways but remained focused on the model class, while others have resorted to using complexity notions that are tied to the chosen instance rather that the whole class. Therefore, we will refer to classical bounds as theoretical, while the newer approaches will be referred to as empirical. We detail these approaches, their settings and assumptions below. This breakdown not only has the role of fitting each approach in a particular category but, more importantly, helps us clearly define the scope of each method. Many of the previous directions have clear strengths within the settings they address. We aim to emphasise them and motivate why each particular perspective was chosen. We then discuss their limitations considering the confines of their studies. Note once again that in this discussion we do not refer to the specific directions belonging to each category. Instead, we do so in the dedicated sections of Chapters 2 and 4.

1.1.1 Theoretically Bounding the Generalisation Error

As Dziugaite et al. (2020) noted, within the generalisation approaches two polar opposite directions can be distinguished: one that entirely abstracts away from all the aspects of the problem and learning procedure and one which depends on every detail. The first category is largely represented by *classical* statistical learning, arguably the most prominent theoretical direction for decades. It provides a data-agnostic view on generalisation that is focused on the model space. As such, no particular model is chosen. Results can instead be considered to reflect the *worst-case* scenario given a selected subset of learners. We begin by informally introducing the statistical learning framework on top of which we build the theoretical part of our work in Chapter 2.

In statistical learning we start from the i.i.d. assumption. In the language of statistical learning, the function or rule we are trying to learn is referred to as a *hypothesis*. Thus, the purpose of the learning algorithm is to choose, out of a set of hypotheses, one that is the most representative of the true data. The theory employs information about this set of hypotheses, also referred to as *the hypothesis class* to guarantee that a good enough hypothesis will be selected with a certain probability.

More specifically, the results of classical statistical learning are in the form of bounds which depend on some notion of hypothesis class complexity or capacity. The lower the capacity, the better the *guaranteed* generalisation performance. This type of result formed the basis of a widely spread belief among practitioners that in order to get models with good generalisation, one must reduce capacity. In Chapter 2 we show

that the capacity-based calculations are unable to provide meaningful results because abstracting away from all details of the data and learning procedure leads the bounds to become vacuous when applied to real-world problems.

In short, theoretical bounds are exact in the settings they consider and have great generality. However, the bounds are only of theoretical interest, with no practical relevance. As we will recount in Chapter 2, this is because modern learners have good generalisation performance despite having huge class capacity.

1.1.2 Empirically Bounding the Generalisation Error

The main factor that spurred researchers to move away from the classical framework was the remark that the worst case, which classical theory aims to account for, is very far from the typically observed case. Thus, to address the vacuousness of theoretical bounds, a few directions in statistical learning aimed to incorporate information about the data. Going further, most recently researchers started computing generalisation bounds based on empirically computed quantities . All of these quantities relate more or less directly to notions of complexity. Note that theoretical bounds were also considering complexity. While the notion of "capacity" is automatically associated with the model class, "complexity" has seen a number of different definitions, some relating to the model class and others to the model instance. Until we discuss the technical details in the main body of the thesis, all these terms could be simply thought of as measures of expressive power.

Empirical bounds are highly valued for their rigorously proved results but as we will discuss in Chapter 4, they are still far from helping us capture the generalisation mechanism. Subsequently, empirical evaluation of complexity measures inspired researchers to explore generalisation prediction outside the confines of bounds.

1.1.3 Empirically Estimating Generalisation

Despite the fact that the attention has shifted from bound-like results, the paradigm has long remained that it is through some notion of model *complexity* that one can understand generalisation. However, most recently researchers have started to explore other indicators of good generalisation. Although not always explicitly stated, we argue many of the methods that have the learnt instance at heart can be fundamentally seen as combining representational geometry and some notion of model robustness. An increasingly adopted way of measuring model robustness is through data modification with a particular focus on data augmentation. However, we argue in this thesis that the effects of data modifications on their own are poorly understood. This has led to a rather ad-hoc usage of data modification for the purpose of predicting generalisation.

Empirical predictors could represent a promising way of acquiring the intuition needed to capture the true mechanism of generalisation. However, this direction is in a very incipient stage and at the moment is very far from the rigour of bounding approaches.

Summary

To summarise this discussion so far, researchers have acquired partial intuitions about how machine learning works but have no formal frameworks that allow them to capture and refine their understanding. Conversely, the hard proofs of the theorists are exact in the settings they consider but of no real practical relevance. We believe that addressing this problem from both perspectives is necessary for reaching a *comprehensive understanding that can be formalised*. We believe that there is a need for a better formulation of intuitions and for the construction of a theoretical framework where these could be clearly seen. At the same time, we believe the field should build stronger intuitions based on rigorous empirical findings which can then be used for building relevant frameworks.

1.2 Thesis Overview and Scope

In this thesis we want to start empirically building the grounds and motivating the construction of new, more flexible theoretical formulations. We begin with the classical statistical learning framework and underline its limitations. We propose a new model of generalisation which encourages the reader to see this problem from a *data-centric* perspective.

Having this view in mind, we pursue an empirical understanding of generalisation that is more focused on the data. We limit our exploration to visual classification tasks and propose an incremental understanding by analysing changes that occur when modifying data. More precisely, we evaluate models that were trained on different data augmentations but also evaluate model attributes using data alterations. We believe incorporating such modifications can fuel a data-aware direction for generalisation studies and help uncover a yet unexplored piece of the puzzle. Incidentally, this endeavour highlights a number of incorrect assumptions about the side effects of data manipulations that are commonly adopted in the field. For example, we analyse grid-shuffling for shape-texture bias identification and patch overlapping for occlusion robustness and find that they represent biased methods of evaluating model attributes.

Lastly, we focus on the newest direction in generalisation studies represented by the empirical estimation of model performance. We first provide a comprehensive review of previous approaches. We then start building the foundations of a large-scale study for evaluating such estimators. We aim to build on previous large-scale studies and address

scenarios which had previously been omitted. More specifically, we advocate for the necessity of incorporating variations in the data in the evaluation procedures.

With a few exceptions (e.g. Kaplan et al., 2020), most of the influential experiments and studies of generalisation in deep learning have so far been centred around vision applications. Because of this, as well as stronger intuitions built on personal experience, we exclusively focus on image tasks in the empirical part of the thesis.

Note that although we take a data-centric approach, the generalisation ability as discussed in our empirical study excludes extrapolation. That is, we leave out the settings in which the true distribution used to sample the training data is significantly different to that of the test data. There is no clear threshold starting from which two distributions are different enough for the problem to fall under the "extrapolation" or "out-of-distribution" regime. This issue brings a level of subjectivity to generalisation evaluations and our work is no exception. However, finding such a threshold is a highly complex problem in and of itself and would make up the subject of a whole new study.

Distinguishing between out-of-distribution and within-distribution problems is only one of the challenges associated with studying generalisation. There are many more important aspects which we do not discuss in the present thesis such as how easy it is to navigate the loss landscape for finding a low-error solution, or properties associated with each architecture. We reflect on some of these in Chapter 5, Closing Remarks and Future Directions, and propose ways of integrating them into future studies.

Coming back to the pneumothorax example (Rueckel et al., 2020) we introduced in the beginning of this chapter, a part of the generalisation puzzle is the "quality" of the data itself. Although highly connected to the performance of the learnt model, this constitutes an independent area of research and will not be addressed in this thesis. Challenges in this field range from identifying mislabelled samples to determining if there is sufficient variety in the available data. A related issue is that of data complexity, which we discuss in the Future Work section of Chapter 3.

We would like to end this section with a note on classifying generalisation studies. Earlier in the process of writing this thesis, we created a classification of generalisation studies that was significantly simpler. In a short time, the classification criteria no longer held, as directions became more and more nuanced. First and foremost, we find this encouraging. The field is still swarming with ideas and great collective efforts are being made to solve this problem. Secondly, this rapid change likely tells us that our most updated classification will quickly not be able to account for advances in the field. Thus, our objective is not to provide a final classification of directions in the field. Instead, we aim to provide a clear picture of what has already been considered and how the ideas in the field have evolved up to the moment of writing. Our hope is that this will help researchers identify paths that have not yet been explored.

1.3 Structure and Contributions

With a clearer picture of the scope of the thesis, we proceed to outline the structure and our main contributions. The chapters are written around the theoretical–empirical trade-off. Chapter 2 is highly theoretical, with rigorous calculations, while Chapter 3 is highly empirical, with thoroughly thought-out experiments. Each of these two chapters constituted the backbone of independent research papers. These complementary approaches are then followed by Chapter 4, which focuses on empirical predictors of generalisation and their evaluation. Finally, Chapter 5 proposes broader future directions for the overall goal of understanding and capturing generalisation.

Aspects of the work in this thesis were published and presented at the 2022 International Conference on Machine Learning (ICML), the 2019 Neural Information Processing Systems (NeurIPS) Workshop on Machine Learning with Guarantees, the 2021 Neural Information Processing Systems (NeurIPS) Workshop on Data-centric AI, and 2021 the International Conference on Learnt Representations (ICLR) Workshop on Robust Machine Learning. The calculations in Chapter 2 were presented at the Applied Maths Seminar at University of Southampton, while the work as a whole was presented as an invited talk at the Max Plank Institute for Intelligent Systems, Tübingen. At the beginning of each chapter we highlight the venues where its contents were presented, as well as the list of contributions corresponding to that particular chapter.

The narrative is often question-based. To make the story easier to follow, we also provide a short abstract at the beginning of each chapter. Chapters 2, 3, and 4 have a "Future Work" section with concrete directions and experiment proposals for extending the work presented in each of them. A more high-level vision for the future of this work is presented in Chapter 5, Closing Remarks and Future Directions. Below, we give a chapter-based overview of the main ideas and contributions.

Chapter 2: The Theoretical Approach: the Importance of the Data

This chapter is concerned with the statistical learning view of generalisation. We start by introducing the terminology and set-up of classical statistical learning and highlight the aspects in which we deviate from it. Unlike most studies, we then compute the *expected* generalisation performance. The main contribution of our work is to propose a more accessible way to reason about the generalisation phenomena. Our calculation provides a higher-level perspective, highlighting the importance of the *alignment* between the choice of model class and the data at hand. We term this model-problem alignment "attunement" and argue that capturing it is crucial for reasoning about generalisation in a meaningful way.

Chapter 3: Steps Towards the Empirical Approach: Understanding by Distorting

The previous chapter raises the need for building the necessary intuition to identify and control attunement. In Chapter 3 we take initial steps towards this, by informally investigating what aspects of the data are important for generalisation. Inspired by recent empirical generalisation predictors, we start our exploration by looking at the practice of modifying data. We combat a number of incorrect implicit assumptions in the field and advocate for more awareness of side effects when dealing with data modification. To do so, we create an index to identify instances of unfair evaluation and we then propose an alternative method for assessing robustness to occlusion. We subsequently use our newly proposed method to formulate a number of questions about attributes previously associated with generalisation and propose directions for future exploration of generalisation from an empirical viewpoint.

Chapter 4: Steps Towards a Data-centric Evaluation of Empirical Predictors

Working towards empirically capturing generalisation, this chapter establishes the foundations for principled evaluation of generalisation predictors. We start by reviewing prior art, presenting a comprehensive overview of directions in generalisation. This provides the context in which empirical predictors have emerged, and allows us to highlight the strengths and limitations imposed by studying them. We then propose a standardised way of evaluating these predictors. Building on prior evaluations, we propose a list of requirements and settings that we believe must be addressed. Lastly, we present the concepts that we believe can be used to capture generalisation. We argue that it is in terms of these concepts that new, meaningful predictors could be designed.

Chapter 5: Closing Remarks and Future Directions

In this final chapter of the thesis we return to a more holistic view of generalisation studies. We once again emphasise that understanding machine learning is a career-long pursuit and we reiterate our contributions from this perspective. More specifically, the present work has a foundational role, although an incomplete one. We reflect on the limitations of the thesis. For some of these limitations we propose ways of addressing them in the future, while for others we pose open-ended questions that we hope the wider community will be inspired to help answer. We end with optimism and great enthusiasm for possible future directions.

Supplementary Material

To keep the thesis concise, we provide in this part additional evidence that further corroborates results found in the main body. For example, in Chapter 3 we show that the side effects of modifying data affect evaluation for one scenario; in the Supplementary Material we provide the results for the remainder of the scenarios we analyse, which reinforce the conclusions drawn for the scenario presented in the main body of the thesis. This is to simply show that we did not consider a single case, but instead a more thorough analysis was carried out. We also provide in the Supplementary Material the full experimental details and a description of all the data sets we use. Note that the thesis is stand-alone and this part only addresses additional supporting evidence.

In summary, the contributions of this thesis are:

Chapter 2 - **We show the importance of the data.**

- We propose the β-Risk model for calculating the expected risk for classification under the annealed approximation (Section 2.2.1);
- Using the β-Risk model, we show the importance of accounting for the data when studying generalisation (Section 2.2.1);
- We validate our model in the case of the perceptron (Section 2.3.1);
- We propose a number of avenues for expanding our calculations and their interpretation (Section 2.4).

Chapter 3 - **We call for a more principled use of data distortion.**

- We identify a neglected phenomenon in data distortion which we term "data interference" (Section 3.2);
- We propose an index that shows the presence of data interference (Section 3.2);
- We empirically show that model evaluation methods that do not account for data interference provide biased results (Section 3.2.1);
- We identify three limitations of the classical method of measuring robustness to occlusion (Section 3.3.2);
- We propose a fairer alternative for measuring robustness to occlusion that addresses the identified limitations (Section 3.3);
- We disprove the belief that data augmentations that preserve the distribution are better (Section 3.4);
- We raise a number of open-ended questions to guide future research on data modification (Section 3.6).

Chapter 4 - **We design a more extensive evaluation of empirical generalisation predictors.**

- We contextualise and give an account of the major directions in generalisation studies (Section 4.1.1);
- We provide what is, to the best of our knowledge, the first review of modern empirical estimators of generalisation (Section 4.1.1);
- We identify limitations of prior empirical predictors and then build conceptual arguments as well as preliminary experiments to support our claims (Sections 4.1.4 and 4.2.2);
- We review the evaluation settings for empirical predictors proposed in prior art (Section 4.2.1);
- We identify a number of limitations in evaluation studies (Section 4.2.1);
- We design a new data set for evaluation which addresses the limitations we pointed out (Section 4.2.2);
- We identify promising directions and propose a number of actionable approaches for future work (Section 4.3).

Chapter 5 - **We present our vision for the future of generalisation studies.**

- Based on the understanding built throughout the thesis, we propose a number of directions for the future of generalisation studies (Section 5.1).

Chapter 2

The Theoretical Approach: the Importance of the Data

This chapter covers a new formalism for calculating generalisation. We start with a short introduction to the framework of classical statistical learning theory. Highlighting the limitations of the classical approach, we then motivate and introduce the β-Risk model for calculating the expected risk. Our calculations point towards the importance of a data-centric approach to understanding generalisation.

The objective of the thesis is to better understand generalisation. To do so, we start with classical statistical learning theory, which provides the first attempt to formalise generalisation. Since this classical work has been extensively studied, we do not provide a thorough discussion. Instead, we introduce the framework and direct the reader to established textbooks (e.g. Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014) for a more in-depth overview. In this thesis we choose to recount the evolution of ideas in generalisation studies and focus our review on the *new directions*. We provide this overview and discuss the new directions of generalisation studies in Chapter 4.

In this chapter we propose an alternative formalism to that of classical statistical learning theory. Our framework exposes the inability of the classical theory to meaningfully capture generalisation. We argue that this is caused by not accounting for the data. We show that the generalisation performance is determined by the distribution of risks, which is intricately linked to fitness of the model class to the learning problem, and therefore, to the data. Instead, as we will see both in this chapter and in Chapter 4, the focus of most prior studies is on the class or model expressivity. Therefore, the core message of this chapter is that *the data plays a much more central role in generalisation than it was previously attributed.*

In Chapter 1 we have seen that typically the objective of generalisation studies is either to bound or to estimate generalisation performance. Classical statistical learning opts for

an a priori bounding approach. For a priori methods, the problem of training samples' representativeness is implicitly bypassed. The question becomes one of sample quantity rather than quality. In other words, knowing that the train and test samples are i.i.d. then, given sufficiently many training examples, we can get a good approximation of the underlying distribution. In such a context we want to determine what is referred to in the field as the *sample complexity*. The sample complexity is the minimum required size of a randomly chosen training set that would, with a certain probability, lead to a sufficiently good generalisation performance. Here sufficiently good means within a defined interval of the optimal performance. In other words, in this scenario we want to determine the training set size that would guarantee the learnt hypothesis is *approximately correct* with a certain *probability*.

Such results belong to the Probably Approximately Correct (PAC) learning paradigm (Valiant, 1984*a*) and represent the standard results in the classical theoretical framework. Since statistical learning aims to provide guarantees, then the sample complexity needs to hold for all hypotheses in the class and over all possible data distributions. This generality is the undisputed virtue of classical results. However, as we will recount throughout the chapter, this generality makes the classical results uninformative for practical settings. For this reason, like most of the concurrent approaches, we opt to trade generality for practical relevance. Thus, in this chapter, rather than seeking to determine bounds by eliminating all high-risk hypotheses, we aim to study the *expected* generalisation performance. This perspective allows us to question the role of model class expressivity, which is regarded by statistical learning theory as key to understanding generalisation.

Reiterating, the quantity of interest within the classical framework is the sample complexity. In an infinite hypothesis space this can be determined (Shalev-Shwartz and Ben-David, 2014) by the *Vapnik–Chervonenkis* (VC) dimension (Vapnik and Chervonenkis, 1971). The VC dimension captures *capacity*, a notion that is central to the classical theory. As mentioned in Chapter 1, capacity reflects the expressivity of the hypothesis class. The aim of this chapter is to dispute the central role of capacity in understanding generalisation and to propose a framework that could help researchers gain more insights into the generalisation puzzle. The fundamental conclusion of our calculations is that researchers need to account for the data in order to understand generalisation. This conclusion sets the direction for the rest of the thesis.

This chapter was presented at the Applied Maths Seminar of the Mathematical Sciences department at the University of Southampton. A short version of this work was also presented at the NeurIPS 2019 workshop on Machine Learning with Guarantees. Our contributions in this chapter are:

- We propose a framework for reasoning about generalisation where we look at the expected risk rather than eliminating all bad hypotheses (Section 2.2);

- Making assumptions about the distribution of risks, we propose the β-Risk model for classification (Section 2.2.1)
- We validate the β-Risk model on the realisable perceptron (Section 2.3.1), whose distributions of risks can be calculated;
- Through our framework we emphasise the importance of accounting for the data in generalisation studies.

We provide full derivations of our calculations in Sections 2.2 and 2.3. These use a number of known mathematical tools, such as the Beta and Gamma functions, the Gaussian distribution, Jensen's inequality, etc. For completeness, we include the definitions of these in Section A of the Supplementary Material for the reader's reference.

We start by giving a short informal introduction to the statistical learning paradigm, whose framework forms the basis of our calculations. This allows us to underline the limitations of classical work, which we then seek to address.

2.1 Introduction

In the previous chapter we mentioned that the purpose of the learning algorithm is to choose, out of a *set* of hypotheses, the one that best captures the data. **How do we define this ability to capture the data and which set do we pick a hypothesis from?** We start by choosing a loss function to be minimised. For each hypothesis, the loss defines the ability of the hypothesis to make a prediction about a particular data point. Keeping in mind the assumption on which a priori methods are based, our best attempt to capture the unseen data is to select a hypothesis only out of those that do best on the training data. As such, in the statistical learning framework one computes the loss over the whole training data set for all hypotheses in the class and then chooses a subset composed of the hypotheses with the lowest expected loss. This is also referred to as the *risk*. It is the quality of this subset that makes the subject of classical theoretical studies. Note that this process is usually infeasible in practice, as hypothesis spaces are so vast that we cannot exhaustively explore them; it is simply another abstraction that allows us to reason about generalisation.

Naturally, the larger the training set, the better the estimate of the true data distribution. We can say that a new training sample "eliminates" those hypotheses that do not correctly classify it, iteratively getting a more refined hypothesis subset. As mentioned in Chapter 1, traditional statistical learning theory takes a worst-case approach to generalisation analysis. In this view, the number of training samples should be high enough so that all rules that poorly explain the data are eliminated. Once again, this process relies on the idea that rules with poor generalisation performance (high risk)

will, with high probability, make errors on a sufficiently large randomly chosen training data set (Vapnik and Chervonenkis, 1971; Valiant, 1984*a*; Baum and Haussler, 1989; Blumer et al., 1989; Haussler, 1992; Vapnik, 1992). Thus, given a large enough training set, the only rules left are those with good generalisation behaviour. To formalise this, we assume there is a set, $\mathcal{H}$, of hypotheses. Note once again that we also refer to a hypothesis as an individual *rule* or *learning machine*.

The expected loss for a hypothesis, h, over this distribution of data we term *the risk*, R_h. Thus, our objective is to choose a hypothesis with low risk. For a given training set we can compute the total loss, L_h, on all data points for a particular hypothesis h. The loss associated with the training set is known as the *empirical risk*. We assume that we have an algorithm capable of choosing a hypothesis from the set

$$\mathcal{H}_{\text{ERM}} = \{h \in \mathcal{H} | \forall_{h'} \; L_h \leq L_{h'}\} \; , \tag{2.1}$$

i.e. the set of hypotheses with minimum loss on the training set. This is known as *empirical risk minimisation* (ERM). Using the notions introduced above, in traditional statistical learning theory the aim is to find a worst-case bound on the sample complexity such that with overwhelming probability all $h \in \mathcal{H}_{\text{ERM}}$ will have a risk less than some ϵ. To obtain such a bound there needs to be a finite number of hypotheses. Otherwise, there could still be a high-risk hypothesis that by chance did well on the particular training set we used. In the case where the learning machine has a continuous parameter space (so that the dimensionality of the space is uncountably infinite), we consider the effective size of the hypothesis space to be the VC dimension, in terms of which the sample complexity can be expressed.

The VC dimension is a measure of hypothesis class capacity and revolves around the idea of space shattering. Informally, in a classification setting a hypothesis class can shatter a space if, given a set of points, for any possible labelling there exists a hypothesis in the class that can assign the labels correctly. The VC dimension is given by the maximum size of the set that can be shattered by the hypothesis class. More specifically, for a binary classification problem, a hypothesis class shatters a set of size n when one could arrange the elements in the set in such a way that there exists a hypothesis for all 2^n possible combinations of label assignment.

To more easily visualise set shattering, we give a concrete example. We choose the hypothesis class of axis-aligned rectangles, where everything within the rectangle is classified as belonging to class “A”, while everything outside of it belongs to class “B”. Given a set of four points, one can arrange them as depicted on the left-hand side of Figure 2.1. For any possible labelling, we can find an axis-aligned rectangle that gives the desired classification. Therefore, the class of axis-aligned rectangles can shatter a set of size 4. When adding one more point to the set, no matter how we arrange the five points, there will be a configuration that cannot be correctly classified. We can conclude

FIGURE 2.1: The class of axis-aligned rectangle classifiers has a VC dimension of four. We take the axis-aligned rectangles to classify everything with the rectangles as belonging to class "A", and everything outside to belong to class "B". This is because arranging four points as shown on the left-hand side, one can find a rectangle for each possible class assignment. However, adding a fifth point makes this impossible. For example, on the right-hand side configuration, there is no axis-aligned rectangle that can classify the two red points as belonging to class "A" and the three blue points to class "B". Figure inspired by Shalev-Shwartz and Ben-David (2014).

that the VC dimension is 4, the size of the largest shatterable set. This effective size or capacity lies at the heart of conventional statistical learning theory. By limiting the capacity we can obtain stronger bounds on the generalisation performance. But this central role of capacity has been empirically questioned.

Our work, like many of the generalisation studies proposed over the past five years has as its starting point the influential paper of Zhang et al. (2017). In their study, Zhang et al. perform a randomising label experiment which shows that architectures which were achieving state-of-the-art generalisation performance (Szegedy et al., 2016; Krizhevsky et al., 2012), *had immense capacity*. We will introduce the experiment later in this chapter but focus on their observation for the moment.

Why is this observation important? Firstly, the results under the form of bounds had started to be wrongly interpreted in the machine learning folklore. As such, the bounds were taken to mean that it is necessary for the capacity to be restricted in order to achieve generalisation. The finding that good hypothesis classes had almost infinite capacity was seen by some as puzzling and has resulted in many studies aimed at explaining generalisation in overparametrised regimes and attempts to redefine complexity. Secondly, the finding caused the field to question the relevance of theoretical results for modern deep learning settings. The bounds are constructed such that they hold regardless of the data distribution. However, the elegant generality makes them intangible when applied in practice.

In this work we challenge the traditional approach. Eliminating all high-risk hypotheses is, in our view, too stringent and often leads to weak bounds. Good generalisation can be achieved with high probability so long as the vast majority of hypotheses in $\mathcal{H}_{\text{ERM}}$ have low risk. Thus, provided there is no bias towards choosing high-risk machines we will still, with high probability, choose a low-risk machine. In this scenario, capacity plays a much more minor role. Instead, we need to know the distribution of risks, $\rho(r)$, of a learning scenario. That is, we need to know the proportion of hypotheses with a certain risk. As we will show, the asymptotic generalisation performance is determined by the power-law growth in $\rho(r)$ for small r; a quantity we term *attunement*. This new

perspective solves the apparent paradox that arose as a result of Zhang et al. (2017)'s experiment.

The basic idea of our approach is simple. We assume that we are given a set of hypotheses with a given distribution of risks, $\rho(r)$. We eliminate hypotheses that perform poorly on the training examples. This will, with overwhelming probability, remove more hypotheses with high risk, thus the expected risk of hypotheses in $\mathcal{H}_{\text{ERM}}$ will decrease as the size of the training set is increased. Although the idea is simple, computing the expected ERM risk *exactly* from $\rho(r)$ alone cannot be done. We would require information about the correlation between hypotheses, which depends on other details of our learning algorithm. However, by making an assumption about the independence of losses for the hypotheses, we can obtain an approximation to the expected ERM risk from $\rho(r)$ alone. We do this in Section 2, where we propose the β-Risk model for classification. Section 2.3.1 derives an exact expression for the expected risk of a realisable perceptron — this result is data set dependent. This allows us to show that our proposed β-Risk model can closely capture the expected risk of a perceptron.

How does our calculation relate to prior art? One of the main changes to the classical framework of statistical learning that we make is to focus on the expected generalisation performance. As we will present in more detail in Chapter 4, this perspective can also be found in the PAC-Bayesian literature. Our calculations however, are more closely related to the largely forgotten statistical mechanics work on learning (Engel and den Broeck, 2001). This developed out of Gardner (1988)'s calculations of the expected generalisation performance for a perceptron. Gardner's results are believed to be exact in the limit when the number of features becomes infinite. Although we are forced to use approximations, our intent is to develop a more general framework. The approximation developed in the next section is equivalent to the *annealed approximation* in statistical mechanics. The work presented in this chapter was done independently and we only became aware of the field after drafting the calculations. Although, as mentioned in the introduction of this chapter, statistical learning literature is extensively covered by prior studies and textbooks, the connections between statistical learning theory and the work done in the physics field of statistical mechanics are rarely mentioned in the generalisation literature. For a review of the work on learning in statistical mechanics, we refer the reader to Engel and den Broeck (2001).

Another similar approach to ours has also been put forward by Scheffer and Joachims (1999). For a number of classes of hypotheses, they estimate empirically the distribution of error rates from which the expected error of an ERM hypothesis from each hypothesis class can be obtained. However, they used this approach to propose a model selection algorithm rather than introducing a new framework for reasoning about generalisation.

What is the contribution of our calculation? The novelty of our work stems from its focus on *simplicity* and the clear connection with statistical learning theory. Under

the annealed approximation, we assume a distribution for the risks which then allows us to study the relationship between this distribution and the generalisation performance. This allowed us to grasp the importance of accounting for the data, which leads us to pursue a data-centric analysis of generalisation in the upcoming chapters. Therefore, the β-Risk model we propose allows us to gain a higher-level understanding of generalisation compared to prior art. Thus, the present chapter focuses on carrying out calculations that can provide a better intuition on the quantities of interest for generalisation studies.

Before presenting our calculations, we introduce Zhang et al. (2017)'s label randomisation experiment in more detail and relate it to the view on generalisation that we discuss in this chapter.

The Label Randomisation Experiment

In our opinion, the label randomisation experiment has fuelled the development of modern generalisation theories. Given its seminal role, we will refer back to this experiment throughout the thesis. Zhang et al. (2017) challenged the relevance of classical theories by studying some of the most successful deep learning architectures at the time: AlexNet (Krizhevsky et al., 2012) and Inception (Szegedy et al., 2015). They used CIFAR-10 (Krizhevsky et al., 2009), a 10-way image classification task consisting of 50 000 training images, and ImageNet (Deng et al., 2009), a 1000-way classification task with over one million training images. The CIFAR-10 images are of size 32×32, while ImageNet ones are centre-cropped to 224×224 pixels. They then trained AlexNet and Inception models on modified versions of these data sets, where the images were assigned random labels.

Interestingly, despite this unnatural setting, they were still able to find a set of parameters that for CIFAR-10 (Krizhevsky et al., 2009) perfectly classified all the training examples, while for ImageNet (Russakovsky et al., 2015) they found network instances with very low errors on the randomly labeled training data. This experiment shows that the capacity of the AlexNet and Inception architectures is incredibly vast; so vast that they are able to shatter spaces of dimensionality at least as high as those of the considered data sets. In fact, these networks have such a large capacity that conventional statistical learning theory can provide no useful guarantee of generalisation performance. Notably, the AlexNet and Inception networks consisted of many fewer parameters than some of the successful architectures which have been proposed since. Thus, for modern deep learning architectures, the bounds of classical theory become vacuous. Nevertheless, when trained on the real, unmodified data, such vast-capacity architectures achieve state-of-the-art results. In our approach described below, this provides no contradiction. If we consider the set of parameters that perform well on the training set, then an overwhelming proportion of those parameters corresponds to low-risk hypotheses.

Although in the thesis we focus on the VC dimension, we must mention that this is not the only measure of expressiveness proposed prior to Zhang et al. (2017)'s experiment. Another such measure is the Rademacher complexity which, in essence, measures the ability of a hypothesis class to fit randomly assigned binary labels. Zhang et al. (2017)'s experiment questions the relevance of this notion of complexity as well.

Summarising, since hypothesis classes that are commonly being used have almost infinite capacities, the practical relevance of capacity-centric approaches is being questioned. We believe that by looking at the *expected* ERM risk we can get a more tangible model of generalisation. Although we lose the generality of the statistical learning view, as will be made clearer later in this chapter, being able to control the distribution of risks underlines its central role and brings to the forefront the importance of accounting for the data distribution.

2.2 Computing the Expected ERM Risk

We seek to obtain more informative results than those obtained by considering the capacity of a learning machine. To do so we require more information about the learning problem than in classical learning theory. In particular, we assume that we are given a problem with a fixed data set for which we know the distribution of risks, $\rho(r)$.

Following conventional a priori theory, we imagine that we are given a training data set of size m, where each training example is drawn independently at random from the distribution of data that defines our problem. We model our learning machine by a set, $\mathcal{H}$, of hypotheses. In our formalism the set of hypotheses may be finite or infinite. Each hypothesis, $h \in \mathcal{H}$, will have a loss associated with it. In this section, we consider classification problems where we take the loss function to be 1 for a misclassification and 0 otherwise. As each training data point is sampled independently, for any hypothesis, h, with risk R_h, the loss over the entire training data, denoted L_h, will be binomially distributed

$$\mathbb{P}\left(L_h = \ell | R_h\right) = \text{Binom}(\ell | m, R_h) = \binom{m}{\ell} R_h^{\ell} \left(1 - R_h\right)^{m-\ell} .$$

However, the hypotheses may be correlated. For the moment we will assume that the correlation can be ignored without the end results being significantly affected. This makes the analysis relatively straightforward. The approximation we get by ignoring the correlations is equivalent to the annealed approximation in statistical mechanics (Engel and den Broeck, 2001). We analyse the case in which zero approximation loss can be achieved. That is, among the hypotheses in the training set there exists one which fully captures the data. Such a case is termed the *realisable* case and it is commonly considered in theoretical studies. In this context, the chance correlations between training examples

lead to a systematic correction in the typically observed generalisation performance. As we will show, the annealed approximation gives a reasonable qualitative description of the generalisation performances but it is overly pessimistic. Since this is not relevant to the central story of this thesis, we do not present these calculations in the main body of the thesis but include them in Section C.2 of the Supplementary Material for completeness.

We remind the reader that we consider the scenario known as *Empirical Risk Minimisation* (ERM) where we choose a hypothesis from the $\mathcal{H}_{\text{ERM}}$ subset. Importantly, we assume that every hypothesis in $\mathcal{H}_{\text{ERM}}$ is equally likely to be chosen. This is sometimes referred to as *Gibb's learning* and in the context of training a perceptron, it provides a good approximation to the performance of the learning algorithm.

We are interested in the generalisation performance of the model obtained as a result of the learning algorithm. The algorithm consists of reducing the hypothesis space to a subset, the ERM subset, and then uniformly choosing one of the hypotheses in this set. Thus, in the language of our framework it is the expected risk of such a randomly chosen hypothesis that gives the generalisation performance of our learning algorithm. Throughout the thesis we will refer to this quantity as the expected ERM risk. Formalising, let $R_{\text{ERM}} \in \{R_h | h \in \mathcal{H}_{\text{ERM}}\}$ denote the risk of a randomly sampled hypothesis from the set of hypotheses with minimum loss on the training set. Under the assumptions of our framework, the expected ERM risk is

$$\begin{aligned}\mathbb{E}\left[R_{\text{ERM}}\right] &= \frac{\sum_{h\in\mathcal{H}} R_h \left[\!\left[h \in \mathcal{H}_{\text{ERM}}\right]\!\right]}{\sum_{h\in\mathcal{H}} \left[\!\left[h \in \mathcal{H}_{\text{ERM}}\right]\!\right]} = \sum_{\ell=0}^{m} \frac{\sum_{h\in\mathcal{H}} R_h \left[\!\left[L_h = \ell\right]\!\right]}{\sum_{h\in\mathcal{H}} \left[\!\left[L_h = \ell\right]\!\right]} \left[\!\left[\ell = L_{\text{ERM}}\right]\!\right] \\ &= \sum_{\ell=0}^{m} \mathbb{E}\left[r|\ell\right] \left[\!\left[\ell = L_{\text{ERM}}\right]\!\right] ,\end{aligned}$$

where $\left[\!\left[\text{predicate}\right]\!\right]$ denotes an indicator function equal to 1 if the predicate is satisfied and 0 otherwise, and $L_{\text{ERM}} = \min\{L_h | h \in \mathcal{H}\}$ (i.e. the minimum empirical risk). Note that above we take the expectation over the choice of hypotheses. Following classical statistical learning, we would like to be able to compute the expected generalisation performance over the choice of data set as well. Making the strong assumption that $\mathbb{E}\left[r|\ell\right] \approx \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\left[r|\ell\right]\right]$ (i.e. that there are no significant fluctuations between data sets) then

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\left[R_{\text{ERM}}\right]\right] \approx \sum_{\ell=0}^{m} \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\left[r|\ell\right]\right] \mathbb{P}\left(\ell = L_{\text{ERM}}\right) .$$

In the rest of this section we write $\mathbb{E}\left[\cdots\right] = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}\left[\cdots\right]\right]$ (i.e. the expectation both with respect to the data set and over all hypotheses in $\mathcal{H}_{\text{ERM}}$). Computing $\mathbb{P}\left(\ell = L_{\text{ERM}}\right)$ is inherently problematic in this formalism as it depends on the correlations between

our hypotheses. In the spirit of the approximation we are making we can ignore any correlation between hypotheses. If we do this and define $f_L(\ell) = \mathbb{P}\left(L_h = \ell\right)$ for a random hypothesis $h \in \mathcal{H}$, and let $F_L(\ell) = \mathbb{P}\left(L_h \leq \ell\right) = \sum_{\ell'=0}^{\ell} f_L(\ell')$, then

$$\begin{aligned}\mathbb{P}\left(\ell = L_{\text{ERM}}\right) &= \prod_{h\in\mathcal{H}} \mathbb{P}\left(L_h \geq \ell\right) - \prod_{h\in\mathcal{H}} \mathbb{P}\left(L_h \geq \ell + 1\right) \\ &= (1 - F_L(\ell - 1))^H - (1 - F_L(\ell))^H \,,\end{aligned}$$

where $H = |\mathcal{H}|$, that is, the size of the hypothesis space. For an infinite hypothesis space, we should take H to be some effective size of the hypothesis space (e.g. $2^{D_{VC}}$, where D_{VC} is the VC-dimension). This is the one area in our formalism where capacity plays an important role. We remind the reader that the realisable regime is that where there exists a hypothesis that correctly captures that data. That hypothesis must naturally be among the ERM hypotheses. Thus, for realisable models we do not need to evoke capacity, as we will see below.

From Bayes' rule $f(r|\ell) = \mathbb{P}\left(\ell|r\right)\ \rho(r)\ /\ \mathbb{P}\left(\ell\right)$ so that

$$\mathbb{E}\left[R|\ell\right] = \frac{\int_0^1 r\,\mathbb{P}\left(\ell|r\right)\ \rho(r)\,\mathrm{d}r}{\int_0^1 \mathbb{P}\left(\ell|r\right)\ \rho(r)\,\mathrm{d}r} = \frac{\int_0^1 r^{\ell+1}\,(1-r)^{m-\ell}\,\rho(r)\,\mathrm{d}r}{\int_0^1 r^{\ell}\,(1-r)^{m-\ell}\,\rho(r)\,\mathrm{d}r}\,.$$

Putting together the results above we obtain

$$\mathbb{E}\left[R_{\text{ERM}}\right] = \sum_{\ell=0}^{m} \frac{\int_0^1 r^{\ell+1}\,(1-r)^{m-\ell}\,\rho(r)\,\mathrm{d}r}{\int_0^1 r^{\ell}\,(1-r)^{m-\ell}\,\rho(r)\,\mathrm{d}r}\left((1 - F_L(\ell - 1))^H - (1 - F_L(\ell))^H\right)\,. \tag{2.2}$$

Since in the realisable regime $L_{\text{ERM}} = 0$, the generalisation performance in this case becomes

$$\mathbb{E}\left[R_{\text{ERM}}\right] = \mathbb{E}\left[R|\ell = 0\right] = \frac{\int_0^1 r\,(1-r)^{m}\,\rho(r)\,\mathrm{d}r}{\int_0^1 (1-r)^{m}\,\rho(r)\,\mathrm{d}r}\,. \tag{2.3}$$

To be able to analyse the generalisation performance from this point onward, we would need to know the distribution of risks, $\rho(r)$. In Section 2.3.1, we look at how to compute $\rho(r)$ for a specific problem, namely the realisable perceptron. In general, however, this is a hard task. We can obtain an estimate of this quantity in practice through MCMC sampling (Belcher et al., 2022).

By examining how the generalisation performance depends on $\rho(r)$, we can obtain a better understanding of what is required to improve generalisation. For this, we have to make assumptions about the distribution of risks. In the following section we study the generalisation performance where we assume the risks are distributed according to a Beta distribution. We call this the β-Risk model of generalisation.

Is assuming a Beta distribution reasonable? In Section B.1 we look at the asymptotic generalisation in a realisable setting and show that for an infinite hypothesis space, generalisation is given by the exponent of the risk. In other words, we show that it is the *power-law growth* in the distribution of risks that is the driving factor of the generalisation performance. Therefore, our calculations have generality beyond problems with a Beta-distributed risk. Subsequently, in Section 2.3.1 we show that our model allows us to capture the generalisation performance of the realisable perceptron, which further validates our proposal. Nonetheless, the ability of our model to capture all aspects of generalisation as well as possible alternatives to β-Risk remain open research questions.

2.2.1 Classification: β-Risk Model

Starting from Equation 2.3 we can numerically compute the expected ERM risk from a knowledge of the distributions of risks, $\rho(r)$. In this section, we consider a special form of $\rho(r)$ that allows us to compute the integrals in closed form. That is, we take $\rho(r)$ to be Beta-distributed,

$$\rho(r) = \text{Beta}(r|a,b) = \frac{r^{a-1}\,(1-r)^{b-1}}{B(a,b)}\ . \tag{2.4}$$

For a balanced data set where we perform a binary classification task we would choose $b = a$, while for k-way classification we would set $b = a/(k-1)$ so that $\mathbb{E}\left[R_h\right] = (k-1)/k$. Note that this distribution is unbiased, so, for example, in the binary case, there are as many poor hypotheses as good ones. We call this the β-Risk model.

The parameter a measures the degree of "attunement": the smaller a the more attuned the hypothesis class $\mathcal{H}$ is to the problem being solved. The β-Risk model allows us to obtain an intuitive understanding of the generalisation performance in this framework. This seems a very particular functional form for $\rho(r)$. However, as mentioned earlier, for large m the expected ERM risk is dominated by the power-law growth in $\rho(r)$, so that the β-Risk model provides a reasonably accurate approximation for many different learning scenarios. To further verify the validity of our model, we explicitly compare the results obtained for the perceptron using the true $\rho(r)$ and a β-Risk model with the same asymptotic behaviour in Section 2.3.1.

For the β-Risk model the distribution of learning errors is given by

$$f_L(\ell) = \mathbb{E}_R\Big[f(\ell|R)\Big] = \binom{m}{\ell} \frac{B(a+\ell, b+m-\ell)}{B(a, b+m)} \,. \tag{2.5}$$

The conditional probability of a risk, r, given an empirical loss of ℓ is

$$f(r\,|\,\ell) = \frac{\mathbb{P}\left(\ell\,|\,r\right)\,\rho(r)}{\mathbb{P}\left(\ell\right)} = \frac{r^{\ell+a-1}\,(1-r)^{m-\ell+b-1}}{B(\ell+a, m-\ell+b)} \,, \tag{2.6}$$

from which we find

$$\mathbb{E}\left[R|\ell\right] = \frac{\displaystyle\int_0^1 r^{\ell+a}\,(1-r)^{m-\ell+b-1}}{B(\ell+a, m-\ell+b)} = \frac{B(\ell+a+1, m-\ell+b)}{B(\ell+a, m-\ell+b)} \tag{2.7}$$

$$= \frac{\Gamma(l+a+1)\,\Gamma(a+m+b)}{\Gamma(a+m+b+1)\,\Gamma(l+a)} = \frac{a+\ell}{m+a+b} \,. \tag{2.8}$$

The β-Risk model is a *realisable problem* in the limit $H \to \infty$ since $\inf_{h\in\mathcal{H}} R_h = 0$. That is, there exists a learning machine with arbitrarily small risk. In this case, the expected ERM risk is $\mathbb{E}\left[R_{\text{ERM}}\right] = a/(a+b+m)$.

Next, by considering a finite hypothesis space, a common abstraction in statistical learning theory, we can use β-Risk to model unrealisable problems. Unrealisable problems are those where all hypotheses have a risk that is greater than zero. If we assume that our hypothesis space corresponds to samples drawn from a continuous parameter space of a learning machine then such an unrealisable problem would be one where $\rho(r) = 0$ for all $r < R_{\text{min}}$. If we sample from $\rho(r)$ then all hypotheses will have a risk greater than or equal to R_{min}. Inserting Equation 2.8 into the expected ERM Risk given in Equation 2.2 we obtain

$$\mathbb{E}\left[R_{\text{ERM}}\right] = \sum_{\ell=0}^{m} \frac{a+\ell}{m+a+b} \left((1-F_L(\ell-1))^H - (1-F_L(\ell))^H\right) \,. \tag{2.9}$$

Figure 2.2 shows the expected ERM risk versus m plotted on a log-log scale for the case when $a = 10^2$ and $a = 10^3$ with different sized hypothesis spaces. This allows us to get a quick intuition about the generalisation behaviour for unrealisable problems.

We see in Figure 2.2 that, for a given attunement a, we can obtain better results for larger hypothesis spaces. This is because larger hypothesis spaces are likely to include lower-risk hypotheses. Of course, in a concrete scenario increasing the size of the hypothesis space would most likely implicitly modify the level of attunement. This is simply to say

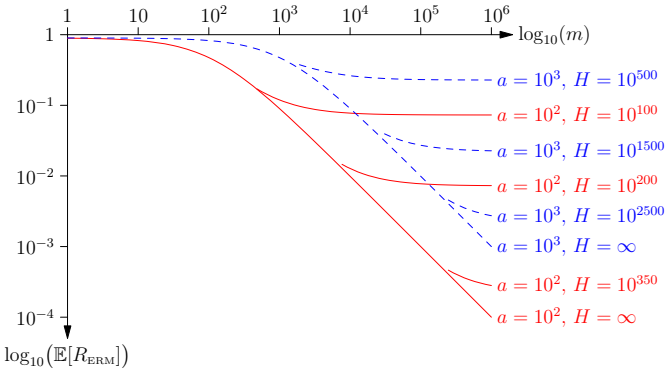


FIGURE 2.2: Expected ERM risk versus the number of training examples plotted on a log-log scale for $a = 10^2$ (blue) and $a = 10^3$ (red), with $b = a/9$ (i.e. for a 10 class problem) for different values of H.

that given the same level of attunement, one is more likely to find lower-risk hypotheses in higher hypothesis spaces.

When providing bounds on the asymptotic behaviour, standard statistical learning theory makes a strong distinction between the realisable and unrealisable learning scenarios. In our framework, we observe that there is a zero-loss phase and a nonzero-loss phase in the generalisation curves. For small m some proportion of the learning machines are able to perfectly classify the training examples. If $\rho(r)$ is well approximated by a Beta distribution around $\mathbb{E}\left[R_{\text{ERM}}\right]$ then $\mathbb{E}\left[R_{\text{ERM}}\right] \approx a/(a+b+m)$ — this characterises the zero-loss phase. When $\mathbb{E}\left[R_{\text{ERM}}\right]$ approaches the minimum risk R_{min} (the risk of the best learning machine in $\mathcal{H}$) then $\mathbb{E}\left[R_{\text{ERM}}\right]$ will converge towards R_{min}. For realisable scenarios, $\mathbb{E}\left[R_{\text{ERM}}\right]$ will remain in the zero-loss phase for all m.

In our framework the role of the VC-dimension is played by the attunement parameter a. This captures a quite different concept from class expressiveness, namely how quickly does $\rho(r)$ fall off as $r \to 0$. If the learning machine is well attuned to the problem we would expect this to fall off relatively slowly. Note that, whereas the capacity depends only on the learning machine architecture, the attunement also heavily depends on *the distribution of data*.

In the following section we study the case of a well-attuned perceptron. We calculate its risk probability density and relate back to our β-Risk model to analyse changes in attunement as a result of feature reduction.

2.3 The Distribution of Risks: Case Study

Key to our formalism is the need to know the distribution of risks, $\rho(r)$, for a learning problem. To validate our model, in this section we compute $\rho(r)$ for a realisable perceptron. Using the inferred distribution of risks we match the generalisation performance

of the perceptron with that modeled by β-Risk and show that our model can capture reasonably well the generalisation performance for the concrete scenario we study.

2.3.1 Realisable Perceptron

We consider a very simple learning scenario. Our training set corresponds to m pairs $(\boldsymbol{x}_i, y_i)$ where $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y_i = \text{sgn}(\boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}^*)$. That is, $y_i = 1$ if the data is positively correlated with the target vector $\boldsymbol{w}^*$ that defines the separating plane and $y_i = -1$ otherwise. We consider learning this with a perceptron with p-dimensional weights $\boldsymbol{w}$ such that $\|\boldsymbol{w}\| = 1$. That is, the weights live on a hypersphere. If we consider sampling uniformly from the set of weight vectors then the distribution of weight vectors with an angle θ to $\boldsymbol{w}^*$ is

$$f_\Theta(\theta) = \frac{\sin^{p-2}(\theta)}{B(\frac{1}{2}, \frac{p-1}{2})} \,. \tag{2.10}$$

For this problem the risk is given by $r = \theta/\pi$ so that $\rho(r) = \pi\, f_\Theta(\pi\, r)$. This is a realisable model for which the expected ERM risk, under the assumption of the annealed approximation, is

$$\mathbb{E}\left[R_{\text{ERM}}\right] = \mathbb{E}\left[R|\ell = 0\right] = \frac{\int_0^1 r\,(1-r)^m\, \sin^{p-2}(\pi\, r)\, \mathrm{d}r}{\int_0^1 (1-r)^m\, \sin^{p-2}(\pi\, r)\, \mathrm{d}r} \,.$$

We can compute this numerically. However, when m is large the dominant contribution to the integral comes from where r is small. In this region $\rho(\pi\, r)$ grows as r^{p-2} (since $\sin(\pi\, r)$ grows linearly with r for small r). Thus we can approximate $\rho(r)$ by a Beta distribution $\text{Beta}(r|p-1, p-1)$ for which $\mathbb{E}\left[R_{\text{ERM}}\right] = (p-1)/(2\,p-2+m)$.

In Figure 2.3, we show $\mathbb{E}\left[R_{\text{ERM}}\right]$ as a function of the number of training examples, m, for the realisable perceptron and the β-Risk model with $a = b = p-1$. Note that the expected risk for the realisable perceptron is computed numerically. We see that the β-Risk model provides a good approximation to the realisable perceptron in the annealed approximation.

For this simple scenario, the distribution of risks, and hence the attunement, is directly determined by the dimensionality of the vector $\boldsymbol{w}^*$. If $\boldsymbol{w}^*$ is orthogonal to some of the features, then they can be removed, improving generalisation. Traditionally, this would be attributed to reducing the size of the hypothesis space. However, we see that this also leads to an improvement in attunement. Comparing the solid curves in Figure 2.3, we can see the improvement in the expected risk when reducing features that do not affect the generalisation performance. In the example depicted in this figure, removing

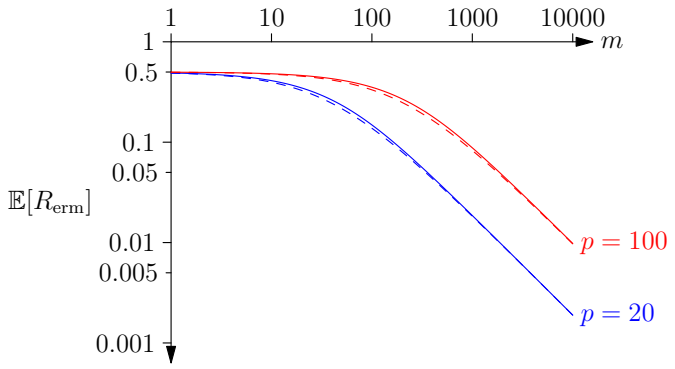


FIGURE 2.3: Expected ERM risk for the realisable perceptron for the cases $p = 20$ and $p = 100$. We also show as dashed curves are $\mathbb{E}\left[R_{\text{ERM}}\right] = (p-1)/(2p-2+m)$ corresponding to a β-Risk model with $a = p-1$. Removing 80 features leads to a better generalisation curve. Fitting a β-Risk model to the two perceptron cases, we note that the value of a has decreased from 99 to 19. Therefore removing the 80 features has increased attunement. Note once again that a smaller value of a indicates better attunement.

80 features leads to better generalisation performance. Fitting our β-Risk model to the two perceptrons, one with 100 features and the other with 20 features, we observe that removing the 80 features has led to *an improvement in attunement*.

In Section C.2 of the Supplementary Material we do the same calculation for the unrealisable perceptron and once again note the central role of attunement and, in turn, of the data. To ensure our calculations are correct, in Section C.2 we take a closer look at the assumptions we make. This allows us to compare our results against those obtained through Gardner (1988)'s replica calculation. The corrections we obtain give very similar asymptotic generalisation results to Gardner's.

We believe that the good performance of modern deep learning algorithms can be explained by their attunement. It is important to note that we do not yet have a good understanding of exactly how changes in the data or model class shape the attunement. This is what we attempt to build an intuition for in the second half of the thesis. Once we have an understanding of how the attunement changes, we can use our model to further reason about generalisation. However, the important insight gained from this result is the *importance of the data*, which we believe has not been sufficiently accounted for in prior studies.

Coming back to Zhang et al. (2017)'s experiment, we can informally reason about it in terms of attunement. Because the training data has random labels, then all the hypotheses in the ERM subset will have high risk, since they can have no better than random generalisation power. In our β-Risk model, this corresponds to very low attunement. Therefore randomising labels changes the attunement. Although the architectures Zhang et al. experiment with have high capacity, they *can* have good generalisation performance provided that they are well-attuned to the problem. When the labels are not randomised, the attunement will be significantly better than that on randomised labels. A similar argument is given by the concurrent work of Wilson and Izmailov

(2020) who take a probabilistic approach to generalisation. They argue that it is the "fitness" between the model's inductive biases and the data distribution that determines generalisation.

2.4 Future Work

The most important endeavour for future work is to understand how to capture and formalise attunement. This, however, would imply solving the generalisation problem and it is therefore a difficult problem. The purpose of this chapter was to emphasise the importance of accounting for the data and the limitation of classical studies that comes from abstracting away from the data. In the rest of this thesis we advocate for a *data-centric empirical approach* to understanding generalisation. Although such an approach does not have formal rigour, the hope is that it will guide us towards a theoretical formulation once we have a better handle on what drives generalisation.

Regularisation. An immediate direction to consider is the effect of regularisation on attunement. It is believed in the community that regularisation improves performance because it reduces the capacity of the model (Neyshabur et al., 2015; Hernández-García and König, 2018). We believe, however, that the capacity is imperceptibly affected by regularisation, if at all and wish to study how adding regularisers such as the L_2 impact the distribution of risks for small risks. Aiming to get an intuition of a realistic setting for this problem, we have considered empirically determining the risks obtained when performing linear regression for the Maximum Satisfiability Problem. For this, we chose the predictor to be a discrete Fourier transform and we calculated the risks as we restricted the complexity of the function. However, the distribution of risks for this particular problem is too specific. Thus, to ensure that our findings are sufficiently representative, analysing more problem classes is necessary.

Separability. One other natural question to ask is how the attunement changes with the increase in class separability. We believe the reason why deep learning architectures are so successful is because the powerful feature extractor is increasing attunement by facilitating class separation. Through our framework, we can analyse this effect very easily for the case of the unrealisable perceptron by varying the distance of the two class means.

Class Imbalance. Similarly, we believe that studying the attunement in the case of unbalanced data sets will further our understanding of the different mechanisms that determine attunement and has the potential to hint towards more practical approaches to alleviate the effects of the imbalance.

As we have mentioned in Chapter 1 and will reiterate throughout the thesis, deeply engaging with the theoretical approach has allowed us to understand that in the absence

of better practical intuitions, we cannot formalise generalisation in a meaningful way. Although we can do calculations for simple models, we cannot yet scale them to the types of architectures and high-dimensional data sets used in practice. Therefore, the understanding that we can get from such calculations is also limited. Although the above directions for future work are interesting from a purely theoretical point of view, we believe that pursuing them is unlikely to help us find the missing ingredient in generalisation studies, which is the goal of this thesis. As such, in the following chapters we steer towards a more experimental approach.

As we will detail in Chapter 5, at the time of drafting this chapter, we became aware of Zhang et al. (2018a)'s work, which introduced us to the field of mixed data augmentation. Studies centred around data modification have then piqued our interest, since, like our calculations, they were attesting to the importance of the data. Albeit at the cost of losing formal rigour, the data modification literature is providing a more practical way of building intuitions. Therefore, in Chapter 3, we focus on the practice of modifying data. The hope is that the understanding gained from this pursuit will take us one step closer to building a theoretical framework that provides informative results.

2.5 Conclusions

Traditional machine learning theory has universal applicability in that it provides bounds on the generalisation gap that depend only on the capacity of the learning machine and are independent of the problem being tackled. This apparent strength is also its weakness. A learning machine with a large capacity may or may not generalise well depending on the distribution of the data. We know there exist distributions of data for which we cannot get any tighter bounds, so obtaining tighter ERM risk bounds requires us to include information about the data distribution. We have done this by considering the distribution of risks, which depends on the alignment between the learning machine and the problem. Therefore, the distribution of risks implicitly depends on the data.

The cost of considering the distribution of risks is that we lose a lot of the elegance of traditional machine learning. Instead of hard bounds, we are left with approximate results for the expected ERM risk. There are, however, advantages: we know that a poorly attuned problem will require a large number of training examples and we have a model for the generalisation performance rather than just the generalisation gap. We can improve on the annealed approximation, but this requires additional information about the learning machine. However, the annealed approximation provides a qualitatively accurate model that captures many of the generalisation properties of the exact system. The most important takeaway, in our view, is that generalisation is heavily determined by attunement.

The attunement measures the power-law behaviour of $\rho(r)$ around $r = 0$. This determines the asymptotic learning behaviour for realisable models. We believe this provides a new language for understanding generalisation performance. To design a successful learning machine it is not necessary to limit the capacity but to obtain good attunement (i.e. ensure that there is a relatively high proportion of low-risk machines). We believe this shift in thinking will aid the design of learning machines in the future.

A natural question is what is the exact mechanism behind attunement. Although one can try to formulate a number of intuitions, finding an exact description is a very difficult pursuit. For the rest of the thesis we seek to build stronger intuitions with the hope that these will inform researchers how to start theorising about this intricate mechanism. Until then, one of the immediate takeaways of our theoretical work that is going be the driving force of our empirical endeavour is that *data is a core aspect of generalisation.* Although the observations we make around our calculations point towards integrating information about the problem, in one way or another this has been attempted in other studies. From the theoretical end, such studies can be found in the PAC-Bayesian literature, which we discuss in more detail in Chapter 4.

In the PAC-Bayesian framework, estimators are defined in terms of data-dependent probability measures from which predictors are either drawn or aggregated. As such, McAllester (1999) bounds the risk of a randomly drawn estimator using the Kullback–Liebler divergence between the distribution over estimators and a fixed prior probability distribution. For a simple introduction, see Alquier (2021). These bounds have been recently applied in the context of overparametrised deep learning models (e.g. Dziugaite and Roy, 2017), albeit on architectures significantly smaller than those used in practice. Despite being tighter than many other bounds, they are still not sufficiently descriptive (Neyshabur et al., 2017). We will discuss this in more detail in Chapter 4.

In turn, the most recent and allegedly successful empirical estimators of generalisation have resorted to modifying the data. We will study such estimators in Chapter 4. Before doing so, we take a closer look at data modification as a practice in the next chapter. We want to better understand the implications data distortion has on model training and evaluation. This is because they are highly used in the field without being thoroughly researched but also because we believe we can better understand the role of the data by approaching it in an incremental manner. Thus, the next chapter is dedicated to data distortion.

Chapter 3

Steps Towards the Empirical Approach: Understanding by Distorting

In this chapter we focus on data distortion as a way of studying learned representations. We start by highlighting the neglected side effects of data distortion when both training and evaluating models. Our empirical findings challenge core assumptions in the field, raise a number of important questions and subsequently motivate a distortion-based approach to understanding generalisation.

The previous chapter has emphasised the importance of accounting for the data. While in that particular theoretical setting we were able to do so by looking at the distribution of risks, in practice we have no tools to integrate information about the data or properly describe it and capture its underlying attributes.

We believe one way of grasping complex concepts is by breaking them down and this is what this chapter is striving to achieve. Instead of aiming to understand data as a whole, we focus on studying the effects of *incrementally changing the data*. Despite its great potential, this topic has largely remained unexplored. A case in which it has been somewhat studied is that of data augmentation. But as remarked in Harris et al. (2020), previous studies have a number of inconsistencies and, as a field, we are far from having a clear description of how distorting data affects learning. Being able to fully describe this phenomenon is in itself highly challenging and beyond the scope of this work.

As we have briefly mentioned in the Foreword and Chapter 2, empirical predictors of generalisation are starting to incorporate data modification in their approaches. Therefore, rather than aiming to fully understand data distortion, this chapter starts building the intuitions necessary for understanding *how data modification can be used in a more*

principled way for the purpose of empirically predicting generalisation. During this process, we propose new promising directions for understanding distortions and we expose a number of issues that arise when distortion is not thoroughly studied.

The contributions in the first half of this chapter were presented at the Robust ML workshop as part of the International Conference on Learned Representations (ICLR) 2021. The second half was included in a lightning talk presentation at the Data-Centric AI workshop at the NeurIPS 2021 Conference. As a whole, a condensed version of this chapter was submitted to the NeurIPS 2021 Conference. Despite having three out of four reviewers in favour of acceptance, it was finally rejected on grounds which we had already addressed during the rebuttal period. The same work was subsequently submitted and accepted for publication as a spotlight paper at the International Conference on Machine Learning (ICML) 2022. The contents of this chapter along with our vision for the future of generalisation studies, which we will introduce in Chapter 4, were also presented as an invited talk at the Max Planck Institute for Intelligent Systems Tübingen.

3.1 Context and Prior Art

Augmentation is commonplace when training models. It is a form of data modification where samples are artificially distorted to create larger training sets. Apart from augmentative purposes, data modification is also used for a wide range of model analysis methods. Most recently, distortion-based approaches have been adopted when trying to answer key machine learning questions. To this end, MixUp-like distortions (Zhang et al., 2018a) were proposed for empirically predicting generalisation (Schiff et al., 2021; Natekar and Sharma, 2020; Lassance et al., 2020). Thus, data modification is becoming increasingly popular, but little attention is paid to the secondary effects of this practice. As we will demonstrate, *our current understanding of the effects of data modification lies on fundamentally flawed assumptions.* This impacts not only our perception of what features are important to our models, but also the correctness of the distortion-based approaches we propose as a field.

In this chapter we study the implicit assumptions made when artificially distorting data and their implications. From the model analysis perspective, we take occlusion robustness and shape bias identification methods as examples of where modified data is used. On the training side, we focus on some instances of Mixed Sample Data Augmentation (MSDA), where two images are combined to obtain a new training sample. Visual illustrations of each can be found in Figure 3.1. In this chapter we delve into some of the side effects of data modification and point out that this practice has resulted in the creation of biased model interpretation tools and poorly informed theories. More specifically, we study a number of assumptions which we show are erroneous and which lie at the heart of the methods we briefly introduce below. Contesting these assumptions has broader

FIGURE 3.1: Examples of image distortions on ImageNet (224 × 224 pixels) samples. Occlusion and Shuffled grid are typically used for evaluating models, while MixUp, CutMix and FMix are distortions used for model training. For test-time distortions, only Image 1 was used. For mixing augmentations, the first row was generated with a mixing factor of 0.2, while the second one with 0.5.

implications on the community's perception of what aspects of the data are important when learning.

Shape-texture Bias. Deep models are known to be sensitive to distribution shifts (Alaiz-Rodríguez and Japkowicz, 2008; Cieslak and Chawla, 2009; Engstrom et al., 2019) and interventions that are imperceptible to humans (Szegedy et al., 2014; Goodfellow et al., 2015). It has been argued that this is intimately linked to networks tending to use texture rather than shape information (Brendel and Bethge, 2019; Geirhos et al., 2019). Recently, input distortions have become a popular way of assessing a model's texture bias. To this end, images are divided into a grid and the resulting patches are randomly shuffled such that information is preserved locally, while the global shape is altered (Shi et al., 2020; Mummadi et al., 2021; Luo et al., 2019; Zhang and Zhu, 2019). *It is implicitly assumed that patch-shuffling does not introduce misleading shape or texture that could affect model evaluation.* As such, if a model's accuracy drops when evaluated on patch-shuffled images, this degradation in performance is entirely attributed to the model's bias for shape information. Thus, any side effects of the data manipulation process are considered negligible.

Occlusion Robustness. Commonly, occlusion robustness is concerned with the amount of information that can be hidden from a model without affecting its ability to classify (e.g. Tang et al., 2018; Rajaei et al., 2019). A widely adopted proxy for measuring occlusion robustness is through the raw accuracy obtained after superimposing a rectangular patch on an image (Chun et al., 2020; Fawzi and Frossard, 2016; Yun et al., 2019; Zhong et al., 2020*b*; Kokhlikyan et al., 2020). We refer to this approach as CutOcclusion throughout the thesis. Just as with shape bias, *this method relies on the introduced information not to interfere with a model's learnt representations* such that a decrease in performance can be directly attributed to a lack of robustness. Thus, using CutOcclusion, one implicitly assumes that artefacts do not interfere with the results of robustness evaluation.

Data Augmentation Studies. In statistical learning, training with augmented data is termed Vicinal Risk Minimisation (VRM) (Vapnik, 1999*a*; Chapelle et al., 2001) and it is seen as injecting prior knowledge about the neighbourhood of the data samples. The intuition behind augmentation caused researchers to interpret its effect through the similarity between original and augmented data distributions. This perspective is often challenged by methods which, despite generating samples that do not appear to fall under the distribution of natural images, lead to strong learners. Gontijo-Lopes et al. (2021) argue it is the *perceived* distribution shift that needs to be minimised, while maximising the sample vicinity. Formalising these concepts, they introduce augmentation "diversity" and "affinity". Diversity is defined as the training loss when learning with artificial samples, while affinity quantifies the difference between the accuracy on original test data and augmented test data for a reference model. The latter penalises augmentations that introduce artificial information to which the model is not invariant, *implicitly assuming that training with that information is detrimental to generalisation.* Thus, in contrast to the evaluation methods mentioned above, the artefacts are considered non-negligible when training with distorted data.

In summary, it is currently assumed that the artefacts introduced by changes in the data are negligible when evaluating models, while those introduced when training are important and undesirable. These assumptions implicitly shape the community's perception of how machine learning works. Does the artificial information added by analysis methods not have major side effects or does it lead to biased results? Conversely, are the artefacts important when training with modified data? Do they cause models to learn better or worse representations?

We set out to answer these questions. We find that results can be misleading when not accounting for the secondary effects of data manipulation, especially in comparative studies. Taking an oversimplified example to illustrate this for robustness, we can imagine a binary cat–truck image classification problem and two models: model A, which identifies cats solely by the presence of ears and model B, which has a more holistic approach. Generally, masking out the ears will cause model A to misclassify cats and we would consider this model not robust to occlusion, while model B will continue to correctly classify them. However, if the ears are covered with a large rectangle that introduces horizontal and vertical edges strongly associated with the "Truck" class, this will cause model B to also misclassify. In this case, because the misclassification is not caused by the absence of a feature but rather by the presence of a distractor, we would still consider model B robust to occlusion, although its performance degrades. In such a case, CutOcclusion would be unable to distinguish between a model that incorrectly classifies because of lack of information or because of the presence of confounding artefacts, making it an incorrect proxy for measuring occlusion robustness. Thus, the side effects of data distortion must be taken into account to create fair evaluation methods.

We start by introducing the Data Interference index, a measure that highlights the existence of such side effects. We then use the index to disprove the negligibility of data distortion artefacts when evaluating models. Subsequently, using the existence of artefacts, we disprove common beliefs in the augmentation literature through empirical counter-examples. This further motivates the importance of understanding the changes data manipulation introduces. Our contributions are:

- We introduce a quantity for highlighting the interference of artificially introduced visual information with a model's learnt representations (Section 3.2);
- We show that increasingly popular model interpretation and analysis methods are biased, relying on unfounded assumptions (Section 3.2);
- For measuring occlusion robustness, we propose a fairer alternative (Section 3.3);
- We show that, in contrast to what is widely assumed, not preserving the data distribution can lead to learning better representations (Section 3.4).

While the impact of our practical contributions is relevant to the community, we believe more important for the future development of the field is combatting erroneous research directions. Correctly understanding the increasingly popular mixed-sample augmentation is essential for trusting its usage in sensitive applications where the data can be out of distribution. Moreover, studying the effects of image distortion can shed further light on how vision models perceive changes to elemental features. But most importantly, we believe this could set a new direction in capturing the relationship between data and learned representations, which could ultimately play a role in understanding generalisation.

3.2 Are Artefacts Negligible when Analysing Classifiers?

In this section we show that artificially introduced artefacts may not be negligible, and distorting data at evaluation time could have side effects not previously considered. Specifically, the artefacts can interfere with the representations learnt by the model, which in turn leads to incorrect evaluation. We highlight this interference by showing that *the distortion can be consistently associated with a particular class in an image classification task*. We do so by looking at the increase in misclassifications per predicted category; from the number of incorrect predictions of a model evaluated on modified data, we subtract the incorrect predictions when testing on original data. If, across multiple training runs, there is a significant increase for a specific class, this indicates that the distortion introduces features the model associates with that class. We refer to this phenomenon as "data interference".

By a model "run" we refer to a model instance trained with a different seed for the initialisation of weights and for the randomised data augmentation. For computing the DI index we train 5 different instances of the same model and only consider that data interference occurs when runs consistently display a bias for the same class. Considering only positive differences, we denote the increase in the percentage of misclassifications for predicted class c for a run r by c^r. Note that the class is taken to be that predicted by the classifier. To keep the score within a consistent range across data sets, we scale c^r by the number of classes. We define the Data Interference (DI) index as

$$\mathbb{E}_r \left[\frac{c^r_{max}}{\sum_c c^r} \mathbb{E}_{r'} \left[c^{r'}_{max} \right] \right], \tag{3.1}$$

where c_{max} is that of the class with the highest mean increase across all runs. Note that the inner expectation does not depend on r. We chose to write the index this way to make it easier to match the informal definition, which is that the DI index measures the proportion represented by the dominant class weighted by its average increase across runs. A high index value indicates a sharp increase for a particular class which is consistent across runs. We associate this with an overlap between introduced artefacts and learnt representations, thus highlighting the *side effects of distorting data for model evaluation.*

We also experiment with an alternative index, where we weigh by the highest increase of a model across the 5 runs, so as to obtain a worst-case analysis. As expected, we observe a stronger bias in this case, evidenced by an increased gap in the DI index. We will exemplify this in the case of shape bias in Section 3.2.1. Although the worst-case formulation makes our arguments more evident, for the rest of the chapter we present the results using the more restrictive definition so as to show that our findings hold in more general settings.

To obtain models with different behaviours in a controlled manner, we make use of data augmentation. Since it is sufficient to identify some common cases in which models are disfavoured, we choose to reduce our environmental impact by restricting the analysis to simple MSDAs that combine images without incurring additional computation time or external models. As will be argued in Section 3.3, we expect the unfairness to be present in most settings, thus the exact choice of augmentation is irrelevant. We focus on two popular MSDAs, MixUp (Zhang et al., 2018a) and CutMix (Yun et al., 2019). MixUp linearly interpolates between two images to obtain a new training example, while CutMix masks out a rectangular region of an image with the corresponding region of another image. Besides the aforementioned methods, we also employ FMix (Harris et al., 2020), a mixed-sample augmentation that samples masks from Fourier space. We choose to use FMix due to its irregularly shaped masks, which will play an important role in our analysis.

Note that although the masking methods sample the size of the occluding patch from the same distribution, in CutMix part of the rectangle can be outside the image, which leads to less occluded samples overall compared to FMix. A second important observation is that interpolating methods (i.e. MixUp and its variants) modify every pixel of the image, while masking methods (in this thesis CutMix and FMix) are more localised. We will come back to this difference later on in the chapter.

We refer to models by the augmentations they were trained with and use "basic" to label the models trained without MSDA. Throughout this chapter, we do five runs of each experiment with PreAct-ResNet18 (He et al., 2016*b*) as the default architecture. We explicitly state when other architectures (i.e. BagNet (Brendel and Bethge, 2019), VGG (Simonyan and Zisserman, 2015)) and ResNet-101 (He et al., 2016*b*) are used. The main data sets we report results on are CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), Tiny ImageNet (Karpathy et al., n.d.), FashionMNIST (Xiao et al., 2017), and ImageNet (Russakovsky et al., 2015). For ImageNet we use pretrained ResNet-101 models made publicly available by Harris et al. (2020). Note that the only experiments for which we are unable to run repeats are those on ImageNet, since only one model per augmentation is provided. For all other experiments, we give the average and standard deviation results obtained across the five runs. For full experimental details, see Section D.1.

As we will discuss below, the DI index has a high standard deviation. For this reason, for computing the DI index, besides taking the average across the five runs we also experimented with performing five distortion iterations for each image in the data set. That is, we computed the increase in incorrect predictions of each model instance when randomly occluding every image five different times. However, we did not see a difference in results since the bias for associating distortions with a particular class seems to be very consistent for each model. The high deviation stems from the difference in the learnt bias between model runs.

DI Confidence Evaluation and Gaussian Noise

Before presenting our findings, we briefly focus on the observation that for the models we evaluate, the DI index has a high standard deviation. For example, in the case of patch-shuffling, the modification for assessing shape bias, the value of the DI index is $2.40_{\pm 0.59}$ for the basic model (i.e. the model trained without mixed data augmentation) on CIFAR-10. In the case of the CutMix-trained model, we measure a DI of $0.31_{\pm 0.10}$. As we will emphasise when presenting the full results, the DI index must be interpreted in a comparative manner.

How can we trust that the results have statistical significance? To answer this question, we want to determine to what extent the DI we measured is a result of chance. We,

therefore, calculate the result of DI for a *random* increase in misclassifications. For this, we compute the increase in CIFAR-10 misclassifications for the basic model just as we do for computing the DI index. Then, we redistribute each misclassified sample to a random class and compute the DI index on this random assignment of the increase in misclassification. We repeat this experiment 10^5 times. We obtain an average DI of 0.10 with a standard deviation of 4×10^{-5}. This result, highly concentrated around 0.10, tells us that there is an infinitesimally small chance that the DI index we observe for the basic model is a chance result of a random phenomenon. Thus, despite the high standard deviation, our index can inform us when one model is more affected than another by the artefacts of a distortion.

To make our baseline more challenging, we then randomly assign **all** misclassified examples to one of the classes and compute the index across 5 iterations. This is to simulate the case where each specific model instance consistently associates artefacts with a certain class, but the class is not necessarily consistent across runs. We perform this experiment 10^5 times and obtain a DI of $0.10_{\pm 0.30}$. Once again, there is a significant difference between the DI obtained by chance and the true DI index of the basic model ($2.40_{\pm 0.59}$).

Finally, we compute the DI index for distorting with Gaussian noise. We distort the samples after the image normalisation step, when the pixel values are between 0 and 1. For each image, we uniformly sample the standard deviation of the Gaussian noise from the interval $[0, 0.1]$, with a mean of 0. For the basic model trained on CIFAR-10, we obtain a DI value of $0.09_{\pm 0.02}$. This is significantly lower than the index we observe in this chapter in the case of patch-shuffling and rectangular occlusion. We have therefore seen through three different experiments that our index captures a real phenomenon that cannot be replicated by chance. The DI index helps us identify in a comparative way when a distortion introduces features that are associated with the learnt representation of a class. Although this is sufficient for the purpose of this thesis, it would be interesting to explore the idea of data interference outside the scenario we are concerned with. We reflect on the limitations of applying this index in wider contexts in Section 3.6.

3.2.1 Shape Bias Measurement

Using the DI metric, we want to show the existence of side effects that occur when measuring shape bias based on the accuracy after patch-shuffling images. We argue that these side effects make the patch-shuffling evaluation method unreliable. For assessing shape bias through sample manipulation, the standard procedure is to choose between dividing the image in 4, 16 or 64 patches to be shuffled. Since FashionMNIST images are smaller, we choose a 2×2 grid, while for CIFAR-10/100, Tiny ImageNet and ImageNet we use a 4×4 grid. However, similar results are obtained for different grid sizes (see Section D.2 of the Supplementary Material).

TABLE 3.1: DI index for PreAct-ResNet18 on grid-shuffled images for four different types of models. Results with the highest average are given in italic and the lowest in bold. Information introduced when shuffling tends to interfere less with the representations of FMix and CutMix models, as indicated by the lower DI values.

	basic	MixUp	FMix	CutMix
CIFAR-10	$\mathit{2.82}_{\pm \mathit{0.44}}$	$2.40_{\pm 0.59}$	$0.59_{\pm 0.12}$	$\mathbf{0.31_{\pm 0.10}}$
CIFAR-100	$\mathit{0.99}_{\pm \mathit{0.27}}$	$0.88_{\pm 0.24}$	$0.18_{\pm 0.10}$	$\mathbf{0.09_{\pm 0.04}}$
Fashion	$1.23_{\pm 0.15}$	$\mathit{2.42}_{\pm \mathit{0.92}}$	$1.06_{\pm 0.23}$	$\mathbf{0.68_{\pm 0.11}}$
Tiny	$\mathit{1.28}_{\pm \mathit{1.13}}$	$0.57_{\pm 0.11}$	$0.67_{\pm 0.10}$	$\mathbf{0.25_{\pm 0.11}}$
ImageNet	0.82	*1.49*	**0.58**	–

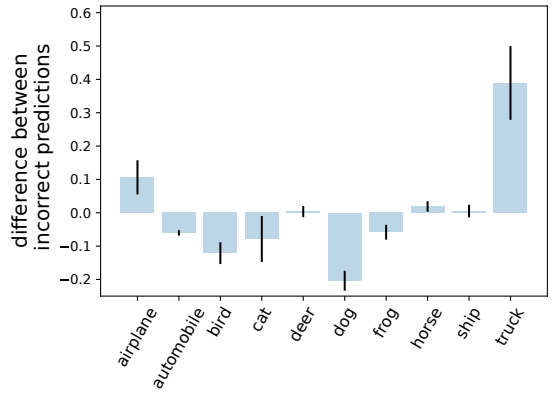
TABLE 3.2: Alternative DI index for PreAct-ResNet18 on grid-shuffled images for four different types of models. Again, a gap in bias can be noted for all considered data sets, with the basic and MixUp models typically showing higher data interference when evaluated on patch-shuffled images.

	basic	MixUp	FMix	CutMix
CIFAR-10	$\mathit{3.52}_{\pm \mathit{0.56}}$	$3.31_{\pm 0.82}$	$0.76_{\pm 0.16}$	$\mathbf{0.43_{\pm 0.13}}$
CIFAR-100	$\mathit{1.40}_{\pm \mathit{0.38}}$	$1.09_{\pm 0.29}$	$0.38_{\pm 0.21}$	$\mathbf{0.16_{\pm 0.08}}$
FashionMNIST	$\mathit{1.56}_{\pm \mathit{0.39}}$	$3.57_{\pm 1.35}$	$1.65_{\pm 0.35}$	$\mathbf{0.82_{\pm 0.13}}$
Tiny	$\mathit{3.88}_{\pm \mathit{3.43}}$	$0.66_{\pm 0.20}$	$0.47_{\pm 0.07}$	$\mathbf{0.19_{\pm 0.09}}$
ImageNet	0.82	*1.49*	**0.58**	–

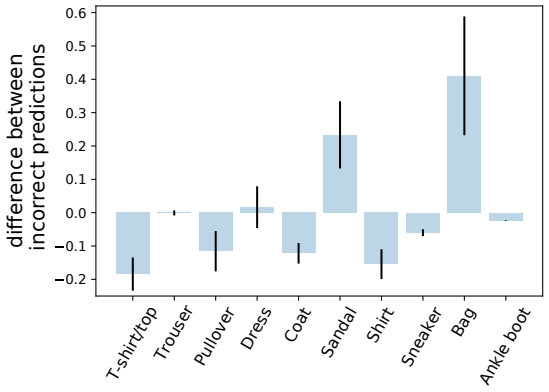
We remind the reader that a high DI value indicates that the model consistently associates distorted images with a particular class. Note that there is no fixed threshold that indicates data interference. Instead, the results are meant to be interpreted in a comparative manner. As such, we are interested in determining if there exists a clear gap between different models across the same task. If such a gap exists, then the model with higher DI will artificially appear to be more shape biased than a model with low DI index, as we will see later in this section. We compute the DI values for patch-shuffling images of the five standard data sets mentioned above. We present the results of this experiment in Table 3.1. We observe that for all the data sets we consider, such a gap exists, with the basic and MixUp models generally having high index values. The large index indicates that they tend to associate the features artificially introduced by patch-shuffling with a certain class.

We see the gap more clearly in the worst-case scenario captured by the alternative DI index (Table 3.2), where we consider the highest increase out of the five runs. In this case, the gap between the average index for the basic model and the CutMix model is more than 20% higher than that measured with the standard index defined in Equation 3.1. Although a clearer data interference can be observed when considering the highest increase across runs, it is enough to use the more restrictive definition to see that the introduced artefacts are affecting the model's predictions. Therefore, for the remainder of this chapter we will use the standard DI index.

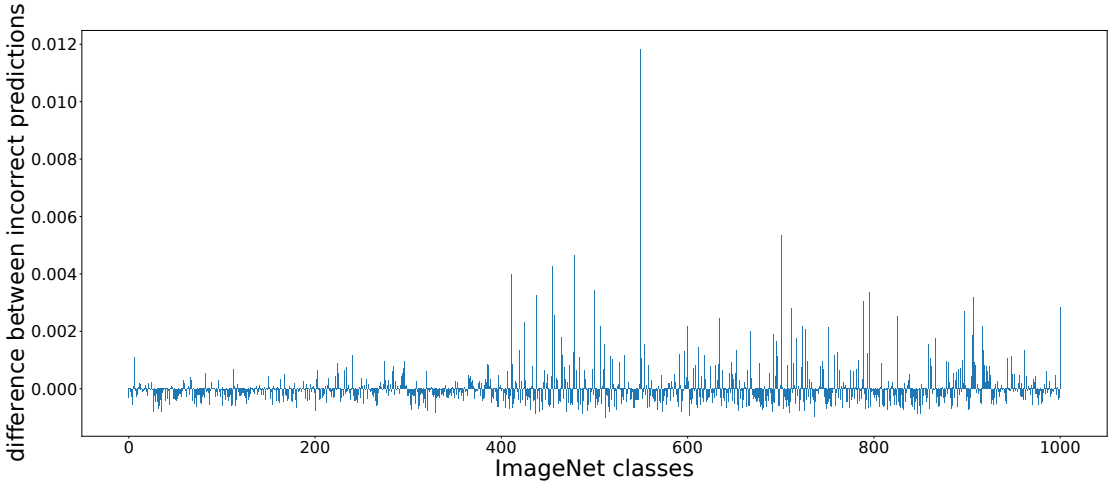
To better interpret the results, we take a closer look at the distribution of misclassifications for CIFAR-10 and notice that the basic model tends to wrongly predict the class



(A) Increase in incorrect predictions on CIFAR-10 for the basic model. Patch-shuffled images tend to be incorrectly associated with the class "Truck".



(B) Increase in incorrect predictions on Fashion MNIST for the MixUp model. Patch-shuffled images tend to be incorrectly associated with the class "Bag".



(C) Increase in incorrect predictions on ImageNet for the basic model. There is a sharp increase for class 550, "Espresso Maker".

FIGURE 3.2: Difference between the number of times a class was wrongly predicted when presented with original data samples compared to patch-shuffled data. For all data sets, we can observe a considerable increase for one of the classes, a class that is visually characterised by strong horizontal and vertical edges.

"Truck" (Figure 3.2a). This is not at all surprising, given that the strong horizontal and vertical edges are highly indicative of this class. Similar observations can be made for other data sets. For example, Fashion-MNIST models tend to identify shuffled images as belonging to the class "Bag" (Figure 3.2b), while on ImageNet they are associated with the "Espresso Maker" category (Figure 3.2c).

Given the observed tendency, we believe the grid-shuffling approach is causing models which are not invariant to strong horizontal and vertical edges to *falsely appear* to rely more heavily on shape information. A model not affected by this transformation could be considered texture-biased if we accept the larger definition of texture as local information. However, there is a question about the extent to which the reciprocal is true; a model could be invariant to the artificial edges because it relies on texture information or because it uses different shape-related features. Since patch-shuffling implicitly penalises models that are sensitive to artefacts, we question if this sensitivity

TABLE 3.3: Accuracy of ImageNet and Tiny ImageNet models on the GST data set when the label is taken to be either the shape or texture depicted in each sample. No clear correlation can be drawn between masking methods and low texture bias.

	ImageNet		Tiny ImageNet	
	Shape	Texture	Shape	Texture
basic	20.31	53.28	$10.56_{\pm 0.65}$	$26.04_{\pm 1.77}$
MixUp	24.14	60.31	$12.02_{\pm 0.33}$	$27.77_{\pm 1.56}$
FMix	21.25	53.43	$10.40_{\pm 0.39}$	$19.90_{\pm 2.12}$
CutMix	—	—	$10.54_{\pm 0.38}$	$23.72_{\pm 2.42}$

directly implies increased shape bias. Ultimately, we want to verify if models with similar shape bias can have different DI index values; if so, patch-shuffling would be an unfair basis for evaluation of the shape bias of such models.

Is a model necessarily more affected by patch-shuffling if it has a higher shape bias? We will show that the side effects of patch shuffling captured by our DI measure are not necessarily caused by a higher shape bias. To do this, we can use another method of determining shape and texture bias to find a counter-example. We analyse the ImageNet models on the Geirhos Style-Transfer (GST) (Geirhos et al., 2019) data set. The GST data set was specifically designed for shape and texture bias identification and it represents a gold standard in the field. The limitation associated with it is that it can only assess the bias of models trained on data sets with which it is compatible. For this reason, universal alternative methods such as patch-shuffling were proposed.

The GST data set contains artificially generated images where the shape belongs to one class and the texture to another. For example, an image could depict the shape of an elephant and the texture of a cat. There are 16 coarse classes that encompass a number of ImageNet categories to which they are mapped. The bias of the models is given by the accuracy obtained when the label is set to either the shape or texture information. Using this well-known method of identifying shape bias we want to find models which have similar biases but different DI indices when patch-shuffling. This would indicate that sensitivity to shuffling is not necessarily linked to increased shape bias, which in turn would mean that models evaluated using patch-shuffling can artificially appear more shape biased.

Table 3.3 gives the accuracy on the GST data set when setting the label to indicate first the shape, then the texture class. The results in Table 3.3 show that the basic model does not have a higher shape bias than models trained with masking augmentations, although it has a significantly higher DI index, as we have seen in Table 3.1. We repeat the same experiment on the Tiny ImageNet data set. Geirhos et al. (2019) use WordNet (Miller, 1995) to map the 1000 ImageNet categories to the 16 classes of the GST data set. We used the same method to create a mapping between Tiny ImageNet and GST. A number of ImageNet categories that belong to the 16 higher-level classes of GST

TABLE 3.4: Shape and texture accuracy of BagNet9 models versus the *DI* index on the GST data set. Compared to the masking-trained models, the basic model shows a clear bias when evaluating predictions on grid-shuffled data. Their shape and texture biases as measured with the GST approach, however, do not differ outside the margin of error.

	Shape	Texture	*DI*
basic	$11.29_{\pm 0.15}$	$18.90_{\pm 0.66}$	$1.16_{\pm}0.07$
MixUp	$11.04_{\pm 0.29}$	$12.56_{\pm 1.26}$	$0.89_{\pm}0.27$
FMix	$11.06_{\pm 0.48}$	$17.47_{\pm 1.74}$	$0.64_{\pm}0.10$
CutMix	$10.76_{\pm 0.27}$	$20.28_{\pm 0.88}$	$0.21_{\pm}0.07$

TABLE 3.5: Accuracy obtained on patch-shuffled images. A considerable gap can be noted between the basic model and mask-augmented ones although they exhibit a similar shape bias according to the GST method (Table 3.3).

	ImageNet	Tiny ImageNet
basic	49.49	$13.53_{\pm}2.02$
MixUp	52.16	$17.81_{\pm}0.16$
FMix	56.20	$34.53_{\pm}0.62$
CutMix	—	$42.30_{\pm}0.29$

are missing. For this reason, a poorer overall performance is expected and the results could differ slightly given a better fit between the sets. Nonetheless, we find again no significant correlation between masking augmentation and texture bias.

We also perform the same experiment for BagNet9 models (see Table 3.4). Note that we do not present results for ImageNet since no pretrained BagNet models were made publicly available. We consider this architecture since it has smaller receptive fields and so models are forced to use more local information. Even in this case, we find a high DI for the basic model and no difference in texture bias compared to MSDA. Thus, a model which is more affected by the side effects of patch-shuffling is not necessarily more shaped-bias. In other words, the artefacts introduced by patch-shuffling can cause models to have different accuracies on these distorted images, albeit not differing in their shape and texture bias.

We further confirm this conclusion using the GST approach on the two data sets with which this is compatible. Table 3.5 gives the accuracy obtained when patch-shuffling on Tiny ImageNet and ImageNet. In both cases, for comparable levels of shape and texture bias, different accuracies are obtained. This confirms that *models can appear to have vastly different shape bias when evaluated on randomly rearranged patches, when in reality their actual shape bias is similar*. The sensitivity of the patch shuffling approach to artefacts makes it an unfair and unreliable measure of shape bias and our *DI* index can help expose this limitation.

3.2.2 Occlusion Measurement

We next want to determine whether the same issue identified in the case of shape bias evaluation applies to occlusion robustness measures. We focus on CutOcclusion, where a rectangular black patch is superimposed on test images and the robustness is given by the resulting accuracy. We perform the same experiment as for shape bias identification, where we evaluate the *DI* index for four types of models trained on the five main

TABLE 3.6: DI index when occluding with black patches. The highest results are given in italic and the lowest in bold. For each data set, there exists a non-negligible gap in the DI index.

	basic	MixUp	FMix	CutMix
CIFAR-10	$\mathit{1.25}_{\pm \mathit{0.17}}$	$0.47_{\pm 0.11}$	$\mathbf{0.11_{\pm 0.04}}$	$\mathit{2.20}_{\pm \mathit{0.81}}$
CIFAR-100	$\mathit{1.24}_{\pm \mathit{0.35}}$	$0.34_{\pm 0.09}$	$\mathbf{0.12_{\pm 0.10}}$	$\mathit{1.06}_{\pm \mathit{0.32}}$
FashionMNIST	$0.21_{\pm 0.08}$	$\mathit{0.38}_{\pm \mathit{0.06}}$	$0.16_{\pm 0.05}$	$\mathbf{0.12_{\pm 0.05}}$
Tiny ImageNet	$\mathit{0.52}_{\pm \mathit{0.17}}$	$0.39_{\pm 0.03}$	$\mathbf{0.14_{\pm 0.04}}$	$3.46_{\pm 2.45}$
ImageNet	**0.50**	*1.50*	**0.50**	–

standard data sets we consider. The only difference is that the index is now measured when testing on rectangle-occluded images rather than patch-shuffled. We once again look for a gap in the *DI* index values across each data set.

There is no standardised distortion when measuring CutOcclusion, with the size and positioning of the obstructing patch varying between studies. Most often in prior art, a lack of robustness is noted for large occluders (e.g. Chun et al., 2020; Zhong et al., 2020*b*). For this reason, we uniformly sample the size of the patch from $[0.7, 1]$, allowing the occluding patch to lie outside the image (as it is done for augmenting with CutMix and CutOut (DeVries and Taylor, 2017)). This allows us to capture both the cases in which either the centre or the border area is masked out but requires a non-uniform distribution to counter for the patches existing outside the image. We also experiment with sampling from the interval $[0.1, 1]$ where the occluder is restricted to be positioned within the image boundaries and obtain similar results (See Section D.5).

Table 3.6 gives the result of measuring the *DI* index for occluding images with rectangular patches as described above. A significant gap in the DI index can be identified for each of the data sets. This indicates that some models will again be disadvantaged. Additionally, we find data interference to also occur when overlapping patches sampled from external images (Table D.2), using differently shaped masks (Table D.3), and for different architectures (Section D.4). This confirms that data interference is commonly occurring in a variety of settings. Thus, the result of CutOcclusion and its variants is highly dependent on the problem at hand. Just as for randomly shuffling tiles, by occluding images using a particularly shaped patch, one implicitly measures a model's affinity to certain features. In other words, models for which a distinctive feature is artificially introduced when performing CutOcclusion are going to be penalised by this method even though the feature is perfectly valid in a normal classification setting. This deems such methods inappropriate for fairly assessing robustness and texture bias.

A related observation was made by Hooker et al. (2019), who note the pitfalls of manipulating data to determine feature importance. They point out that when simply superimposing uniform patches over image features, it is difficult to assess how much of the reduction in accuracy is caused by the absence of those features and how much is due to images becoming out-of-distribution. To address this, the most important features

identified by an estimator are masked out both on train and test data, closing the gap between the two sets. Hooker et al. then train and evaluate models on the newly generated images. Unlike for interpretability methods, the subject of occlusion robustness studies is the model itself, which makes training with a modified version of the data an inviable option. In the following section we explore ways of overcoming this bias when measuring occlusion robustness.

3.3 What are Fairer Alternatives?

We have shown that the results of CutOcclusion depend on whether the artefacts interfere with the learnt representations or not. As we will illustrate in this section, another limitation of CutOcclusion is its sensitivity to the overall generalisation ability of the model. To better reflect how much information can be hidden from a model without affecting its performance, a fair alternative should be invariant to the characteristics of the occluder and to the model's goodness of fit. We propose a simple, more carefully defined measure that aims to decouple the machine's edge bias and generalisation ability from the occlusion robustness. We refer to our measure as "interplay occlusion" (iOcclusion). Interplay occlusion reflects the change in the interplay between performance on seen and unseen data. Formally,

$$iOcclusion_i = \left| \frac{\mathcal{A}(\mathcal{D}^i_{train}) - \mathcal{A}(\mathcal{D}^i_{test})}{\mathcal{A}(\mathcal{D}_{train}) - \mathcal{A}(\mathcal{D}_{test})} \right|, \tag{3.2}$$

where $\mathcal{A}(\mathcal{D})$ denotes the accuracy on a given data set $\mathcal{D}$, and $\mathcal{D}^i$ is the data set resulting from removing $i\%$ pixels of each image. The intuition is that on train data, robust models are less sensitive to the artefacts of the occlusion policy for small levels of occlusion, resulting in a large difference in accuracy from that on unseen data. Note that this is an implicit assumption that although we argue it makes intuitive sense, we cannot verify. We discuss this in more detail in Section 3.6. The performance of both train and test gets close to random as the percentage of occluded data approaches 90% and we expect the gap to fall off quicker for less robust models. This change in interplay is taken with respect to the generalisation gap of the model, such that the quality of the model fit in itself does not interfere with the robustness measure. We next discuss the methods we consider for masking out $i\%$ of the pixels.

3.3.1 Choosing a Masking Method

Although iOcclusion reduces data interference, other factors have to also be considered when choosing a masking method for computing $\mathcal{D}^i$, such as the number of contiguous components or the amount of salient information that is masked out. In this section we illustrate these points by comparing the results obtained with four different masking

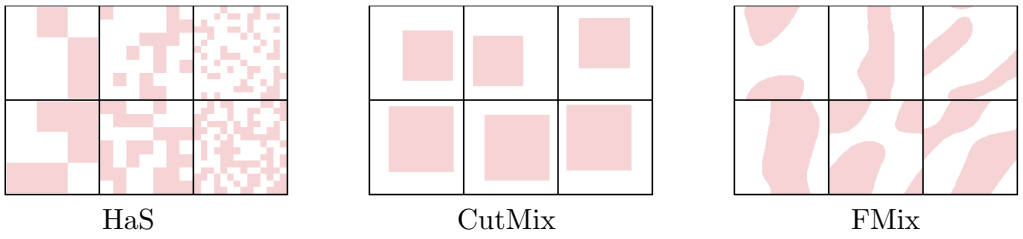


FIGURE 3.3: Examples of random masks used for ImageNet-sized samples (224×224 pixels) obtained with each of the three methods we consider. We refer to the masking approach based on the augmentation they were inspired from. For each method, the top row corresponds to masks that occlude 30% of the image pixels, while the bottom one provides examples of masks that occlude 50% of the pixels. For the HaS masking the grids used were, from left to right, 16×16, 32×32 and 64×64.

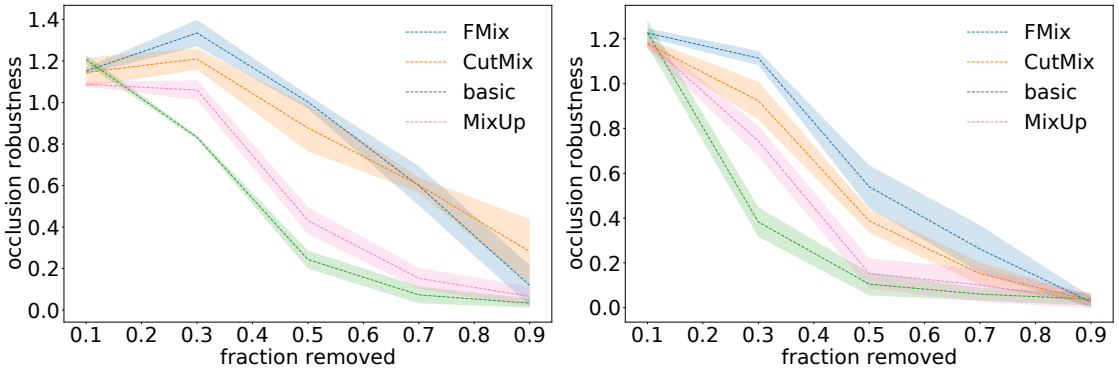


FIGURE 3.4: iOcclusion results when using an 8×8 (left) and 4×4 (right) grid mask. Models that were not trained with masking augmentations appear to be more robust for small levels of perturbations when the masking grid is more fine-grained.

methods. Three of these methods are inspired by the masking approaches in the FMix, CutMix and Hide-and-Seek (HaS) (Singh et al., 2018) augmentations. Similar to the patch-shuffling approach, HaS divides the image in a rectangular grid and masks out a fraction of the resulting tiles, selected at random. We give examples of masks obtained with the three approaches in Figure 3.3. The final masking approach we consider is based on Grad-CAM (Selvaraju et al., 2017). Grad-CAM uses the gradient information to identify which parts of the input have the largest impact on the final classification prediction.

Mask Granularity. The level of mask granularity has two important and closely-linked aspects: the number of occluders and whether it mimics a local or global distortion. We show the importance of granularity by masking as in the HaS augmentation. As mentioned earlier, HaS uses a lattice to mask out random rectangular regions in an image. Naturally, the more fragmented the lattice is, the more the distortion resembles noise addition rather than occlusion. In other words, a smaller number of contiguous regions measures resilience to *local* distortions, while a very large number of contiguous regions is closer to measuring resilience to a form of *global* perturbation.

We compute the iOcclusion results obtained when masking with an 8×8 and 4×4 grid respectively and give the results in Figure 3.4. We observe a difference in the curves obtained in these two regimes. The models appear more robust when the mask is more

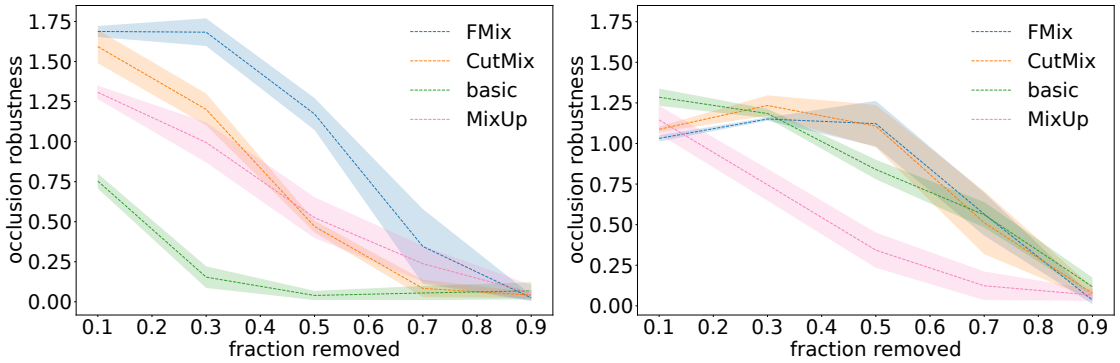


FIGURE 3.5: iOcclusion results when occluding a percentage of the pixels containing the most (left) and least (right) salient information. Note that the basic model is very robust to occluding the least salient information, while performing very poorly when as little as 10% of the most salient information is masked out. This indicates that the basic model classifies based very few key features.

fine-grained (8×8). This is particularly noticeable for the basic and MixUp models. For small fractions of occluded pixels, their robustness remains close to that of the FMix and CutMix models. That is because the distracting information is sufficiently diffused to be considered as negligible noise. For higher percentages of removed information the lack of robustness of the basic and MixUp models becomes clearer. A natural question is why is the robustness of MixUp models so low when the distortion is "spread" out across the image. An intuitive argument is that MixUp models are trained to recognise global patterns, which are now locally fragmented by this distortion. This interpretation will be reinforced later in this section. We believe the explanation lies in the fact that MixUp is a *global* distortion.

Additionally, it is important to notice that for broad-grained masks, FMix appears to be more robust than CutMix for small levels of occlusion. This is because CutMix models are trained with a single obstructor, whereas FMix masks force the model to learn more distributed representations. Thus, the mask granularity must be taken into account when assessing robustness.

Saliency. We next focus on illustrating the importance of accounting for information saliency when assessing robustness. For this, we use Grad-CAM, a model-dependent approach. For each image, Grad-CAM generates a heatmap that indicates the magnitude played by each image region on the model's prediction. We will refer to regions that lead to the highest Grad-CAM activations in an image as "most salient". Therefore, using the Grad-CAM method, we generate masks that occlude the most salient and the least salient information in the train and test images. It must be noted that the Grad-CAM masks are computed with respect to each of the models being evaluated rather than one reference model. This is because saliency is subjective and we want to occlude the information that is most or least relevant for each model in turn. The results we obtain are presented in Figure 3.5. We first observe that *the results we obtain when occluding the least salient information are significantly different from their counterparts.* This confirms that saliency is important when choosing a masking policy. The most notable

difference between the most and least salient regimes is for the basic model. For the case of occluding the most salient information, the basic model shows very poor robustness to occlusion, significantly below all other models. Conversely, when the least salient information is being masked out, the basic model appears to have a level of robustness comparable to the masking models.

Apart from observations relating to mask choice, this experiment also provides interesting results with respect to the level of distribution of the learnt representations for the models we consider. More specifically, it is interesting to observe that the basic model is heavily affected by occluding the most salient features, while being very robust to occluding less relevant pixels. This suggests that the basic model relies on a very small number of features and those features are crucial for the classification. On the opposite side of the spectrum, the FMix model appears to have a very "distributed" interpretation of the images.

Compared to the other models, FMix is very resilient to masking the most relevant pixels, while not showing outstanding robustness to occlusion of the least relevant information. This would suggest that FMix relies more heavily on contextual information, which is expected given the fragmented masks this model was trained with. These insights confirm that the difference between CutMix and FMix-like masking is largely explainable given the difference in masking granularity and expected saliency occlusion. It is also interesting to note that the behaviour of MixUp is fairly similar across the two scenarios. Our interpretation of this is that this stems from MixUp being a global distortion; as such, its perception of information saliency might be less localised.

What mask are we going to use? Since random masking makes the process noisier, we choose to generate masks using Grad-CAM when computing iOcclusion. Given that methods could be sensitive to either occluding contextual or core information, we get an average performance by choosing with equal probability between occluding the most or least salient i% pixels. It must be noted that this method implicitly assumes there could be multiple occluders and has the downside of being more computationally intensive. Another important observation is that Grad-CAM suffers from a number of limitations, including its inability to properly account for multiple object occurances or imprecise object localisation (e.g Chattopadhay et al., 2018; Omeiza et al., 2019; Belharbi et al., 2021). However, for the purpose of this study we are not interested in an exact saliency map, but rather a new randomly shaped masked. Although FMix already provides this to some extent, to ensure fairness we prefer to validate our approach on a masking method that has not been used for training purposes. We provide examples of masks obtained when occluding the least respectively most salient pixels in Figure 3.6. Note that for a fair comparison, throughout this section we do not allow the obstructing patch when measuring CutOcclusion to lie outside the image. As such, the fraction removed is exact and comparable to that of iOcclusion.

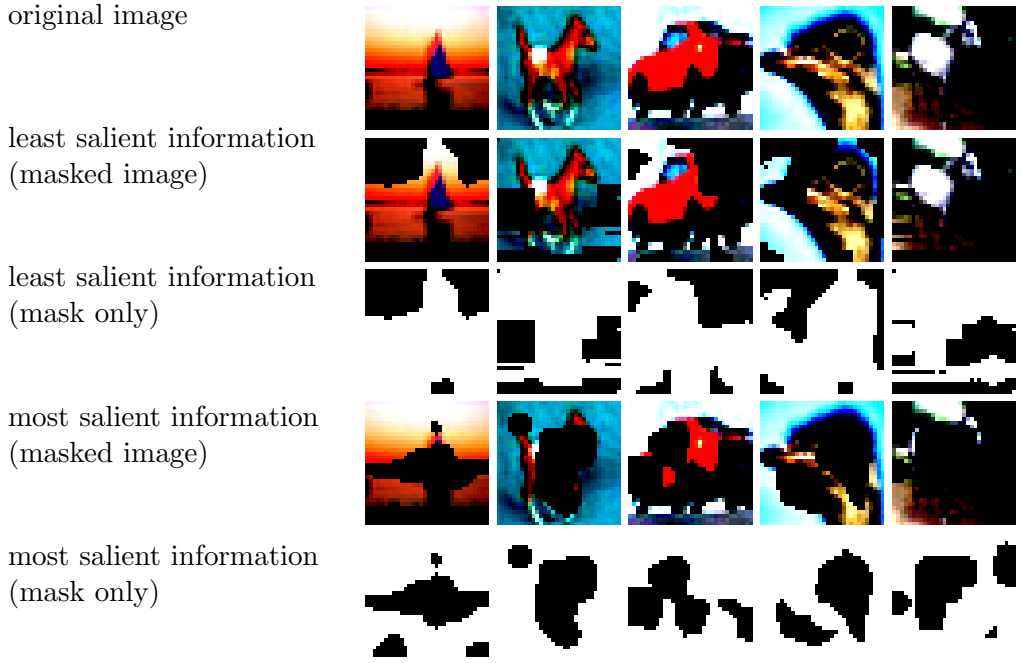


FIGURE 3.6: Examples of masks obtained on CIFAR-10 (32 × 32 pixels) for the basic model for an occlusion level of 30%. Here we are interested in observing the variety in mask shapes. According to Grad-CAM's notion of saliency, the basic model appears to pay little attention to contextual information. Although this latter observation is not relevant at this point in our study, it will play an important role in understanding the predictions of the basic model later on in this chapter.

3.3.2 iOcclusion Results

Assessing the correctness of such a measure is difficult in the absence of a baseline. For the remainder of this section we will build varied experiments to attest the validity of our method, highlighting important limitations of CutOcclusion that our approach addresses: sensitivity to the specifics of the occluding patch and sensitivity to the model's performance on unseen data.

Sensitivity to the Colour Pattern of the Occluding Patch

Since occlusion in real-life scenarios could be caused by non-uniformly coloured objects, an appropriate measure must generalise across colour patterns. We show that the results of CutOcclusion are sensitive to the colour pattern of the occluding patch, while iOcclusion is more invariant. To see this, we superimpose patches from images belonging to a different data set when computing iOcclusion and CutOcclusion, and compare the results to those obtained when occluding with black patches only. The reason why we choose to occlude with images from other data sets is to try to avoid some of the confounding factors. We present here the results for models trained on CIFAR-10. In this case, for evaluating robustness to non-uniform occluders we superimpose patches taken from CIFAR-100 on top of the CIFAR-10 images. The obtained occlusion curves are presented in Figures 3.7 and 3.8.

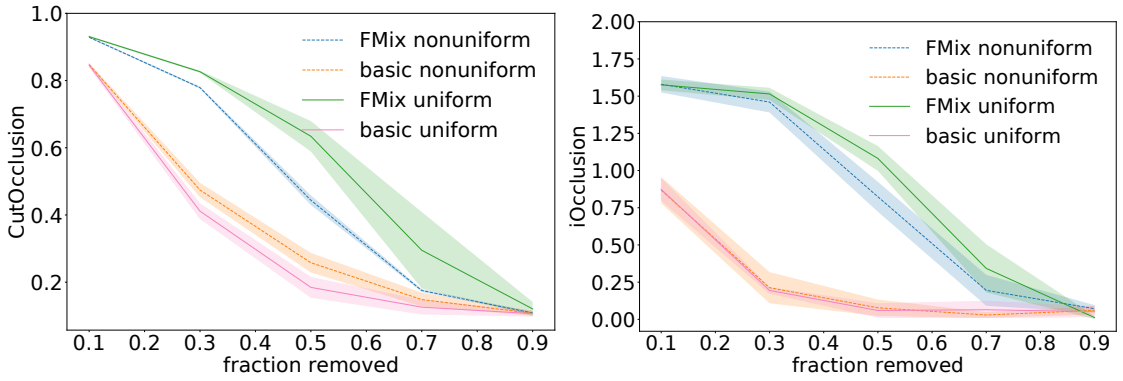


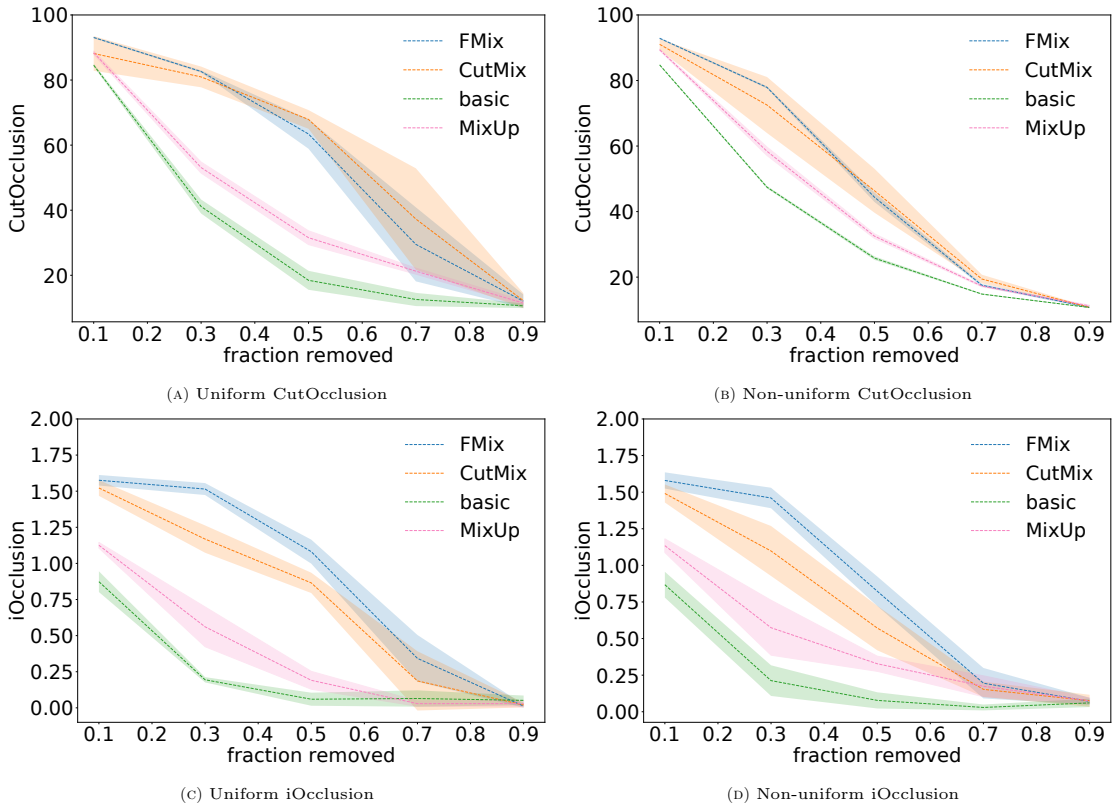
FIGURE 3.7: CutOcclusion (left) and iOcclusion (right) when occluding with black patches (uniform) and patches taken from other images (nonuniform). For both the FMix and the basic model, the curves obtained with iOcclusion in the two scenarios (i.e. uniform and nonuniform occlusion) are mostly overlapping. This corroborates the idea that iOcclusion is less sensitive to the information contained inside the occluding patch.

For visual clarity, Figure 3.7 only presents the results for the basic and FMix models, which are the least and most robust models respectively. Figure 3.8 gives a full comparison. The randomness introduced by the texture is naturally making the process noisier. Nonetheless, we find iOcclusion to better rule out the specifics of the occluding patch. For iOcclusion, using uniform occluders gives similar results to its non-uniform version, whereas the CutOcclusion measure provides an inconsistent model evaluation. In Figure 3.7 this can be seen by the mostly *overlapping curves* obtained for iOcclusion when occluding with uniform and non-uniform patches. In the case of CutOcclusion, however, there is a big gap between the results obtained with one type of patch and the other. In Figure 3.8 iOcclusion's consistency of results can be seen by the similarity in the curves obtained in the uniform (Figure 3.8c) and non-uniform cases (Figure 3.8d). On the other hand, the results obtained with CutOcclusion for the two cases (Figures 3.8a and 3.8b) are visually more distinct.

Sensitivity to the Patch Shape

As we have argued, in addition to not being sensitive to the colour pattern of the patch, a fair measure must also be invariant to the shape of the patch. To empirically confirm iOcclusion reduces the importance of edge information, we could compare the results obtained when masking with differently shaped masks. However, as we have seen in Section 3.3.1, there are multiple factors to account for and we are not aware of simple masking methods that account for all of them.

To avoid issues caused by confounding factors, we aim to compare models which exhibit different characteristics *using the same masking method*. For this, we want to obtain a model that is robust to occlusion, but at the same time has a high DI index (i.e. it is sensitive to edge information). To this end, we create a variation of FMix, Random Masks (RM), where at the beginning of the training process three masks are randomly



(A) Uniform CutOcclusion

(B) Non-uniform CutOcclusion

(C) Uniform iOcclusion

(D) Non-uniform iOcclusion

FIGURE 3.8: Left-hand side versus right-hand side. Comparison of metric sensitivity to textured occlusion. Uniform occlusion refers to superimposing black patches over CIFAR-10 images, while nonuniform refers to superimposing part of CIFAR-100 samples. Nonuniform CutOcclusion provides significantly different results to its regular counterpart, while iOcclusion remains consistent.

TABLE 3.7: DI index and occlusion robustness for models trained on CIFAR-10 when obstructing 30% of the image pixels with non-uniform patches. When measuring the robustness with CutOcclusion, RM appears significantly less robust than CutMix due to its sensitivity to patching with rectangles, while iOcclusion highlights the robustness specific to training with FMix-like masks. Given in bold is the closest mean result to that of RM for each evaluation.

	basic	MixUp	CutMix	FMix	RM
DI index	$1.67_{\pm 0.17}$	$0.98_{\pm 0.21}$	$0.14_{\pm 0.08}$	$0.15_{\pm 0.01}$	$0.39_{\pm 0.05}$
CutOcclusion	$47.97_{\pm 0.52}$	$\mathbf{58.65_{\pm 1.01}}$	$72.56_{\pm 8.55}$	$78.00_{\pm 0.45}$	$60.79_{\pm 5.03}$
iOcclusion	$0.21_{\pm 0.10}$	$0.57_{\pm 0.18}$	$\mathbf{1.09_{\pm 0.17}}$	$1.46_{\pm 0.07}$	$1.20_{\pm 0.23}$

sampled from Fourier space. For each batch, one of the three is chosen uniformly at random. We obtain a model that has higher DI index than FMix ($0.39_{\pm 0.05}$ compared to $0.15_{\pm 0.01}$), as desired. Table 3.7 gives the *DI* index, CutOcclusion, and iOcclusion results for the four main models we consider in this chapter, along with the RM model. For clarity, we only provide results for a 0.3 occluding fraction, although our observations hold for other fractions as well (see Figure E.1 of the Supplementary Material for the full results). Because CutOcclusion implicitly penalises models with high DI index, according to this measure, RM appears almost as sensitive to occlusion as MixUp. On the other hand, our measure reflects the robustness of training with RM, situating it close to the models trained with masking augmentations (FMix and CutMix).

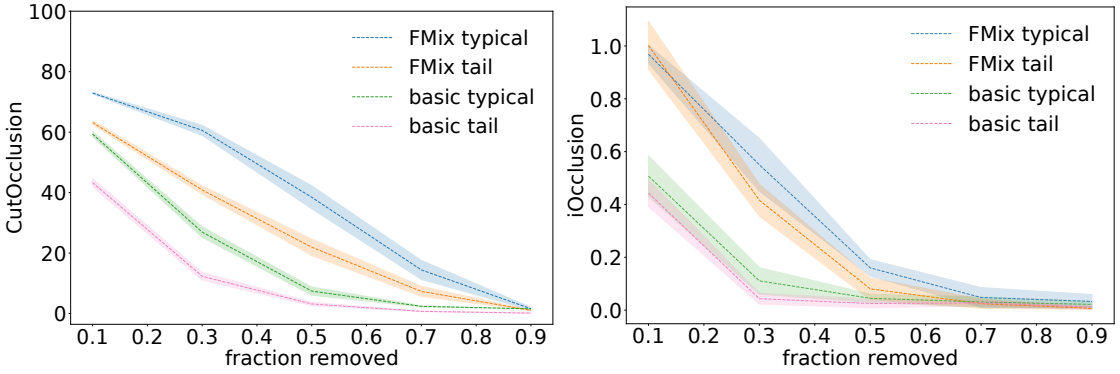


FIGURE 3.9: CutOcclusion (left) and iOcclusion (right) for the basic and FMix models on two subsets of the same data set: tail and typical. Evaluating the models with iOcclusion on the two types of samples leads to mostly overlapping robustness levels. That is, they do not differ outside the margin of error. On the contrary, CutOcclusion incorrectly finds the models to be less robust on tail data.

Sensitivity to the Model's Overall Performance

Another problem that occurs when purely looking at post-masking accuracy is weaker models would erroneously appear less robust. We show this by reversing the problem: we evaluate the same model on two different subsets of the CIFAR-100 data set: typical and tail images as categorised by Feldman and Zhang (2020). They consider a train-test sample pair to belong to the tail of the data distribution if the test sample is correctly classified when a model is trained with the train sample and incorrectly without it. Each tail example from the train set has a corresponding one in the test set.

Once again, for visual clarity we present results for the basic and FMix models alone. These are given in Figure 3.9. When evaluating robustness with iOcclusion, the robustness curves when evaluating models on typical and tail examples are largely overlapping, being within each other's margin of error. On the other hand, when evaluating robustness with CutOcclusion, there is a clear gap, with robustness on tail data falling off quicker than on the typical counterpart. Thus, CutOcclusion would indicate that models are significantly more robust to occluding typical examples. However, a closer analysis makes us doubt this conclusion. The raw accuracy on test data for tail examples is lower than for the typical ones. For example, for the basic model, the accuracy on original test data drops from $76.23_{\pm 0.69}$ on the typical subset to $67.46_{\pm 0.48}$ on the tail subset. In fact, the performance when occluding images decreases at the same *rate* for the two subsets, indicating similar robustness. The raw accuracy of the models is higher on typical examples, making it natural that the accuracy when occluding will also be higher on this subset. By way of definition, iOcclusion allows a fair comparison of robustness regardless of the overall performance of a model.

To further assess the sensitivity of CutOcclusion and iOcclusion to the overall performance of the model, we also experiment with randomising all the labels of the CIFAR-10 data set, just as in Zhang et al. (2017)'s label randomisation experiment presented in Chapter 2. We train a basic and an FMix model on this altered version of the data

TABLE 3.8: Robustness to occluding with patches covering 50% of each image. The models are trained with and without masking augmentation on data with randomised labels (referred to as "basic random" and "FMix random" respectively). For reference, results for FMix-training with original data ("FMix clean") are provided. iOcclusion captures the increased robustness to occlusion of the FMix model, while CutOcclusion makes no difference between regular and augmented training.

	basic random	FMix random	FMix clean
CutOcclusion	$10.24_{\pm 0.27}$	$9.78_{\pm 0.18}$	$63.63_{\pm 4.54}$
iOcclusion	$14.63_{\pm 1.12}$	$47.94_{\pm 19.84}$	$82.36_{\pm 10.06}$

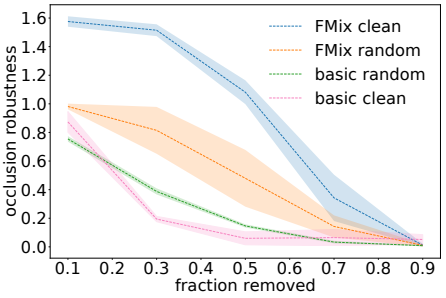


FIGURE 3.10: iOcclusion results for training basic and FMix models with randomly assigned labels (denoted "basic random" and "FMix random" respectively). For reference, results for FMix-training with original data are provided. iOcclusion captures the added robustness of training with random masks when the labels have been randomised.

using our standard training procedure. When evaluated on the randomly labeled training data, all the basic model runs achieve 100% accuracy, while the FMix model runs reach $99.99_{\pm 0.01}$%. Since all labels are corrupted, the accuracy on the test set before and after occlusion is no greater than random. However, the augmentation-trained model is robust to occlusion. This is captured by our metric. Figure 3.10 gives the robustness curve obtained for the basic and FMix models trained on original and randomised labels. As desired, on the randomised data, the FMix model has a higher robustness than the basic model.

On the other hand, CutOcclusion makes no distinction between learning with regular and augmented data. In Table 3.8 we give the CutOcclusion and iOcclusion results for the same models as in Figure 3.10. Namely, the basic and FMix models trained with randomised labels and the FMix model trained on the original data as reference. CutOcclusion incorrectly registers the same level of robustness for the basic and FMix models, while iOcclusion reflects the robustness gained by training with masking augmentation.

Despite being such a peculiar case, this experiment once again shows the comprehensiveness gained by accounting for the degradation on test data relative to the training data. We return to this experiment in Section 4.2.2 in the context of predicting generalisation based on a model's robustness to distortion on training data. In light of this experiment, in Section 4.2.2 we argue that the robustness of a model to a specific distortion *on training data* is not necessarily predictive of its generalisation performance. It is also very striking to observe that training with random labels makes the basic model more robust to occlusion. More precisely, for a 30% level of occlusion the accuracy for the basic model trained on original data drops to $0.37_{\pm 0.017}$, while for the one trained on random labels it only drops to $44.61_{\pm}0.01$. This shows that learning random labels does not necessarily entail "memorising", or in other words depending on every single pixel. Instead,

the model uses more contextual information. As with other experiments throughout the thesis, the network used for this experiment is a PreAct-ResNet18. Naturally, the observed phenomenon could be an artefact of the network architecture. Nonetheless, for the purpose of predicting generalisation, the observation remains relevant in the context of comparing instances of the same architecture. As we will discuss in Section 5.1, leveraging contextual information can help a model generalise in an i.i.d. setting. This raises questions of how one can determine a priori if the contextual information learnt by the model is beneficial or detrimental for generalisation. We will return to this question throughout the rest of the thesis.

Notes on iOcclusion

Alternative ways of measuring robustness to occlusion that have previously been considered in the literature imply cropping out objects from natural images and superimposing them on the data set of interest (e.g. Osherov and Lindenbaum, 2017; Zhu et al., 2019). This method incurs a high computational cost associated with cropping out objects and overlapping them on each data set we want to evaluate robustness on. Further, this approach does not address the limitations of CutOcclusion that we highlight in this chapter: sensitivity to the overall performance of the model and to potential data interference that could be caused by an insufficient variety in the types of objects considered.

As we evidenced through controlled experiments, there are many cases that CutOcclusion does not properly address. The unbiased nature of iOcclusion could lead to a better understanding and development of training procedures. It is equally important to note that it has applicability for real-world deployments where no prior knowledge exists about the possible shapes of the obstructions. While this aspect of generality is the strength of our approach, it must be stressed that when there exists a limited set of known possible occluders, evaluating robustness specifically to them could be safer. For example in an industrial setting or a clinical environment there could be a certain set of objects that could interfere with the subject of interest.

Incorrectly assessing robustness can have severe effects, especially when applied to sensitive applications such as autonomous vehicles or medical imaging. We do not propose a universal solution, but rather suggest an alternative to the biased approach for the common scenario in which the environment is not controlled and little is known about all the potential occluders. However, even in this case, our metric should be taken as a guide when analysing models. Although iOcclusion aims to address data interference, since a ground truth does not exist, it cannot be guaranteed that this method provides fair results in the absolute.

The strength of the bias will depend on the data in question and some applications will be more heavily affected than others. We have seen that for natural images this bias

TABLE 3.9: DI index measured for non-uniform occlusion when training without the class with the highest increase in incorrect predictions. Again, a gap can be noted, supporting the idea that data interference is not specific to peculiar cases.

	CIFAR-10	CIFAR-100
MixUp	$0.39_{\pm 0.15}$	$1.22_{\pm 0.19}$
FMix	$0.08_{\pm 0.06}$	$0.50_{\pm 0.21}$

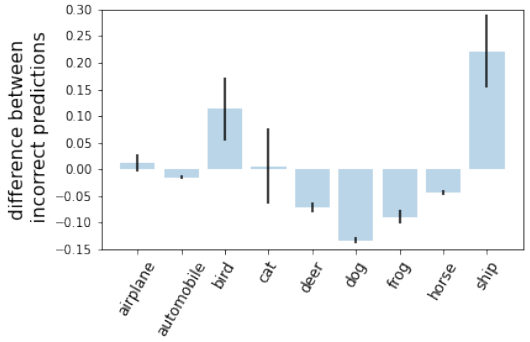


FIGURE 3.11: Difference in incorrect predictions for the basic model trained on a variant of the CIFAR-10 data set where the "Truck" class has been removed. A consistent bias exists for the "Ship" class.

does exist. To confirm that we have not just identified isolated cases, for CIFAR-10 and CIFAR-100 we remove the class that has the highest increase in mispredictions and retrain the models on the remaining classes. We give the index obtained for the MixUp and FMix models when occluding with rectangular patches in Table 3.9. We once again identify a non-negligible gap in the *DI* index for these models. Therefore, the bias is still present but with respect to another class. For example, in the case of CIFAR-10, after removing the "Truck" class, the basic model mispredicts rectangle-occluded images as "Ship" (see Figure 3.11). Thus, the edge artefacts are very likely to interfere with learnt representations since they are such fundamental features. From an evaluation perspective, as we have seen, this impacts assessment methods and must be accounted for. From the training perspective, such a widespread data interference of masking distortions would indicate a large perceptual shift in the data when performing MSDA. However, large perceptual shifts are believed to have a negative effect on generalisation, while MSDAs are known to improve performance. **Is then the perceptual shift caused by artificially introduced information really detrimental to learning?** In the following section we investigate the importance of the artefacts in this case and their implications.

3.4 Is the Magnitude of the Distribution Shift Important?

Note once again that in this thesis we are interested in data augmentation in the classical sense, where distortions are simply used to enlarge the data set. Augmentation has also been used in contrastive representation learning (e.g. Bai et al., 2022; Chen et al., 2020; He et al., 2020), but we are concerned with the scenario in which the model learns to perceive distorted and original data as belonging to the same distribution. In this case, it was traditionally believed that a good augmentation should have minimal distribution shift. However, increasingly many approaches propose heavy distortions (e.g. Yun et al., 2019; Summers and Dinneen, 2019), questioning the traditional view.

TABLE 3.10: Augmentation comparison on CIFAR-10. We consider two variants when calculating diversity. One is computing the cross-entropy loss using the label of the majority class (Diversity), as for mixing in Inoue (2018). The alternative, MixDiversity, takes a linear combination of the two cross-entropy losses.

	Affinity	Diversity	MixDiversity
MixUp	$\mathbf{-12.58_{\pm 0.14}}$	$\mathbf{0.41_{\pm 0.01}}$	$\mathbf{0.84_{\pm 0.00}}$
FMix	$-25.55_{\pm 0.26}$	$0.34_{\pm 0.01}$	$0.65_{\pm 0.00}$

Most recently, it has been argued that it is the degree of shift as *perceived by the model* that determines augmentation quality. Gontijo-Lopes et al. (2021) propose to measure this shift by the difference between the performance of the model when presented with original test data and augmented test data. They refer to this gap as "affinity". We show that *the magnitude of the distribution shift does not determine augmentation quality* of MSDA. We start with the perceptual gap of training with MSDA, as proposed in Gontijo-Lopes et al. (2021). We first argue that high affinity and high diversity are not necessarily desirable. Indeed, on CIFAR-10, we find FMix, a better performing augmentation, to have both lower affinity and lower diversity than MixUp (Table 3.10). We remind the reader that Gontijo-Lopes et al. define diversity as the loss on the augmented training data. For diversity, we compute the cross-entropy loss where the label is taken to be that of the majority class. Similar results are obtained with the MixUp loss, where a weighted average of the true labels is taken.

While intuitively for a high level of affinity, high diversity could correspond to better methods, the converse does not hold. We argue this is because affinity is rather an analysis of the learnt representations of the reference model and cannot give an insight into the quality of the augmentation or its effect on learning. The limitations of affinity are intimately linked to those of CutOcclusion. We have seen in Section 3.3 that the bias of the basic model is present not only when occluding an image with a uniform patch, but also when occluding with patches taken from other images, just as in mask-mixing. As such, an augmentation will have a lower affinity if it introduces artefacts that could otherwise lead to learning better representations when used in the training process. We believe this issue extends to other approaches that aim to motivate the success of MSDA through reduced distribution shift. Henceforth, we focus on bringing further supporting evidence that *the importance lies in the invariance introduced by the shift and its interaction with the given problem rather than its magnitude.*

3.4.1 If It Is Not the Magnitude That Matters, Is It the Direction?

We use empirical evidence to argue against previous assumptions behind the success of MSDA and propose the study of introduced bias as a more informative research direction. Here we use the term "bias" to refer to a drift in the learnt representations introduced by the change in the training procedure. A fundamental difference to classical training

TABLE 3.11: Accuracy on CIFAR-10 (left) and CIFAR-100 (right) upon mixing with samples from a different data set. The baseline is the accuracy when training with a single data set using the reformulated objective. In the interest of space, CIFAR-110 is used to refer to mixing with CIFAR-100 when training on the CIFAR-10 problem and vice-versa.

	MixUp	FMix	CutMix	MixUp	FMix	CutMix
baseline	$94.18_{\pm 0.34}$	$94.36_{\pm 0.28}$	$94.67_{\pm 0.20}$	$74.68_{\pm 0.37}$	$75.75_{\pm 0.31}$	$74.19_{\pm 0.50}$
CIFAR-110	$94.70_{\pm 0.27}$	$94.80_{\pm 0.32}$	$94.66_{\pm 0.12}$	$72.36_{\pm 1.04}$	$74.80_{\pm 0.55}$	$74.47_{\pm 0.39}$
Fashion	$92.28_{\pm 0.28}$	$95.03_{\pm 0.10}$	$94.61_{\pm 0.19}$	$66.40_{\pm 1.86}$	$74.46_{\pm 0.57}$	$74.06_{\pm 0.28}$

is that in the case of augmentation the samples are no longer independent. Mixed-sampling takes this even further. An immediate question is, does the added correlation lead to more meaningful representations? It is claimed that the strength of MixUp lies in causing the model to behave linearly between two images (Zhang et al., 2018a) or in pushing the examples towards their mean (Carratino et al., 2022). Both of these claims rely on the combined images to be generated from the same distribution. We perform inter-data set augmentation and show that good performance can be obtained even when the source images come from different distributions. The same inter-data set experiment further shows that by distorting the data distribution by a similar *magnitude*, we obtain two opposing results. This suggests that it is the *direction* of the introduced bias that is important for understanding the impact of augmentation. We next introduce the inter-data set experiment and present the results obtained.

It has been argued that label mixing has a negligible effect on the final model performance (Inoue, 2018; Huszár, 2017; Harris et al., 2020; Liang et al., 2018). We, therefore, use the reformulated objective setting (Huszár, 2017; Harris et al., 2020), where targets are not mixed and the mixing coefficient is drawn from an imbalanced Beta distribution. This allows us to apply MSDA between data sets. Thus, for training a model on a data set, we use an additional one whose targets will be ignored. As an example, a model that is learning to predict CIFAR-10 images will be trained on a combination of CIFAR-10 and CIFAR-100 images, with the target of the former. This scenario breaks the added correlation between training examples and shows that mixed augmentation does not necessarily rely on the source images to belong to the same distribution. Note that when mixing between data sets we use the same procedure as when performing regular MSDA.

We present the results for the inter-data set experiment in Table 3.11. For the purpose of predicting CIFAR-10 classes, performing MixUp with CIFAR-100 improves the average accuracy form 94.18 to 94.70. A non-trivial increase is also registered for FMix, both when mixing with CIFAR-100 as well as Fashion MNIST images, while CutMix maintains an accuracy similar to the baseline. Thus, this experiment shows that an accuracy similar to or better than that of regular MSDA *can* be obtained by performing inter-dataset MSDA. This invalidates the argument that the power of MixUp resides in causing the model to act linearly between samples, and calls for a broader explanation of its

success beyond that of pushing the examples towards their mean (Carratino et al., 2022). Another observation is that for FMix and MixUp, introducing elements from CIFAR-100 when training models on the CIFAR-10 problem does not harm the learning process. The reciprocal, however, does not hold. Hence, the "distribution shift" is more intimately linked to the problem at hand and aiming to characterise an augmentation based on the distance from the original distribution is a limiting approach, especially when the distance is measured as perceived by a reference model.

We believe an explanation is that the artefacts created when putting together images from CIFAR-10 with those of CIFAR-100 could introduce information that makes the separation of the 10 classes easier. However, if the information happens to interfere with a feature that is important for separating the CIFAR-100 categories, the performance could degrade on this data set. This singular experiment is not sufficient to draw any general conclusions. However, it does show that shifting two distributions by the same amount can have different effects on the model performance. Thus, *the direction of the introduced bias could be more important than its magnitude*. While some level of data similarity has to be preserved when performing MSDA, it is far from being the objective of such data-distorting approaches which, as we will argue further, should be rather seen as forms of regularisation.

3.4.1.1 Augmentation or Regularisation?

We have seen that for all considered data sets, artefacts introduced by masking methods seem to overlap with common features. This led us to believe that MSDA training could help bypass some of the simplicity bias. The simplicity bias refers to the tendency of deep models to find simple representations and, as we will discuss in Chapter 4, has been used to justify the success of deep models (Kalimeris et al., 2019; Valle-Perez et al., 2019*a*). Recent research shows that this propensity causes models to ignore complex features that explain the data well in favour of elementary features, even when they lead to worse performance (Shah et al., 2020; Hermann and Lampinen, 2020).

Although it could seem natural that since MSDAs are not augmentations in the VRM sense, they will increase the complexity of the problem, we design an experiment to support this claim. Similarly to Shah et al. (2020), we combine CIFAR-10 and MNIST (LeCun and Cortes, 2010) samples. Since they have the same number of classes, we can easily associate each class of one data set with a corresponding one from the other. Thus, we stack an image from the kth class of MNIST on top of a sample from the kth class of CIFAR-10. Since CIFAR-10 images are 32×32, while MNIST images are 28×28 pixels, we pad MNIST samples so as to match the dimension of the CIFAR-10 ones. Note that the stacking is done in the spatial, not channel dimension, such that a $3 \times 64 \times 32$ image is obtained.

We randomly combine the test images so as to break the correspondence between the CIFAR-10 and MNIST classes. We then separately compute the accuracy with respect to the targets of each data set. The predictions with respect to the CIFAR-10 labels are no better than random ($10.04_{\pm 0.11}$), while the accuracy with respect to the MNIST images remains high ($99.57_{\pm 0.72}$). Thus, models trained on this combination are mostly relying on MNIST images to make predictions. Similar behaviours have previously been associated with simplicity bias. Subsequently, when training, we perform FMix only on MNIST images and observe that this is enough to reverse the results. The models now rely on the CIFAR-10 images to make predictions. Evaluating against the CIFAR-10 label gives an accuracy of $86.60_{\pm 0.34}$, while testing against the MNIST label only gives $11.61_{\pm 0.30}$. We find that this also holds true for the other MSDAs. Thus, performing these distortions on the simpler data set increases its complexity to the point where it surpasses that of CIFAR-10.

Previously, we presented evidence that masking MSDA does not necessarily promote learning neither more shape nor more texture information. In light of this fact along with the results from this section, we believe image distortions force the model to learn more complex both shape and texture-specific features.

This chapter shows that a greater shift in learnt representations can lead to better models and simply quantifying the magnitude of the distribution shift can be misleading. An open question remains: How can we better capture the bias that is introduced and measure its quality? We believe understanding how a relatively small change in the data distribution impacts learnt representations could lead the way to characterising the relationship between data and model generalisation. However, this is a very complex problem in its own right. In the next section we look at how we can use insights gained from the present chapter to start addressing the relationship between learnt representations and generalisation.

3.5 How Does All This Relate to Generalisation?

As noted in passing in Directions in Generalisation: a Short Introduction, recent studies aiming to empirically estimate generalisation use data modification (Schiff et al., 2021; Natekar and Sharma, 2020; Lassance et al., 2020). The modifications used are preponderantly MixUp-like distortions, which are some of the first and most popular MSDAs. Using the insights gained throughout this chapter, in Section 4.2.2 we highlight the limitation of estimators that use MixUp-like distortions to capture generalisation performance at large. Finally, we discuss generalisation in relation to our findings. This leads to a number of questions that we hope will lead to better-motivated approaches to predicting generalisation.

FIGURE 3.12: Randomly picked MixUp samples obtained from ImageNet for a mixing coefficient of 0.5. For many examples it is difficult to identify the class of both source images and it can be argued that the MixUp samples do not look like natural images.

Why would models that behave linearly between training examples be better? – Interpolation adds structured noise. The idea of evaluating models on MixUp-distorted images for the purpose of predicting their generalisation ability would indirectly imply that models which behave linearly between training examples are better. However, humans would not perceive some of the interpolated samples as belonging to the distribution of natural images (Figure 3.12 provides more randomly selected examples). It is thus unclear why we should expect models that generalise better to necessarily represent the manifold of natural images in such a linear manner.

From a generalisation point of view, the most important observation we draw from the model analysis perspective is that distortion not only hides information, but also adds information. Seeing the above question through the lens of our observation allows us to form an initial intuition. Perhaps the linearity of the space more largely captures the model's ability to make predictions when both salient information is less perceptible *and* structured artificial information is introduced. This is more reflective of generalisation than simply introducing random noise.

We would argue, however, that linearity is not better in the absolute. Care must be taken when trying to predict the generalisation ability starting from it, as it has been proposed in the recent generalisation literature. We will expand on this idea in Chapter 4, where we will show that robustness to MixUp perturbation is not necessarily reflective of generalisation performance.

Local versus Global Distortion

As we alluded to in the beginning of this chapter, from the perspective of training with distorted images we note a difference between learning invariance to local or global distortions. The locality impacts how "distributed" the learnt representations are. We have seen that the fragmented masks of FMix cause models to be resilient to core information being masked out as they are forced to leverage more contextual information. But there is another interesting observation that can be drawn from masking out the least important information: the iOcclusion robustness applied to FMix and CutMix appears to be lower than that observed when masking out very small fractions of the

least salient information. Namely, for an occluding fraction of 0.1, for FMix and CutMix the iOcclusion is $1.57_{\pm 0.03}$ and $1.52_{\pm 0.05}$ respectively, while the one computed when masking out the *least* important information is $1.03_{\pm 0.01}$ and $1.08_{\pm 0.01}$ respectively. At a first glance, this seems contradictory. On a closer inspection, however, we notice that the CutMix and FMix models achieve very high training accuracy ($99.85_{\pm}0.04$ and $99.93_{\pm 0.04}$ respectively). Considering that CIFAR-10 was approximated to have around 3% mislabelled data (Pleiss et al., 2020), this situates them in a regime of minor overfitting. This is then reflected in a significantly lower drop on the training data versus test data when the least salient information is masked out. An important issue is thus raised: **Can a model be too robust on the training data? How do we differentiate between desirable and undesirable robustness in the absence of test data?**

We believe models *can* be overly robust on training data. If a model is robust to extreme levels of random occlusion, for example, when 90% of the image is occluded, then it must make predictions based on information that is in most cases not salient for the class. Therefore, it must learn spurious correlations in order to be more robust on training data. This observation is highly relevant to Chapter 4, where we explore robustness on training data as an indicator of generalisation performance.

Differences between local and global changes can also be seen in the intra-dataset mixing experiment (Table 3.11) introduced in Section 3.4.1. Namely, when the source distributions differ significantly (e.g. CIFAR-10 versus Fashion-MNIST), interpolating methods impact the statistics of the data set, leading to a decreased performance. This once again highlights that masking and interpolative distortions have *different effects* on the learning dynamics and the representations that are learnt. In the context of model training, Harris et al. (2020) argue that their effects are complementary. As we will see in Chapter 4, researchers are aiming to use robustness to distortions to empirically predict generalisation performance. However, we will argue in Section 4.2.2 that robustness to a *specific type* of distortion cannot indicate generalisation performance. **Can this difference between local and global distortions also be used to empirically capture generalisation through a more holistic notion of robustness?**

In summary, distortion for the purpose of predicting generalisation is a nascent field and this chapter brings to light a few directions that are worth exploring: the relationship between adding and subtracting information, between local and global distortions and the balance between too little and too much robustness. However, as we have shown in this chapter, distortions themselves are still not well understood. Therefore, as we will argue in Chapter 4, thorough evaluations of these empirical estimators are needed in order to ensure their correctness.

3.6 Future Work

This chapter represents a rich source for future research. Throughout the chapter we have already raised open-ended questions such as **"How is a small change in the data reflected in the learnt representations?", "How can we better capture the bias introduced by data modification?", "Where is the threshold between local and global distortions and how does their impact differ?"**, etc. Aside from these, we briefly discuss more questions that have surfaced while working with data modification, as well as limitations of the analyses and tools we propose.

Can we design a fairer alternative for measuring shape bias? During the review process for the International Conference on Machine Learning, the question of proposing a measure of shape bias based on patch shuffling was raised. It is important to point out that we do not believe patch shuffling provides a good starting point for evaluating shape-texture bias. This is because texture information is not necessarily equivalent to local information. The rectangles that result when dividing the image using a grid retain *local* information. This could indeed include texture information, but *edge* information is often preserved as well (e.g. a cat's pointy ears might be fully preserved in a patch). It is difficult to see how one could address this issue and construct a method for measuring shape bias starting from patch-shuffling. However, finding a fairer alternative remains an important objective since we believe interesting insights can be gained from studying the shape bias of different models.

Augmentation or Regularisation? An important objective for future work is more rigorously approaching the terms of "augmentation" and "regularisation" achieved through data modification. This distinction might be crucial when proposing a framework for reasoning about the effects of data modification. In this sense, one could think of "augmentation" in the VRM sense, where the distortions are expected to be "natural". That is, the distorted image appears as a natural image to a human. The intuition behind VRM is valuable, however, we do not have good definitions of sample vicinity. Considering only individual pixel values is limiting and does not reflect the complexity of the space of natural images. Nonetheless, assuming we had a good notion of vicinity, we could indeed define augmentation as a method of better capturing the natural space around existing samples. However, a natural distortion could be partially occluding the subject, which could in turn have a regularising effect. Therefore, it is unclear how one could describe natural distortions in a way that clearly separates regularisation from augmentation.

We believe another possible avenue for creating this distinction would be by defining a *data complexity* measure. As such, if a data modification increases the complexity of the data beyond a certain threshold, that modification can be considered to have a regularising effect. There are, however, also problems associated with this proposal. For example, the same type of modification could change the data complexity to different

extents when applied to different data sets. Moreover, we are as yet unsure how one would define data complexity in a formal way. We will return to this issue again in Section 5.1.

Immediate improvements to our work would be to address some of the limitations we have mentioned throughout the chapter. For example, iOcclusion relies on the assumption that models which are more robust to occlusion have a higher train-test gap on occluded images than their counterparts. Verifying this assumption boils down to decoupling the effects of adding and masking out information at the same time, as it is done when masking out regions of images. This leads back to the exact problem that we are trying to solve. **Would it be possible to propose a fair alternative for measuring robustness to occlusion that does not rely on additional assumptions?** While we believe that one always needs to make some assumptions, it is possible that future methods could start from weaker assumptions. However, the alternatives we have considered implied, in our opinion, stronger ones. Once again, it was not possible to verify these assumptions.

Similarly, we could aim to define a DI index that allows a more comprehensive comparison. In this chapter we were only interested in showing that data interference exists. Therefore, we have not considered constructing an index that allows cross-data set comparisons. We have, therefore, not accounted for the fact that the misclassifications for a 10-class problem are going to be less dispersed across classes than those of a 1000-class problem. Proposing a universal DI index could open new avenues for research where one could use this quantity to further study class-wise feature importance or even define a non-relative version of this index.

Lastly, an open problem remains building a framework that allows us to tell when one augmentation will be better than another and why. As mentioned in the introduction of this chapter, this problem is highly challenging. Indeed, we would argue that solving this problem might not be possible until we gain the same intuitions needed to solve generalisation itself. Nonetheless, it remains an avenue worth investigating and a promising source for gaining further practical insights.

3.7 Conclusions

Distorting data is such a commonplace procedure, yet little effort has been devoted to investigating its broader effects. This is particularly problematic when image modifications are applied in analyses. We show a number of cases in which this leads to *biased or incorrect results*. For occlusion robustness evaluation, we propose an alternative measure. The insights we gain from this endeavour point towards the study of data characteristics as a cornerstone of our understanding and raise a number of important questions about mixed sample data augmentation, on which we subsequently focus.

We note that MSDAs interfere with features that are consistently found across a number of data sets and conclude that the methods commonly used are forms of mixed sample *regularisation* rather than augmentation. A limitation of previous studies that aim to explain their success is the focus on trying to argue similarity with original data, rather than explaining the bias introduced by the distortion. Correctly interpreting it is important not only for making the models trustable but also for injecting more informed prior knowledge in future applications. Beyond their practical benefits, we believe MSDAs have the potential to help characterise the interplay between data and learnt representations. Overall, the purpose of this chapter is to encourage better practice when dealing with all forms of data distortions as well as to motivate their *principled* usage in generalisation studies.

Chapter 4

Steps Towards a Data-centric Evaluation of Empirical Predictors

This chapter lays the foundations of the future work needed to make meaningful contributions to the field of generalisation using empirical predictors. It covers a discussion of previously proposed predictors, a series of requirements for large-scale evaluation, and a proposal for a conceptual future direction.

In Chapter 2 we make evident the need to capture the attunement of the model to the given problem and its data. Our insight is that accounting for the data is crucial for this pursuit. We argue, however, that with the field's current level of understanding, it is difficult to propose a concrete and effective way of doing so. In line with recent developments, we believe a valid strategy is to return to practical experimentation to gain stronger insights into the phenomena observed in practice.

Since we began working on this thesis, a number of researchers started embracing the importance of practical experimentation for the purpose of capturing model performance. As such, a new form of studies has emerged: empirically predicting the generalisation performance of a learnt model. That is, having access to the learnt model, can we predict the generalisation performance a priori (i.e. before seeing the test data)? It is such predictors and their evaluation that this chapter is primarily concerned with. We will also refer to the generalisation predictors as *generalisation estimators* or simply *estimators*. Note that in the literature they are sometimes referred to as *generalisation measures*. We use the terms "predictor" and "estimator" since we believe they are more fit to generally describe all approaches that aim to predict generalisation, including the newer ones for which the term "measure" is less appropriate.

We have mentioned in the previous chapter that most recent empirical attempts have incorporated data distortion. Yet, as we have seen in Chapter 3, the field has a poor understanding of the mechanisms behind distorting data. We argued that when evaluating models on distorted data, one might obtain an unfair model comparison if the side effects (i.e. artefacts) of the data modification process are not taken into account. **Do the side effects we have identified earlier affect generalisation predictors in a similar way? Is the field's limited understanding of distortions also impacting empirical predictors? Can our findings so far be used to improve predictors?** These are some of the questions that gave birth to this chapter and which we aim to answer.

We believe that distortions for the purpose of predicting generalisation have been used in a poorly informed manner. Although sensitivity to distortions can signal an overfitting model, there is little foundation for the exact choice of data distortions in prior art. More specifically, it is unclear why the chosen data modifications (e.g. MixUp) would necessarily be correlated with a better generalising model. We argue this conceptually and also provide a small empirical example as initial evidence. However, in the Predicting Generalisation in Deep Learning competition (Jiang et al., 2021), a large-scale competition for empirically predicting generalisation, distortion-based estimators have been reported as successful, being adopted by the winning as well as runner-up solutions. **So why are reported results good when we argue distortion has been used inadequately?** We dedicate the second half of the chapter to answering this question. We argue that the setting in which the estimators were evaluated is too limited to expose the issues we underline. We believe this calls for a more critical examination of prior evaluation settings.

What scenarios have not been considered in previous evaluations? Chapters 2 and 3 have emphasised the need for a data-centric approach to understanding machine learning. With this perspective in mind, we note that most of the previously proposed empirical predictors have *not* been evaluated on variations of the training data. As we will detail in Section 4.2.2, variation could include removing specific subsets of the data set, changing the number of classes, or even distorting samples. We argue that considering settings where data is varied would reflect the limitation of state-of-the-art predictors and provide a more accurate picture of estimators' true ability to capture generalisation performance.

Are there other benefits of a data-centric evaluation? We believe the implications of a data-centric evaluation go far beyond exposing the limitations of distortion-based estimators. A good measure of generalisation should appropriately reflect changes in data, not only training procedure or architecture. We also argue that when models trained with different variations of the data set result in different generalisation performance, good generalisation measures should reflect this. Therefore, an evaluation that constructs a variety of data-focused scenarios would help ensure that the field does not

propose measures that can only explain hyperparameter changes without truly capturing generalisation performance at large.

Are empirical predictors of generalisation worth studying? In Section 4.2.2 we identify a number of challenges associated with pursuing this direction and we argue that in order to obtain truly fruitful results, a strong foundation must be established first. Clearly discussing implicit assumptions or pointing out the limitations of the empirical a posteriori perspective are some of the aspects missing from the literature. We argue that these are crucial for understanding what the realistic potential of this approach is and subsequently reaching it. Therefore, we believe there is value in studying empirical predictors but the field needs to systematically address the current limitations before impactful contributions can be made.

Our initial goal for this chapter was to provide a large-scale, data-centric evaluation of empirical estimators. Given the time constraints, this quickly turned out to be infeasible due to the many challenges that we will highlight throughout the following sections. This chapter, therefore, lays the ground for carrying out such a study. As mentioned earlier, we strongly believe this is important for a principled development of the field.

In the first part of this chapter we provide an extensive overview of directions in generalisation with a focus on empirical methods of generalisation estimation. To the best of our knowledge, we are the first ones to review the recent approaches that moved away from the complexity-centric view and the first ones to recompose the evolution of ideas in modern generalisation studies. With a large-scale study of empirical predictors in mind, we interweave the literature review with reflections and preliminary results for some of the methods. These are those methods which we would not incorporate in a future large-scale study. This is to avoid unnecessary computation by discarding directions that are not promising. However, we believe they should be discussed since studying the field's evolution gives us a more comprehensive view of the community's current take on generalisation and will hopefully help future researchers avoid unfruitful directions.

In the second part we clearly formulate the framework for predicting generalisation performance, motivating it while at the same time drawing the reader's attention to its limitations. We then focus on how empirical predictors have been evaluated in the past and highlight a number of omitted scenarios that expose issues with prior art. Consequently, we propose a list of desiderata for future evaluation. We finally motivate a conceptual direction for future estimators.

In summary, this chapter is concerned with empirical estimators of generalisation and their evaluation. Our contributions are:

- We present a succinct history of ideas in generalisation studies (Section 4.1.1);

- We review prior approaches in the empirical predictors of generalisation literature, highlighting limitations and inaccuracies (Section 4.1.1);
- For some estimators we provide preliminary evidence against their ability to capture generalisation at large (Sections 4.1.4 and 4.2.2);
- We review prior attempts to evaluate such generalisation estimators, discussing the practices they undertake (Section 4.2.1);
- Lastly, we propose a standard for such studies which accounts for the limitations of prior evaluations (Section 4.2.2). Note that constructing the evaluation setting itself is left as future work.

This chapter combines two different research directions: we provide an overview of perspectives on generalisation; and, we investigate the evaluation of empirical predictors. We will therefore discuss the relevant related work for these two research directions separately rather than having a unified section for related work.

Together with parts of Chapter 3, the contribution on Intrinsic Dimension from Section 4.1.1 was included in a workshop paper and presented as a lightning talk at the Data-Centric AI workshop at the NeurIPS 2021 Conference (Marcu and Prügel-Bennett, 2017).

4.1 Empirically Capturing Generalisation — an Overview

In Chapter 1 we have recounted that we can try to capture a model's behaviour either a priori or a posteriori. In this thesis we limit our exploration to methods that aim to characterise networks a priori. We remind the reader that there are two options when choosing the a priori direction: either estimating the generalisation performance without being able to provide guarantees, or provide guarantees typically by making strong assumptions. In this thesis we advocate for the estimation path. Our belief is that once we can successfully predict generalisation, we will be able to create bounds or notions of certainty that would guarantee performance depending on the given conditions. Our hope is that such a practice-based framework for providing guarantees would be relevant for real-world scenarios.

In this section we review the main ideas in the field of empirical generalisation studies. As a reminder, we consider an approach to be empirical if it moves away from the classical setting of statistical learning theory, which abstracts away from the problem and adopts a worst-case analysis. Due to the belief expressed above, we dedicate more time to discussing empirical predictors of generalisation which are at the heart of this study. These are those approaches that, given a specific instance of a model, aim to predict its generalisation performance.

The general review aims to present the *context* in which empirical predictors have emerged, while at the same time noting the more abstract concepts around which prior approaches are concentrated. The latter is particularly important for avoiding future redundancy, since, as we will see, highly similar approaches have been proposed across the different generalisation subfields. For completeness, we also briefly note some of the a posteriori directions but we do not provide a full review. For both a priori and a posteriori studies, we focus on the motivation behind the proposed approaches, highlighting strengths and limitations of the chosen perspective.

Related Work. As we mentioned in Chapter 1, there is no up-to-date, comprehensive overview of directions in generalisation studies, although reviews of some subfields exist. For example, surveys of classical statistical learning can be found in many textbooks covering generalisation (e.g. Hastie et al., 2009; Shalev-Shwartz and Ben-David, 2014). Similarly, the early advances in empirical studies are covered in depth by a number of reviews (e.g. Alquier, 2021; Mohri et al., 2018; Bartlett and Shawe-Taylor, 1999) and we will therefore only go into detail when it comes to the newer approaches. We will, however, mention the broad ideas in the early empirical studies and subsequently refer the reader to the existing extensive overviews. The broadest introduction to generalisation as a field is provided in the concurrent work of He and Tao (2020). However, He and Tao's study is concerned with bounding approaches only, reviewing improvements in tightness. He and Tao's focus is the precise results obtained, whereas our focus is to explain the setting considered and provide a history of ideas in the field. The two works are therefore complementary in nature. Another concurrent survey is that of Hu et al. (2021), who review the literature concerned with expressive power only. Like He and Tao (2020), they focus on specific results and less on the ideas being proposed and their contextualisation.

4.1.1 A Priori Estimation

A priori estimation relies on the implicit assumption that train and test data are both drawn from the same distribution. One important limitation of such methods is that in real-world applications it is very hard to guarantee that this assumption holds. In a posteriori settings, having a model-centric distribution distance measure can shed some light on the extent of the match between the two distributions, yet this is still prone to failure due to the ultimately subjective notion of distance between distributions, particularly in very high dimensional spaces. We will return to this discussion in Chapter 5.

Assuming that the training and test data are sufficiently similar, the focus falls on ensuring the model does not learn spurious features. The intuition often drawn from the bias-complexity formulation is that such spurious features are caused by an increased expressive power of the model class. As noted in the Directions in Generalisation: a

Short Introduction chapter, the ideas of capacity and complexity, which are measures of expressive power, are the cornerstones of classical generalisation studies.

Zhang et al. (2017)'s randomised labels experiment showed that the highly successful models used in modern machine learning have huge expressive power and therefore called for a rethinking of the field's approach to generalisation. Nonetheless, after Zhang et al.'s study, researchers still aimed to explain the generalisation mechanism working around the notion of expressive power, this time searching for new measures of complexity. We next briefly present the different interpretations of Zhang et al.'s experiment followed by an overview of the new ways of bounding or defining complexity.

Opposing Interpretations of the Label Randomisation Experiment

On the one hand, Zhang et al. (2017)'s work sparked the emergence of a new research direction in machine learning which quickly became a whole field: explaining generalisation in a highly overparametrised regime. This is usually studied in settings where training is carried out until zero loss is achieved, also referred to as the interpolative regime. In this direction, an interpretation is that the label randomisation experiment contradicts the bounds of statistical learning. As we have mentioned in Chapter 2, this is due to wrongly taking the classical bounds to mean that capacity needs to be restricted in order to have good generalisation performance. Another commonly accepted take is that the label randomisation experiment invalidates the complexity trade-off. This belief also occurs as a result of misinterpreting the Double Descent curve (Belkin et al., 2019) which we will discuss later in this chapter.

On the other hand, there are studies that reconcile the label randomisation experiment with classical wisdom. In this sense, studies like our own and Wilson and Izmailov (2020)'s argue that the generalisation performance depends on the "fitness", or attunement, of the architecture to the given problem such that an overwhelming number of hypotheses have a low risk.

An earlier complementary and popular explanation was that although deep learning models correspond to high-capacity hypothesis spaces, the SGD optimisers are biased towards choosing *low-complexity instances* (Neyshabur et al., 2014). Apart from the very strong optimisation component, this has once again brought forward the necessity of defining and capturing complexity. One of the most recent works on this topic provides a data-centric argument (Yang et al., 2022) which is in line with our intuition and that of Wilson and Izmailov. We will discuss Yang et al.'s work in more detail later in this chapter but for the moment we note that they use the structure of the data to motivate the reduced complexity of the learnt models. Thus, most a priori methods revolve in one way or another around the idea of complexity which stems from the classical framework.

These complexity-centric attempts have proven to be insufficient for capturing a model's ability to generalise. It is only most recently that researchers started to shift their attention to alternative ways of estimating generalisation performance. We argue that although not explicitly stated, many of them rely on concepts from information theory. We will next provide a concise review of expressive power-based results in empirical generalisation studies. We then succinctly introduce information theory to which we will relate the more recent research directions.

4.1.2 Measures Based on Expressive Power

We have seen in Chapter 2 that class capacity represents the cornerstone of classical studies. As Neyshabur et al. (2017) recount, the class capacity has previously been bounded by the number of network parameters (e.g. Bartlett et al., 1998c; Harvey et al., 2017). Note that the number of parameters is not an exact indicator of the network capacity (see Maddox et al. (2020) and Dwivedi et al. (2020), for example). However, capacity can be *bounded* using the number of network parameters.

As we will see, early attempts have been made to incorporate information about the data. However, the necessity of doing so became evident with the advent of deep learning. Bounds based on network size were provably tight and yet highly overparametrised models (e.g. Karpathy et al., 2014) were increasing in popularity due to their outstanding reported generalisation performance. Below, we review the attempts to incorporate information about the data both before and after the deep learning era.

The first step towards finding a task-dependent class complexity was the proposition of the Structural Risk Minimisation framework (Vapnik and Chervonenkis, 1971). Just like the classical statistical learning theory on which it is based, this framework is highly theoretical and has little practical relevance. Although important for an in-depth understanding of the techniques to capture generalisation, presenting it in detail does not ultimately add to the story told in this chapter. Its proposal, however, marks a shift in the generalisation studies narrative, with researchers now needing to go beyond the worst-case approach. In the bounding literature, this has largely led to an alternative framework termed PAC-Bayes (McAllester, 1999), and the emergence of new notions of expressive power that incorporate some information about the data, which we discuss next.

PAC-Bayes Bounds. A direction that aims to incorporate information about the data is represented by the PAC-Bayes literature. As mentioned in Chapter 1, this has been extensively covered and we will only briefly mention it for completeness. In this setting, one considers the output of the learning algorithm to be a probability distribution over the hypotheses in the class rather than a single learner. Notably, this distribution can depend on the training data. The tight results derived by Dziugaite and Roy (2017)

have led to a resurgence in interest for the PAC-Bayes framework. It is notable that, as Alquier (2021) and Banerjee and Montúfar (2021) point out, some of the recent PAC-Bayes bounds have strong connections to bounds based on the mutual information between the training sample and the learner. Banerjee and Montúfar provide a unified view of these approaches. For a comprehensive introduction to the PAC-Bayes setting and existing bounds, we direct the reader to the work of Alquier (2021).

Class Expressivity. We introduced the VC dimension in Chapter 2 as a notion of the expressive power of a model class. The data-aware alternatives to the VC dimension are the *Rademacher Complexity* and the *covering number* of the input space. The covering number gives the number of fixed-radius balls that can cover a space; in this case, the space considered is that given by the training data. Informally, the empirical Rademacher Complexity gives an indication of the ability of the functions in a hypothesis class to fit random noise. This quantity is "empirical" because just like the covering number, it is dependent on the training data. To obtain the Rademacher Complexity, one would compute the expected empirical Rademacher Complexity. For formal definitions, see Section A.

What has followed since the introduction of these measures of class expressivity was a great variety of attempts to *bound* them. More specifically, rather than guaranteeing performance based on the capacity of the entire hypothesis class, one could provide generalisation guarantees on a restricted subset of hypotheses. This subset can be constructed based on specific hypothesis-dependent quantities. The goal then became to define such quantities that would create informative bounds. Motivated by Zhang et al. (2017)'s randomised labels experiment, Neyshabur et al. (2017) popularised the idea that the purpose of such quantities must not only be to help bound generalisation but also to explain the performance of individual hypotheses. Therefore studies have recently started transitioning from a class-centric perspective to a more hypothesis-aware view. One of the earliest types of such bounds that aim to bound expressivity is given by margin bounds. Here *margin* is to be interpreted as the notion of the minimum distance between points in the training set and the decision boundary. This notion of margin is encapsulated in a model known as "Support Vector Machine" (Cortes and Vapnik, 1995).

Support Vector Machines represented a big step in the evolution of machine learning technologies. They were achieving high performance showing the over-pessimistic results of the bounds at the time. The fact that the learnt classifiers had good margins was questioning the relevance of the worst-case approach, calling for the need to include some notions relating to the data distribution. As a form of accounting for the data distribution, Bartlett and Shawe-Taylor (1999) expressed previously computed generalisation results for Support Vector Machines in terms of their margin. The margin was taken to be an indication of how peculiar the data distribution is. Subsequently, a number of margin-based bounds had been proposed, seminal works including those

of e.g. Shawe-Taylor et al. (1998*a*); Bartlett et al. (1998a); Bartlett (1998b); Mason et al. (2000); Evgeniou et al. (2000); Seeger et al. (2001); Koltchinskii and Panchenko (2002); Bartlett and Mendelson (2002); Kakade et al. (2008); Balcan and Berlind (2014); Kuznetsov et al. (2015); Neyshabur et al. (2015); Bartlett et al. (2017); Chuang et al. (2021) to name but a few. The studies cited above focus on the Rademacher complexity, while works such as that of Zhang (2002) or Ng (2004) bound the covering number. Examples of quantities considered in margin-based bounds include weight norm (e.g. Koltchinskii and Panchenko, 2002; Bartlett, 1998b), spectral norm (e.g. Bartlett et al., 2017), and path norm (e.g. Neyshabur et al., 2015). Another flavour of margin bounds can be found in the PAC-Bayesian literature (e.g. Langford and Shawe-Taylor, 2002; Herbrich and Graepel, 2002; Biggs and Guedj, 2022). For a textbook introduction to margin bounds we refer the reader to Mohri et al. (2018) and Anthony and Bartlett (1999), while for exact results for some of the above-cited bounds, we refer the reader to He and Tao (2020).

We briefly note that the "luckiness" framework (Shawe-Taylor et al., 1998*a*) was another attempt to provide data-dependent bounds and capture the success of SVMs. Informally, the intent was to use a function to rank hypotheses according to their level of compatibility with the training samples, as measured by notions such as margin. Algorithmic luckiness, however, has not seen as much popularity as the previously-mentioned alternatives due to the rather complicated technical details which make this framework less applicable (Foster et al., 2019).

It must be noted that within the margin literature cited above, one prominent idea was to stop the class expressivity from scaling with increasing network width and depth (e.g. Golowich et al., 2018). The aim was to have a notion of class complexity that could justify the performance of overparametrised models. As such, the objective was to bound the Rademacher complexity of overparametrised models in terms of notions that do not scale with the size of the network.

As briefly mentioned earlier, an alternative take was that the optimisation methods lead to implicit regularisation. For example, it was argued that the success of overparametrised models is given by the use of gradient descent methods, which maximise margins (Poggio et al., 2017; Soudry et al., 2018). Later, Valle-Perez et al. (2019*a*) theoretically argue the existence of a strong implicit bias towards low-complexity solutions and then derive a PAC-Bayes bound by looking at the input-output function space rather than parameter space, as in previous PAC-Bayes bounds. They empirically evaluate the bound, albeit on binary versions of benchmark vision data sets such as MNIST and CIFAR-10 and under the label randomisation setting only. As we will argue in Section 4.2, limited evaluation settings can result in misleading conclusions. Therefore, as with most of the methods we will discuss in this section, it is difficult to grasp how well their empirical evaluation reflects the true ability of the estimators to capture generalisation.

This idea of accounting for the learning algorithm dates back to the early works on algorithmic stability. Stability of the learning algorithm, associated with low variance in the bias-variance trade-off, measures how dependent a learnt hypothesis is on the given training set. There are a number of possible definitions for this. The core idea is measuring how much the loss changes when changing (or, more restrictively, removing) one of the training points. Just as with notions of expressivity, the randomisation experiment of Zhang et al. (2017) highlights the inability of stability measures to explain generalisation. Notice that there is an element of learnability captured by this notion. Bousquet and Elisseeff (2002) give a concise history of stability approaches, the origins of this field, and connections with the VC-centric perspective. For the reader interested in a more in-depth history of early ideas around margin bounds, PAC-Bayes, and connections with the stability literature, Boucheron et al. (2005) provide an extensive overview. In line with the belief that the optimization process is often biased towards low-complexity solutions, stability-based generalisation bounds have also been studied under Stochastic Gradient Descent (see e.g. Hardt et al., 2016; Kuzborskij and Lampert, 2018; Zhou et al., 2018; Lei and Ying, 2020).

Another field linked to stability that empirically bounds generalisation, but which has not become equally popular among theoreticians, is algorithmic robustness (Xu and Mannor, 2012). The quantity of interest for obtaining the bounds is the covering number of the input space, which in real-world scenarios is too vast for the bounds to be relevant (Neyshabur et al., 2017). Highly related to the stability perspective are also those measures centred around the properties of the loss landscape. The intuition for these methods goes back to early classical works and has mostly surfaced in the optimisation area of modern machine learning literature. Such examples are statistical physics studies of critical points of the energy surface in high dimensions (e.g Fyodorov and Williams, 2007) which subsequently inspired optimisation advancements (e.g Dauphin et al., 2014). Although the notion of flatness (or, alternatively, sharpness) of the minima was informally introduced by Hochreiter and Schmidhuber (1997), it was only through works such as Keskar et al. (2016)'s that flatness came back to the attention of generalisation researchers.

The ability of flatness to capture generalisation, however, has been contested (Dinh et al., 2017), as we will discuss later in this chapter. Subsequently, still centred around the idea of capacity reduction, Neyshabur et al. (2017) and Dziugaite and Roy (2017) concurrently note that sharpness must be coupled with some notion of weight norm to be able to control capacity. Neyshabur et al. (2017) goes on to empirically demonstrate the limitations of this joint measure as well. Aiming to address the limitations of previous flatness approaches, Liang et al. (2019) propose a norm-based solution that is invariant to the rescaling issue of prior flatness approaches. However, Liang et al.'s geometry-based approach to bounding complexity was later found to correlate poorly with generalisation performance (Jiang et al., 2020). Another attempt to overcome the

limitations of previous norm, margin, and sharpness approaches proposes a hidden unit-wise metric (Neyshabur et al., 2019). Neyshabur et al. (2019)'s metric weights the path norm of each hidden unit by its "impact" represented by the norm of the outgoing weight from that unit. They then use the sum of the unit-wise results to bound the Rademacher complexity of a model. This method was also found by Jiang et al. (2020) not to correlate well with generalisation. For more examples of such bounds as well as connections with PAC-Bayes and stability results, see the work of He and Tao (2020). For another view on the connection of flat minima to PAC-Bayes bounds and information-theoretic perspectives, see Achille and Soatto (2018).

Subsequently, Maddox et al. (2020) propose the eigenspectrum of the training loss Hessian as a notion of a model's *effective dimensionality*. They relate it to notions of compression and flatness and find that it can better correlate with generalisation performance when compared against path-norm bounds (Neyshabur et al., 2017), as well as the PAC-Bayes results of Jiang et al. (2020) and Dziugaite and Roy (2017). However, their experimental setting is very limited. They only vary the width and depth of convolutional networks and present experiments on the CIFAR-100 data set alone. Moreover, they do not compare their method against the most recent predictors.

Also relating to stability and algorithmic robustness is the belief that the *sensitivity* of a network *to perturbations* can be used to capture its complexity. Novak et al. (2018) measure robustness to perturbations by computing the norm of the Jacobian of the logits. They claim that this quantity, in conjunction with the number of linear regions in the network provides a good notion of complexity, which correlates well with the generalisation performance. Prior studies such as those of Montufar et al. (2014) and Raghu et al. (2017) have also used the number of linear regions to define complexity, and have noted the connection of this quantity with a network's stability to perturbations. However, Novak et al. (2018) were the first to evaluate the correlation with generalisation performance. Once again, their evaluation is limited and they do not compare against other empirical estimators.

Note that Novak et al. (2018)'s proposal deviates from the rest of the studies we discuss in this section from two points of view. Firstly, their objective is not to bound generalisation, but rather to find a notion of complexity that directly correlates with generalisation. This situates Novak et al.'s approach among empirical estimators of generalisation, which we will discuss in Section 4.1.5. Secondly, it computes the norm on the *test* data, therefore it is an *a posteriori* study. Although the a posteriori setting is outside the scope of this thesis, we will briefly mention a few such studies in Section 4.1.5. Despite being an a posteriori estimator, we believe Novak et al.'s quantity deserves to be mentioned in this section because it aims to directly capture expressive power.

Coming back to the notion of path norm, a concept that takes further the idea of characterising generalisation of a network based on the distance of the learnt weights from their initialisation is criticality. This was introduced by Zhang et al. (2022) as a layer-wise attribute. According to Zhang et al., a layer is considered critical if, after training, resetting the weights of that specific layer to their initial value affects the overall performance of the trained network. In their analysis they found a large number of layers to be robust to reinitialisation. Consequently, Chatterji et al. (2020) studied criticality at different levels and defined a complexity measure that they claim correlates well with generalisation but bring little supporting evidence.

Yang et al. (2022) propose a data-centric perspective. They computed the eigenspectrum of correlation matrices for a number of quantities considered in prior generalisation bounds. They noticed that the spectra of the considered quantities all follow a certain pattern which they argue is given by the structure of the data, whose eigenspectrum follows the same trend. The pattern is given by a small number of large eigenvalues followed by a large number of small eigenvalues which steadily decrease across orders of magnitude. Based on the eigenspectrum of the Hessian, Yang et al. propose to compute an effective dimensionality of a model, which is a very small fraction (less than 0.5%) of the weight count for the models they consider. Using this, they then propose numerical and analytical methods for computing tighter PAC-Bayes bounds than previously proposed. As such, Yang et al. (2022) argue that there exists a data-induced capacity control. Note, however, that Yang et al.'s view does not give a sense of when a specific model instance will be better than another and why.

Which Directions Relating to Expressivity Are Promising?

Most of the quantities we have discussed above were found not to correlate well with actual generalisation performance. Looking at some of the above propositions, Martin et al. (2021) categorise weight-based metrics into "shape" and "scale" metrics. They argue that the latter reflect broad changes in architecture such as changes in depth, while the former are more appropriate for understanding the relationship with hyper-parameters and the optimisation process. They note that neither of the two categories alone can explain generalisation. Moreover, Martin et al. (2021) observe a limitation of previous weight norm-based methods which they term *scale collapse* and reflects a drastic scale change in specific layers when the model is perturbed. Martin et al. also draw to the reader's attention that weight-based methods must be used with care when comparing architectures with a different number of parameters.

Spectral norm bounds as well as measures based on weight path (Neyshabur et al., 2017) or distance from initialisation (Nagarajan and Kolter, 2019*a*) have been found not to correlate well with generalisation (Jiang et al., 2020). The limitations of norm and eigendecay-based methods at large have been brought to the attention of the community

by various studies (Belkin et al., 2018; Nagarajan and Kolter, 2019*b*; Wei et al., 2022), with Wei et al. (2022) also arguing for the necessity of capturing a form of model-problem alignment.

Most similar in spirit to our calculations in Chapter 2 is the work of Wilson and Izmailov (2020). They propose a Bayesian perspective of generalisation where one marginalises over the set of possible hypotheses. They refer to the expressive power of the model class as the "model support". Akin to us, they argue that we would benefit from a larger support as long as there is an overwhelming number of good hypotheses. In their framework, they refer to attunement as the model's "inductive bias". Wilson and Izmailov (2020) note that PAC-Bayes bounds worsen for an increase in the number of parameters and reward reduced model class expressivity. This is in contrast with what one can observe in practice, where models with a vast number of parameters outperform their smaller counterparts.

A recent study by Jiang et al. (2020) compared a number of these methods and more, concluding that sharpness and optimisation-based methods represent good directions for future studies, while norm-based methods can sometimes negatively correlate with generalisation. However, sharpness methods have most recently been questioned (Dinh et al., 2017). The biggest limitation of sharpness-based approaches is the definition of sharpness itself. As a proxy for flatness of minima, different noise stability measurements have emerged (Langford and Caruana, 2002; Arora et al., 2018; Nagarajan and Kolter, 2019*c*) with more recent adaptations such as that of Morwani et al. (2020). But Dinh et al. (2017) shows that the flatness of minima can be greatly altered by simple reparametrisations, calling for more principled definitions than the ones previously proposed. Interestingly, Arora et al. (2018) link the notion of noise stability to network compression, which they in turn use to guarantee generalisation.

The idea of compression has seen many faces in machine learning; from model compression to manifold compression, it is generally presented as a desirable attribute. For an overview of the former, we refer the reader to the work of Neill (2020), while we will focus here on compression of learnt representations. More precisely, we will dispel the central role of compression starting from studies arguing that lower intrinsic dimension of learnt representations necessarily means better generalisation. To better grasp why compression in itself is insufficient to capture generalisation, we first introduce the Information Bottleneck theory (Tishby and Zaslavsky, 2015) which will help us build an intuitive argument.

4.1.3 Information Bottleneck

The Information Bottleneck theory is an alternative line of work that studies generalisation of deep learning models and is rooted in information theory. Although disputed (Goldfeld et al., 2019), we believe the intuition behind this theory provides a fresh and valuable perspective. However, just like statistical learning theory, the framework's generality impedes it from providing any concrete practical insights. For this reason, in the thesis we will focus on the general argument, rather than going into the technical details of this paradigm.

The Information Bottleneck principle highlights the underlying notion of information "relevance" in the information theory framework (Tishby et al., 2000). Informally, the relevance of a variable X can be established *in relation to* another variable Y. To illustrate this concept, Tishby et al. provide speech compression as an example. Compressing the speech audio signal requires significantly more bits of information than a transcript of the speech, which would completely capture its meaning. The relevance of waveform details such as the pitch is entirely dictated by the purpose of the speech compression. That is, if the purpose of the compression algorithm is to simply retain the message, then waveform details are irrelevant. If the purpose is to compress in a way that preserves the voice inflections, then the details become important. We will refer back to this idea later on in this section, clarifying it further in the context of machine learning.

Under the Information Bottleneck perspective, the goal of learning becomes to find a compressed representation of X that preserves as much information about the target variable Y as possible. This simple sentence summarises the Information Bottleneck principle, through which Tishby and Zaslavsky (2015) propose analysing deep learning. Tishby and Zaslavsky argue that *the objective of deep learning is to find a trade-off between compressing the model's internal representation and its predictive capabilities*. Viewing the successive intermediate representations of a model as a Markov chain, Tishby and Zaslavsky give a layer-wise measure of optimality, where each layer can be compared to the optimal Information Bottleneck limit. However, this limit is dependent on the true data distribution, which in real-world applications is not known. Although there exist empirical estimators of mutual information to account for this, they quickly become intractable when applied to real-world problems.

Ultimately, Tishby and Zaslavsky (2015) go back to a traditional perspective and abstractly argue for a trade-off between the generalisation gap and the complexity gap, with no concrete measure of generalisation performance. From the perspective of this thesis, the important takeaway from their study is the conceptual formulation of the Information Bottleneck goal, which we will further use for either disproving or motivating the success of some of the newly-proposed methods. Finally, we would like to point out that instead of aiming to empirically estimate the mutual information itself, a more

tractable, albeit less formal, solution could be to find attributes that simply indicate a high or low mutual information. This will play an important role in the final part of the thesis.

An important observation to make is that Goldfeld et al. (2019) have criticised Tishby and Zaslavsky (2015)'s binning approach to estimating mutual information, claiming that the observed compression phase is in fact a result of geometrical class clustering, which is the true quantity of interest in Goldfeld et al.'s view. They, however, leave the study of the relationship of class clustering with generalisation for future work. Since estimating the mutual information for the types of models we are interested in is infeasible, it is less relevant to our study whether binning is or not the right way to measure mutual information. Instead, we argue that clustering implies a level of compression *coupled with knowledge about the target variable*. Thus, the conceptual argument, which is the part our thesis is concerned with, remains valid in both views. We will next use this conceptual argument to support our belief that lower intrinsic dimension of learnt representations (equivalent to higher compression in information theory) alone cannot be reflective of generalisation performance.

4.1.4 Intrinsic Dimension

Another direction in the generalisation estimation literature is centred around the notion of the intrinsic dimension of learnt representations. We will briefly introduce this notion and unlike with the other approaches discussed so far, we will construct an empirical counter-argument to the previously proposed method of Ansuini et al. (2019). We do so because the Information Bottleneck perspective allows us to easily construct a case to illustrate the limitation of this approach. Secondly, as we will discuss in the latter part of this chapter, the notion of intrinsic dimension is a promising one and could provide a strong base for future work.

It is commonly believed that data lies on a low-dimensional manifold of a high-dimensional space (e.g. Pearson, 1901; Brand, 2002). The manifold's dimension reflects the minimum number of variables required to describe the true data. In practice, we do not have access to the true, low-dimensional data manifold. We can instead try to estimate it in the representational space. Note that the manifold dimension is dependent on the task at hand. For example, for reconstruction we might need more information than for classification. Therefore, the same original data could have a lower effective dimensionality in the latter case. Although this observation does not play a role in the present discussion, we believe it is important for constructing an understanding of the general notion of intrinsic dimension of the data. Note once again that in this part of the thesis we focus on classification tasks alone.

While it is difficult to know the true Intrinsic Dimension (ID), a number of estimation methods have been proposed (e.g. Granata and Carnevale, 2016; Duan and Dunson, 2021; Facco et al., 2017; Denti et al., 2021). Given a good model, we can estimate the ID based on its representations by measuring how much its embedding space can be "compressed". This is dependent on the quality of both the model and the estimator. Such estimators of intrinsic dimension have been used to analyse the relationship between local ID and adversarial robustness (Amsaleg et al., 2017; Ma et al., 2018). The connection between low ID and robustness that was discovered has motivated Ansuini et al. (2019) to investigate if ID can predict generalisation performance.

Based on the belief that the lower the dimension of the representation manifold, the easier it is to generalise, Ansuini et al. (2019) propose a training-time generalisation estimate. Specifically, using the TWO-NN (Facco et al., 2017) algorithm, they estimate the global ID of the last hidden layer manifold based on its representations of the training data. The TWO-NN algorithm computes the ratio of the distances to the closest two neighbours of each data point. Ansuini et al. claim that the generalisation performance can be predicted based on this quantity. As such, better performance should correspond to a lower ID value.

Is lower ID the driving factor behind better learners? We show through a counter-example that higher ID representations can lead to better generalisation performance, thus disproving the above correlation. We train the same model architecture on different versions of the training data so as to obtain representations with different ID and then compute the estimate of the obtained representation dimension. Following Ansuini et al. (2019), we use the TWO-NN estimator introduced by Facco et al. (2017) to approximate manifold dimension. We then argue that models with lower ID of learnt representations do not necessarily have a better generalisation performance than their high-dimension counterparts. Therefore, the ID is insufficient to determine a model's generalisation performance. To give an intuition for this, we can go back to the Information Bottleneck argument. In the Information Bottleneck framework, the idea is to find a compressed representation of the input variable X *while preserving as much information about the output variable Y as possible*. The ID perspective focuses on the former, without accounting for the latter. Although in a practical setting when training we could try to minimise the intrinsic dimension while maintaining good predictions, we do not have any notion of performance *outside* the training set. Thus, by simply minimising ID we cannot tell whether or not we are discarding information that is predictive of the *true* output variable Y.

To construct our counter-example, we once again make use of MixUp's reformulation (see Section 3.4.1). Ignoring one of the targets when mixing inputs is expected to create a data set where the instances can be represented in a more compressed manifold, decreasing separability at the same time. We train a VGG-16 network on the CIFAR-10/100 data sets using no mixed data augmentation (basic), original MixUp augmentation, and

TABLE 4.1: ID and accuracy of VGG models trained with MixUp, Reformulated MixUp and without mixed augmentation (basic) on CIFAR-10 and CIFAR-100. The basic and Reformulated MixUp models can have worse test performance but lower ID than the MixUp model. This provides a simple counter-example to the argument that lower ID necessarily gives better generalisation performance.

	CIFAR-10		CIFAR-100	
	ID	Accuracy	ID	Accuracy
basic	$7.80_{\pm 17}$	$93.04_{\pm 0.17}$	$12.18_{\pm 1.30}$	$71.70_{\pm 0.37}$
MixUp	$\mathbf{9.14_{\pm 0.31}}$	$\mathbf{93.79_{\pm 0.18}}$	$\mathbf{14.11_{\pm 1.31}}$	$\mathbf{72.60_{\pm 0.63}}$
Reformulated MixUp	$7.80_{\pm 16}$	$92.40_{\pm 0.34}$	$10.71_{\pm 0.21}$	$69.00_{\pm 0.41}$

reformulated MixUp augmentation, obtaining three different types of models. Table 4.1 shows the test accuracy and the estimated ID we obtain for these models. Note that although the models were trained differently, the ID is estimated on the *original* training data so as to ensure fairness. The MixUp model has the highest test accuracy while having a significantly higher ID compared to the Reformulated MixUp model. This directly contradicts the idea that a model with minimum ID of learnt representations is necessarily better.

One question that is immediately raised is if our counter-example would still hold given a more accurate method of capturing manifold dimension. We argue that even with further estimator refinements, the hypothesis that generalisation performance can be predicted based on learnt representation's ID lacks a strong basis, and it is unlikely to hold *in practice*.

Caveat. Does our argument invalidate all approaches that are solely based on compression or complexity? No. Theoretical studies centred around these notions are valid within the setting they address. When assuming that training and test data are drawn from the same distribution, the amount of information the model has about the true variable Y is implicitly determined by the level of compression in the intermediate representation. In practical settings, however, mislabelled samples or poor data collection practices cause this assumption to almost never hold. As we have seen in this section, using these quantities outside of the setting which they were constructed for leads to incorrect generalisation predictors.

4.1.5 Qualities of Learnt Representations

An idea that is becoming more popular among generalisation studies is to move away from notions of class complexity and focus on qualities of learnt representations. That is, study a fixed model instance rather than the model class. While we strongly agree with Dziugaite et al. (2020) that a good theory of generalisation should ultimately be able to abstract away from all details of the learning problem, we have strongly embraced the recent beliefs in the field independently of the concurrently published research. This is because we believe that the intuitions we have as a community are as yet insufficient

to allow us to build a comprehensive framework for understanding generalisation. Our belief is that closely studying the relationship between learnt representations and generalisation ability would allow us to build a principled understanding that when refined, could represent the starting point of a meaningful theory of generalisation.

Researchers have not yet managed to identify those qualities that capture generalisation performance. We will next review the ideas that have been explored so far. As we will see, many approaches focus on notions of representation geometry or robustness. Most of them informally relate to some of the ideas discussed in Section 4.1.2 and merely propose different ways of defining or capturing them in practice. In essence, compression and stability seem to be the desired qualities.

Many good notions of learnt representations quality have emerged during the "Predicting Generalization in Deep Learning" competition (Jiang et al., 2021). Most of the reported methods directly or indirectly propose measures of learnt representation quality. We focus on the top three methods according to the final ranking of the competition and note in passing the solutions of other participants that published a report of their approach.

Henceforth, we will use the term "estimator" or "predictor" for methods that are used to estimate or evaluate the generalisation performance of a model. Here and in the following section we present a number of such methods. These generally lack a strong theoretical justification and have largely been supported through empirical evaluation. We focus for now on the proposed estimators and discuss the evaluation settings in the second part of this chapter. As a general observation, it must be pointed out that the evaluation settings only consider limited scenarios. This, coupled with a lack of a standard evaluation setting and baseline, make it difficult to put the reported results in perspective.

Since our initial goal was to provide a large-scale study and comparison of empirical estimators, we carried out initial experiments with the first four methods outlined below. For this, we have implemented or reimplemented them in PyTorch where a Tensorflow version was publicly available. Where the code was not provided and the report did not clearly specify the details of the approach, we briefly discuss the design decisions we have made. As we will outline in the Future Work section of this chapter, these methods will be included in a future study.

- The winning solution of Natekar and Sharma (2020) takes the weighted sum of two scores: 1. the within-cluster to between-cluster ratio for learnt representations also known as the Davies-Bouldin Index (DB Index) (Davies and Bouldin, 1979) in the clustering literature; 2. the accuracy after performing MixUp between same-class examples. Through the DB Index, Natekar and Sharma, aim to quantify representation consistency. They compute all their measures on the first layer representations and do not report on results computed on the other layers.

Nonetheless, they mention that doing so leads to weaker results for the purpose of the competition. Another quality of learnt representations they consider but do not include is class separability as measured by an approximate distribution of margins. They argue that margins computed on MixUp samples are more predictive of generalisation than those computed on original data alone. In Section 4.2.2 we will come back to this method and argue that it cannot capture generalisation at large. Just as the other distortion-based approaches, it relies on intuitions that do not necessarily hold outside of the competition's evaluation setting.

- Similar to the winning solution, Kashyap et al. (2021) secure the second place by applying distortions to the training samples. A method is thus penalised if it changes the prediction when presented with altered data as opposed to the original images. The image manipulations they consider are random centre crop, flip, saturation change, Sobel filter application, Virtual Adversarial Perturbation (Miyato et al., 2018) and Random Erase (Zhong et al., 2020*b*) – a rectangular version of the CutOut augmentation.

- The third place approach, called *label variation of penultimate layer with MixUp* (VPM) (Lassance et al., 2020), couples the separability of learnt representations with robustness to MixUp interpolation. Lassance et al. (2020) measure the separation of representations using Latent Geometry Graphs, where edges are weighted by the level of similarity of the representations. They compute the graph on the training data set on which MixUp has been applied and then measure the label variation by summing up all the edges that connect vertices from different classes.

- Schiff et al. (2021) also resort to MixUp-like distortions to estimate the performance of neural networks. Compared to the winning solution, they evaluate the effect of MixUp perturbation for a range of mixing coefficients. Computing the accuracy along different perturbation intensities allows them to compute what they refer to as the Perturbation Response Curve. They then compute a Gini coefficient and a Palma ratio-inspired score. The former is given by the area between the cumulative Perturbation Response Curve and the curve of an idealised network (i.e. one that is completely robust to the MixUp perturbation). The latter is computed by taking the ratio of response to the largest 60% perturbations and lowest 10%. This is inspired by the Palma-ratio income inequality metric used in economics, where the income of the richest 10% of a population is divided by that of the poorest 40% (Palma, 2011). Schiff et al. (2021) do not justify the change of ratio.

- Carbonnelle and De Vleeschouwer (2020) propose four measures for the presence of intraclass clusters which they believe occur as implicit forms of regularisation. However, since most of these measures rely on the existence of hierarchical labels, they only propose one for predicting generalisation: the variance of pre-activations

for samples of the same class compared to that of all samples. The rationale is that distinctive features would have a high such ratio. It is important to note that they only consider those activation maps with the k highest variance.

At the time of writing, Carbonnelle and De Vleeschouwer had not yet published their code. We are thus unsure if the original design was to compute the ratio for the neurons that are most activated for all classes or only those most activated for each individual class. However, the latter is more informative and so we chose to implement this variant. Therefore, we take the preactivations to be the feature maps after the batch normalisation layers and consider a neuron to be the entire feature map. For selecting the top k most activated neurons for each class, we collect the activations of neurons across layers and select the ones with the highest standard deviation. Then, using the most activated neurons for each individual class we compute the ratio.

- Mežnar and Škrlj (2020) explore 7 different metrics and their binary combinations. Note that they allow a metric to negatively contribute. Some of the quantities bear some similarities with previously proposed measures, such as the distance to weight initialisation, the percentage of weights that changed more than a certain amount, or the margin, defined here as the percentage of instances where the predictions for the top two classes is greater than a threshold. More interestingly, they continue the training for five more epochs using learning rates of different magnitudes and compute statistics on the models that were obtained with each learning rate. They then use the statistics as indicators of the generalisation performance of the source model.

While we focus here on attributes explicitly linked to generalisation, we note that there are cases in the literature where qualities of learnt representations were informally linked to generalisation. One such example can be found in the work of Geirhos et al. (2019), who argue that increasing the amount of shape information in a network increases both the robustness and generalisation performance of a model. Another example is the belief that maximising the margin at intermediate layers improves generalisation (Elsayed et al., 2018). We will introduce more such informal propositions in the following subsection where we discuss miscellaneous attempts which do not directly fall in any of the categories outlined so far.

What are the recurring ideas? The majority of methods use data modification in their evaluation as a means of identifying and penalising overfitting. This can be used as an informal way of measuring robustness to perturbation. Some of the methods combined this notion of robustness with different notions of representation clustering. Going back to the Information Bottleneck theory perspective, the former can be interpreted as a crude approximation of the mutual information between the representations and the output variable, while the latter crudely approximates the level of compression, or

the information between the representations and the input variable. Therefore, at an intuitive level, the combination has the potential to capture a relevant phenomenon. However, as we will argue, the distortions considered are too limited to provide reliable estimations.

How are they different from what has been proposed before? While the ideas presented here bear similarities to the concepts proposed in the literature centred around expressive power, they are more flexible in their definitions. For example, they do not aim to mathematically define vicinity and then measure sensitivity to perturbations in that strictly defined vicinity. Instead, they focus on intuitive notions of vicinity. Although the approaches we presented above are not well justified, as we will argue later in this chapter, we believe these studies open a new avenue for research and call for the community to reason about robustness from a different perspective.

Other Directions

Interesting proposals for capturing generalisation have been also made in the adversarial robustness literature. For example, Zahavy et al. (2016) argue that the adversarial robustness of an *ensemble* of models correlates well with generalisation. To this end, they perform multiple runs of the same hyperparameter configuration, the only difference between the runs stemming from the stochasticity of the SGD optimiser. They argue that although individual hypotheses might be sensitive to adversarial examples, it is the robustness of the *average* hypothesis that determines the generalisation performance. We will encounter a similar idea in the a posteriori setting, which we discuss later in this section. Zahavy et al. (2016) then use this correlation to explain how the models we obtain in practice are generalising well despite not being adversarially robust. Note that this estimator does not fit the evaluation frameworks proposed so far since it strictly requires access to multiple runs of the same network.

Other lines of work that have gained attention but which we do not cover in this thesis include those which aim to relate training dynamics to generalisation performance. A prominent theoretical direction is represented by Neural Tangent Kernel-based methods (Jacot et al., 2018). The Neural Tangent Kernel is the kernel to which an infinite-width neural network would converge when trained with gradient descent. It has recently been noted that an increasing number of Neural Tangent Kernel-based studies (e.g Adlam and Pennington, 2020*a*; Bietti and Mairal, 2019; Belkin et al., 2018) rely on the test data to belong within the convex hull of the training data, which is extremely rarely the case (Balestriero et al., 2021). Nonetheless, the Neural Tangent Kernel remains an interesting theoretical direction to be explored and refined further.

On the practical side, Gutiérrez-Fandiño et al. (2021) use measures related to training dynamics to predict the performance of neural networks. Note once again that we are

interested in predicting the performance of a learnt instance, not taking into account the learning process for the moment. Nonetheless, we also briefly mention this direction for completeness.

Learning dynamics have also attracted the attention of some empirical researchers whose work broadly falls in the subfield of empirical theory of generalisation. The empirical theory field aims to *explain* certain phenomena in deep learning rather than create a framework to *capture* the mechanism behind generalisation at large. We regard this subfield as complementary to our direction and strongly believe both are necessary for solving the generalisation puzzle. Examples of studies in this area include those of Power et al. (2022), Arpit et al. (2017), and Amari et al. (2021).

Another set of studies that aim to explain certain phenomena related to generalisation is represented by the "memorisation" literature. Note that "memorisation" is used ambivalently by the machine learning community. It is either taken to mean achieving zero training loss, typically on noisy data, (e.g. Dherin et al., 2021) or learning to predict atypical samples (e.g. Sagawa et al., 2020). Note that here atypical samples are defined in the same way as in Chapter 3, namely those samples that, when removed from the training data, lead to the misclassification of a sample from the test data. This is considered by Sagawa et al. (2020) as an indicator that the train-test pair belongs to the tail of the data distribution. Both memorisation scenarios pose a number of important questions on how models generalise and, along with all directions in empirical theory of generalisation, complement our perspective.

The double descent literature is also concerned with explaining a specific generalisation phenomenon. The term "Double Descent" was introduced by Belkin et al. (2019). Double Descent is illustrated in generalisation plots as a downward continuation of the typical U-shaped curves used to depict the *complexity* trade-off. Note that the generalisation curve associated with this learning algorithm had been observed before (e.g. Opper, 1995; Engel and den Broeck, 2001) but it was made popular by the work of Belkin et al. (2019). To show the occurrence of the double descent, Belkin et al. performed ERM on two-layered networks with fixed input weights represented as Random Fourier Features (Rahimi and Recht, 2007). Importantly, out of the ERM subset, their learning algorithm chooses the hypothesis with the lowest l_2 norm. Belkin et al. then plotted the generalisation error of the minimum-norm solution for models with an increasingly larger number of possible features.

The setting that Double Descent is concerned with is not reflective of the typical practical learning scenario; it is peculiar and over-regularised. More specifically, Double Descent is mostly characteristic of overfitting models (see Nakkiran et al. (2021); Maddox et al. (2020)) and is a oddity of the learning dynamics (i.e. optimisation process) and *choice of learning algorithm* (i.e. selecting the model with minimum l2-norm) rather than a general observation. That is because a certain hypothesis chosen according to

specific rules is not reflective of the generalisation performance of the model one would normally expect to learn. The observation that Double Descent is a quirk of the choice of learning algorithm has also been recently made by Dwivedi et al. (2020) and Wilson and Izmailov (2020). Taking a probabilistic approach Wilson and Izmailov (2020) show that by marginalising over multiple predictors, the double descent curve disappears.

For completeness, we also mention the "Effective Model Complexity" (Nakkiran et al., 2021), another informal proposition that has not yet seen much popularity in the generalisation community. Nakkiran et al. (2021) introduce an "Effective Model Complexity" measure when empirically studying Double Descent. This measure gives the maximum number of training examples such that the expected risk of a hypothesis obtained with a particular training procedure is smaller than a threshold. Nakkiran et al. then argue that it is this notion of complexity that allows them to show the existence of a Double Descent curve which cannot be captured by the VC dimension or Rademacher complexity. However, this is not the quantity they measure in practice. It must be noted that this measure cannot be used to compare the complexity of two arbitrary models as this would depend on the learnability of the models under the same training procedure. Although studying learnability in an overparametrised regime could shed further light onto generalisation by helping the community understand overfitting better, here we are interested in capturing generalisation performance at large.

Motivated by the phenomena of exploding/vanishing gradient, another proposition is to approximate the level of nonlinearity in the network *at initialisation* (Philipp and Carbonell, 2018). Note that this differs from the setting we consider since Philipp and Carbonell (2018)'s approach is not concerned with a learnt model instance. It is, in a sense, trying to identify unfruitful initialisation states. The nonlinearity coefficient aims to measure the network's output sensitivity to input distortions *relative to the network's overall output variability* as measured by the output covariance matrix. They claim that the coefficient computed before training correlates well with the network's generalisation performance after training. However, the empirical evaluation of this approach is very limited compared to other estimator proposals and does not seem to correlate sufficiently well with generalisation on the explored settings. Note that closely related to Philipp and Carbonell's measure of the level of "nonlinearity" are works that associate the number of linear regions in the network with its complexity (e.g. Novak et al., 2018), which we have mentioned in Section 4.1.2. As noted before, Novak et al. (2018) consider the a posteriori setting.

Before concluding our overview, we would like to briefly explain the a posteriori setting in more detail and mention a number of additional works in this area. We remind the reader that we are interested in the evolution of proposed concepts. Thus, although these studies do not fit the setting we consider, the intuitions behind the approaches are still relevant to our work.

The A Posteriori Setting. The difference between a priori and a posteriori is given by the latter's access to *unlabeled test data* at inference time. In other words, if in the a priori setting we aim to predict the generalisation performance of a model using information about the trained model instance and the labeled training data, in the a posteriori setting we also have access to *a new set of samples for which we are not given the label*. The target variable is the true accuracy of the model on the new set of samples. Therefore, the task is to estimate the accuracy of the model on input samples that we have access to. This problem should be easier than the a priori one since we do not have to estimate the population test accuracy, but instead that on specific instances. We could, in theory, measure the distribution distance between the training and test samples. As we will discuss in Chapter 5, in practice this is not as straightforward.

Note that this setting can have practical relevance if we see it as a way of guaranteeing performance on newly acquired data. For example, we could train a model and validate it on held-out data but ultimately we are interested in classifying *unlabeled* data. Finding a solution for the a posteriori setting would give us a tool that can tell when the predictions on the newly acquired data cannot be trusted. We will next present some a posteriori estimators.

In Chapter 1 we mentioned that raw generalisation error is not all one cares about in practice when evaluating the quality of a learnt model instance. Model calibration measures how well the confidence of the model aligns with the expected accuracy (Niculescu-Mizil and Caruana, 2005; Guo et al., 2017). Although initially not linked to generalisation, recent work aims to connect a small generalisation gap to good calibration (Carrell et al., 2022). Similarly, Jiang et al. (2022) train an ensemble of models using the same training procedure for all models, but different random seeds. Therefore the difference between models stems from the randomness of the SGD optimiser, the order in which the data is presented, and the initialisation point. They then look at the rate of disagreement, also referred to as *predictive churn* (Bahri and Jiang, 2021). It has been claimed that the rate of disagreement closely tracks the generalisation performance (Nakkiran and Bansal, 2020). Jiang et al. (2022) build on this notion and argue that the generalisation performance can be predicted based on models' rate of disagreement on unlabeled test data, provided that the model ensemble is well-calibrated. They then go further and argue that the rate of disagreement can also help predict models' performance when varying batch or network size, therefore reflecting even small changes in the network. However, the calibration assumption might be too strong in the case of modern deep networks (Guo et al., 2017). Jiang et al. (2022) do not compare their method against other empirical estimators, and theoretical concerns with respect to their approach have been raised (Kirsch and Gal, 2022).

In a similar vein, Morcos et al. (2018) performed a small-scale study to argue higher representation similarity among generalising models as opposed to their overfitting counterparts. They refer to the latter as "memorising" networks, which were trained on

randomised labels. In their study, they trained multiple models starting from different initialisations. They computed a similarity measure at each layer and notice different behaviours between generalising and overfitting networks at late layers. It is important to notice that the similarity level at the softmax layer is the same for the two types of networks when analysed on training data. It is only on test data that this measure becomes discriminative. While such a method could constitute a good a posteriori evaluator, it has a significantly higher computational cost than simple churn. Therefore, the question of whether it is more indicative than churn arises. We also note that this approach has not been empirically compared against other estimators.

Finally, a few studies have tried to train linear predictors in order to estimate the generalisation performance of models. The first such approach is that of Deng and Zheng (2021), who use the Fréchet Distance between the train and test data representations to predict the test accuracy. Note that the distance is computed on the feature maps obtained in the penultimate layer of the network evaluated. Similarly, Deng et al. (2021) aim to predict the generalisation performance of a network by evaluating it on the side task of recognising the angle at which images were rotated. To do this, they train networks on two simultaneous tasks, the standard task where the network predicts the class each image belongs to, and a rotation classification task where the network is trained to predict the angle at which the image was rotated, choosing between four different possible angles of rotation. They report a linear correlation between a network's ability to identify the angle of rotation of an image and its ability to correctly classify that image. Note that this approach is more restrictive than that of Deng and Zheng (2021) as it requires training on an additional task.

The methods proposed by Deng and Zheng (2021) and Deng et al. (2021) have little theoretical justification and have not been evaluated in large-scale settings with varied scenarios. They, however, propose a possibly simpler approach to understanding generalisation which is to train a model to learn from the "generalisation data", a *meta* data set. We argue that it is conceivable that none of the proposals in this chapter captures generalisation performance entirely, and one might have to consider *combinations* of these estimators. We believe that once a number of good individual estimators founded on solid intuitions are proposed, one could resort to such a meta data set to further understand the intricate mechanism behind generalisation. We will come back to this idea in the Future Work section of this chapter.

Summary

There is a great variety in modern approaches to capturing generalisation. We believe studying their evolution is important for understanding how ideas came to be and what they are fundamentally trying to model. An interesting evolution, for example, can be observed in the case of weight norm. As we have seen, the idea of norm started with

having a very close link to the notion of margin, as was the case for SVM classifiers. It was much later translated to deep convolutional networks, resulting in increasingly complex notions of weight norm at different layers, units, or even norm of the weight distance from initialisation, as well as various combinations of the aforementioned quantities. It might be difficult to understand their essence if we take the resulting complex combinations in isolation. However, we believe understanding their history makes these concepts more tangible.

Consequently, we believe that our overview can help the community abstract away from the exact techniques of each direction and gain a better high-level perspective. We hope that this will encourage researchers to reflect on the direction of generalisation studies. We believe our work can help future researchers avoid unknowingly taking already explored paths, especially since many directions are based on the same fundamental ideas. For example, separability ultimately aims to quantify how close the network's decision boundaries are to the training samples. We have seen above that separability can be linked to sensitivity, which can be linked to robustness and stability, which in turn can be related to properties of the loss landscape. Therefore, many directions try to capture the *same notions* from different angles, using other definitions and tools. The inherent problem is that it is difficult to reason about and capture the essence of this idea in the dauntingly high dimensional spaces we are concerned with. Additionally, the focus until recently was on using these notions to *bound complexity*, which we see as one of the reasons why the field has not yet managed to correlate good notions of separability with generalisation performance. As we have mentioned before, bounding generalisation remains an important pursuit but the field's understanding is still too limited for creating bounds that are relevant in practice.

As alluded to in Section 4.1.5, we believe the idea of separability, which is behind the above directions is promising and should be explored further. Using the language of robustness, we believe the key to achieving in-distribution generalisation is not robustness to any type of distortion but rather to those that are "meaningful" or "natural". Therefore, although separability is ultimately desirable in all directions in the representational space, it is sufficient to be well separated in the directions that *matter*. We will come back to this idea in Section 4.3, where we discuss future work.

Lastly, we have seen that the field is starting to search for notions that *correlate* well with generalisation performance. We believe the main limitation is that the notions that have been proposed are rather unfounded, and their empirical evaluations are limited. There is no large-scale study that evaluates measures from *all* the directions we have reviewed. Therefore it is difficult to get a clear image of how each estimator performs with respect to all other methods. To get a partial idea, throughout Section 4.1.1 we reviewed and outlined the findings of the existing comparative studies. However, as we will see next, these comparative studies are still limited in the number of settings they

consider. Therefore it is possible that the above findings do not reflect the real ability of estimators to capture generalisation.

4.2 Evaluating Empirical Estimators

In the previous section we have seen the accelerated emergence of empirical generalisation estimators over the past three years. The rampant evolution of the field immediately calls for establishing the validity of such estimators. The first attempt to evaluate quantities relating to generalisation goes back to the complexity-focused literature of bounding approaches (Jiang et al., 2020). In this context, three possible ways of evaluating complexity measures have been considered: comparing the tightness of the obtained bound, using the notion of complexity as an explicit regulariser when training, or determining the correlation between the complexity measure and empirical generalisation performance.

The former does not have an equivalent in the context of empirical estimators and we are therefore left with two possible approaches. Jiang et al. (2020) argues that explicitly regularising complexity is more problematic since regularisers change the optimisation problem, possibly increasing its difficulty. Moreover, it could be hard to differentiate between the regularising effect of the measure and the implicit regularisation of the optimiser. Although when measuring correlation with generalisation, one could accidentally capture spurious correlations, Jiang et al. (2020) chose this option, which became the norm in empirical generalisation studies.

In this thesis we also chose to evaluate the correlation with empirical generalisation performance. To do so, one must evaluate the performance of the estimator on a large number of models. This creates a meta problem with an associated data set. In its simplest form, the input variable is represented by the learnt model instance, the data samples the model was trained on, and any additional variables we want to consider (e.g. details about the training procedure, initialisation point, or even unlabeled test data). The target variable is given by the test accuracy. To differentiate between the data sets on which models are trained and the data set consisting of model instances and their generalisation performance, we call the latter *the generalisation data set*.

There is no set standard for how these generalisation data sets are constructed, with each study defining its own points of interest and constraints. For example, some might only consider as input the model instance and the training samples while others provide more details. Some studies aim to predict the performance of a learnt model instance while others turn the problem into a comparative one: given two model instances, can an estimator rank the generalisation performance of the two models? Naturally, this slightly changes the structure of the data set and the evaluation criteria. Some studies consider a variety of learning scenarios while others are very limited.

We argue for a well-considered and unified analysis and identify a number of limitations of the prior art. Particularly, we highlight important learning settings that are being neglected. Ideally, we would employ our arsenal of trained models used in this thesis so far to show that estimators which were found by previous large-scale comparison studies to correlate well with generalisation would perform poorly in the scenarios we describe. This, however, is infeasible due to the very specific evaluation criteria set by previous comparison studies. We will discuss the criteria in detail in Section 4.2.2 and only note for the moment that it implies systematically changing each hyperparameter value. The reason behind choosing such criteria was to try to rule out spurious correlations. The implication is that we can only argue the importance of additional scenarios in an informal manner, with no experiments to support our claims. We leave these as future work that would build on the setting we propose.

Systematic Changes or a Variety of Scenarios? Ideally, one would create an evaluation setting that systematically varies all hyperparameters in an exhaustive variety of settings. Given the current level of computing power, this is infeasible and we are therefore faced with a tradeoff. So far, the balance has been in favour of systematic changes of hyperparameter values. This, we argue, has led to overoptimistic evaluations caused by the limited pool of scenarios. We, therefore, argue that increasing the number of considered settings is necessary for getting a more accurate picture. Coupling the newly proposed scenarios with systematic changes would significantly scale up the size of the generalisation data set. The societal issue associated with this is the huge environmental cost of validating approaches on such a computationally-intensive data set. Moreover, a practical limitation is storing and making publicly available a very large number of trained models. For this reason, we argue that the community must focus first on identifying those estimates that perform well in a variety of settings. We argue for this approach since we believe it is easier to mistakenly design an estimator that is predictive of generalisation in a very limited number of scenarios than one that spuriously correlates with generalisation in a great variety of scenarios.

Why Estimate the Test Accuracy? In the previous chapter we went from bounding the generalisation performance of models, which has the clear goal of providing guarantees, to empirically estimating generalisation. So far we have broadly argued for practical experimentation as a way to build stronger intuitions. But why would estimating the test accuracy be the way forward? The hope is that predicting the test accuracy would allow researchers to get a better sense of what is important for generalisation and what is not. In other words, the idea is to identify notions that correlate well with generalisation in practice. Studying then those notions and formalising them would hopefully lead to better generalisation theories.

Why Would This Be Any More Fruitful Than What Has Been Attempted So Far? During one of the conversations with other researchers, were asked the following question: "People have been working on this problem for a long time. Why do you

think *you* are going to solve this problem?". Our answer is that we do not necessarily believe in an individual breakthrough but rather in an iterative effort of the community. The research in this area is rapidly growing and we believe with each contribution the community as a whole is one step further to understanding generalisation. But we believe that it is important for the researchers working in the various subfields to be aware of the direction in which the field as a whole is headed. We therefore aim to provide this unified understanding and help researchers reflect on the bigger picture while proposing a new vision for the future of the field. We believe that in order to solve the generalisation puzzle, the field needs to make a coordinated effort through sustained collaboration, rigorous scrutiny of prior contributions, and a good understanding of the concepts already explored. We believe these are currently missing. This is, in our opinion, an important limitation of the generalisation community that must be addressed.

For this reason, we aim to build a publicly available generalisation data set and appeal to the wider machine learning community to help extend it and, through time, expand it to include more challenging settings. Moreover, we aim to publish our overview of the decades-long research done in the generalisation field to make it easier for the community to make informed contributions. We also advocate for transparent assumptions and a more meticulous evaluation of proposed approaches. Therefore, our goal is not to be the individuals that solve the generalisation puzzle, but rather to be part of the community that has done so.

We will next go through the design choices of previous studies which we will draw inspiration from. We highlight the strengths and limitations of each one. Supporting our view using conceptual arguments, we then design a new setting for evaluating empirical estimators of generalisation.

4.2.1 A Brief Overview of Prior Evaluations

In this section we go back to some of the approaches introduced in Section 4.1.5. More precisely, we discuss the *evaluation setting* of those approaches that empirically evaluated estimators. Note that while Section 4.1.5 discussed how these estimators relate to each other, here we are only concerned with the *design* choices of their evaluation.

We focus on the evaluation part of two types of studies: those that aim to provide large-scale *comparisons* of already proposed methods; and, those that propose new estimators and aim to bring empirical evidence in support of their *individual* success. We discuss details such as the data sets and augmentations considered, the architectures, and the hyperparameter choices. We aim to highlight unique decisions made by each study where these exist.

Individual Studies

Most of the works that solely aim to empirically support their proposed bound or estimator tend to be very limited in scope. They all claim to find a correlation between their respective estimators and generalisation performance. Among early attempts, Arora et al. (2018) visually analyse the correlation between their bound and the test error of a single network. One remarkable difference is that they plot the error and the computed bound at different points throughout the learning trajectory. Likewise, Neyshabur et al. (2019) plot the generalisation performance of a two-layer network and the evolution of their complexity measures for an increasing number of hidden units. A similar visual analysis is carried out by Liang et al. (2019) in a label randomisation context. Liang et al. (2019) consider 9 width values, while Neyshabur et al. (2019) look at the correlation for 8 width sizes. They do not vary any other hyperparameters. Neyshabur et al. (2019), however, repeat the plot for a ResNet architecture and consider more data sets (CIFAR-10/100 and SVHN (Netzer et al., 2011)). Nonetheless, this type of evaluation is insufficient for the empirical justification of the estimators.

Later attempts have aimed to investigate generalisation in broader settings. A first example is that of Philipp and Carbonell (2018) who study fully connected networks. They vary the size of the network, the activation function used as well as the presence, location, and strength of skip connections. They also experiment with either batch (Ioffe and Szegedy, 2015) or layer normalisation (Ba et al., 2016). The vision data sets they consider are MNIST and CIFAR-10, alongside the Waveform data set (Breiman et al., 1984). Interestingly, for each network configuration they consider 40 learning rate regimes and only select the best performing one for constructing their generalisation data set. SGD is the only optimiser they use. Out of all the empirical evaluations we present, Philipp and Carbonell's has the highest variety of activation functions, with eight different options considered. This design choice is justified by the intuition behind the estimator they propose, which is to measure the level of nonlinearity in the network based on the activation regions in the network.

Morcos et al. (2018) focus on studying the effect of the learning rate. To this end, they train 200 different networks where the learning rate was varied. The vision data set they chose to experiment with is CIFAR-10 and they have only considered the Adam optimiser.

Gutiérrez-Fandiño et al. (2021)'s estimator requires the representations to belong to a fully connected layer. Nonetheless, apart from purely fully connected architecture, they experiment with pretrained Convolutional Neural Networks on top of which they train a Multilayer Perceptron and use its representations for evaluation purposes. Gutiérrez-Fandiño et al. vary the number of units per layer, but not the number of layers. They consider 5 different learning rates and 5 different Dropout (Srivastava et al., 2014) probabilities but fix the batch size. Most notably, they experiment with modifying the data

sets by removing classes altogether. However, they do not motivate this choice, nor do they provide a separate analysis for this case.

Jiang et al. (2022) evaluate their estimator on ResNet, simple Convolutional Neural Networks, and fully connected models trained until close to zero training loss is achieved on the CIFAR-10/100 and SVHN data sets. They consider a number of different batch sizes and network configurations, as well as different weight decay regimes. The optimiser used is SGD. Jiang et al. (2022) also use data augmentation without specifying the exact augmentation applied. They similarly vary the size of the training set but do not mention the details. Lastly, they report results on some out-of-distribution settings.

Carbonnelle and De Vleeschouwer (2020) consider CIFAR-10/100 and the 20-classes version of CIFAR-100, training Wide ResNets and VGGs models on them. They vary the network size, the learning rate, batch size, the level of weight decay, and Dropout probability. They also consider the Adam optimiser alongside SGD, which few of the individual studies do.

Comparative Studies

One of the first studies that is not concerned with a single estimator is that of Novak et al. (2018). In their evaluation they only considered the CIFAR-10 and CIFAR-100 data sets. Just like some of the individual studies, they vary hyperparameters such as learning rate, batch size, and network size. Importantly, they consider image translation and flip as augmentations, as well as label randomisation.

Subsequently, Jiang et al. (2020) proposed the first large-scale comparison of generalisation measures, which became the backbone of most ulterior comparative studies. It represents a valuable contribution to the community not only through the scientific findings but most importantly by the standard of evaluation they set. In particular, they aim to design an evaluation metric that would allow them to rule out misleading correlations between estimators and generalisation. This implies training a large number of models by systematically changing a chosen set of parameters. Thus, in a sense, the focus falls on the relationship between generalisation and various hyperparameters.

Jiang et al. (2020) train their fully connected models on CIFAR-10 and only present some results on SVHN in their supplementary material. They consider 3 different values for the weight decay, network width and depth, batch size, learning rate, Dropout fraction, and optimiser and train models with all combinations of values. They optimise the cross-entropy loss until it reaches a near-zero value of 10^{-3}.

We see their decision to only include models that reach a close to zero training loss as one of the main limitations of the experiment conducted by Jiang et al. (2020). As we will argue in Section 4.2.2, this is highly constrictive and fails to capture a wide range

of phenomena in machine learning. Similarly to us, Vakanski and Xian (2021) argue that studying measures in a zero-training loss regime only is not reflective of practical settings. The limitation of this regime is also noted by Carbonnelle and De Vleeschouwer (2020).

Another direction that was not included in Jiang et al. (2020)'s analysis is *data variation*. The one case they consider, but do not experiment with, is label randomisation. They argue that altering labels could be misleading since this is not a typically encountered situation in practical settings. We agree that this is true for extreme cases of label randomisation. However, one could expect the data collected in practice to be significantly noisier than the standard vision data sets. For this reason we believe models should also be evaluated under mild label randomisation when the data set in case is known to be well-curated. Nonetheless, we argue that one can include many other scenarios of data modification when evaluating models. Most studies do not investigate this area or consider only a limited setting.

Jiang et al. (2020)'s work has inspired a number of subsequent studies, one of which is that of Vakanski and Xian (2021). They focus on generalisation for medical image predictors and aim to incorporate more architectures, data sets, optimisers, etc., than previous studies. Some of the challenges of the medical imaging setting are the small size of the data sets and the large image resolution. An important attribute of the data Vakanski and Xian consider is that the images were collected in different conditions and on varied populations. This makes it a good candidate for evaluating models in an out-of-distribution scenario. Although such a targeted study highlights that the superiority of a generalisation method is dependent on the type of problem being addressed, we believe an evaluation of the like is insufficient for drawing more general conclusions. Some of the limitations of Vakanski and Xian's study are that only two data sets are considered, the evaluated measures are exclusively complexity-based and no form of regularisation is used.

A highly rigorous evaluation of predictors can be found in the work of Dziugaite et al. (2020). However, we believe the rigour made it difficult for Dziugaite et al. (2020) to explore more settings and data sets. Just like Jiang et al. (2020), they were limited to the CIFAR-10 and SVHN data sets. Apart from the choice of data set, they vary 4 hyperparameters: the learning rate, model width and depth, and the training set size. For optimisation they only consider SGD with fixed momentum, do not use weight decay, and do not change the learning rate throughout training. Dziugaite et al. and Jiang et al. made use of architectures that do not reflect the types of architectures commonly used in the field. This is due to the complexity measures they evaluate, which either quickly become infeasible or are not properly defined for models with skip connections.

Although atypical, another piece of work inspired by Jiang et al. (2020) that deserves mentioning is the "Predicting Generalization in Deep Learning" (PGDL) contest (Jiang

et al., 2021) from which many new estimators have emerged. This contest closely follows the methodology established in Jiang et al. (2020) and, in our opinion, provides a good platform for initial evaluations.

Jiang et al. (2021) train VGG-like and fully connected networks until close to zero training loss is achieved. Note that they do not consider networks with skip connections. The data sets included in their study are CIFAR-10, SVHN, CINIC-10 (Darlow et al., 2018), Oxford Flowers (Nilsback and Zisserman, 2008), Oxford Pets (Parkhi et al., 2012) and Fashion-MNIST. For each data set, they vary the size of the network, the weight decay, batch size, and Dropout rate. Just as we are intending to do, Jiang et al. (2021) make their data set publicly available, facilitating the community's ability to empirically validate generalisation estimators.

It must be noted, however, that this approach to understating generalisation has also received criticism. In particular, Martin and Mahoney (2021) argue that such competitions provide a distorted image of estimators' performance. In their view, it is most often the case that good solutions often get outshined by solutions that are specifically tailored to win in the fixed setting of the competition. Thus, Martin and Mahoney aim to more closely analyse the methods proposed. They only study measures concerned with the norm and shape of the learnt model instance weights. They argue that no metric can fully describe the generalisation performance and that researchers should rather seek a combined approach.

We agree with Martin and Mahoney that having a fixed generalisation data set could lead to the emergence of estimators that "overfit" to the considered setting. By the nature of Jiang et al. (2021)'s study, a thorough justification for the proposed methods was not relevant. As such, retrospectively, the winning solutions could seem to lack soundness in places, as we will discuss later. However, we also believe they deserve attention. This is not only to highlight their limitations but also to understand their strengths and exploit them further.

Importantly, Martin and Mahoney also criticise the use of augmentation for the purpose of the experiment, this being seen as an attempt to artificially get better results. While this could be true of many competitions, we believe in this case Martin and Mahoney omitted the power of data manipulation to capture learnt representations' quality which we see as an important direction for future research.

Martin et al. (2021) follow up on Martin and Mahoney (2021)'s work. In their experiments, Martin et al. (2021) used 17 different architectures. The main data sets they train on are a reduced version of ImageNet, CIFAR-10/100, and SVHN. We could not find details on their choice of hyperparameters such as optimiser or learning rate. One unique design choice is not to use the training data when estimating the generalisation performance. We will come back to this decision later in this chapter.

Summarising, the key quality of previous comparative studies lies in the extensive number of models they consider and in the systematic evaluation under very specific parameter changes. Important limitations include the stringent nature of the ERM setting and the restricted pool of architecture types. To these, we add the lack of data variations and little diversity in the learnt representations. Moreover, to the best of our knowledge, there is no large-scale study that analyses newly proposed measures that go beyond the notion of complexity. This motivates the following section where we take the first steps towards such a study. We discuss the setting we choose for the study that will be part of future work.

4.2.2 Our Proposed Setting

In the future we aim to create an up-to-date study that evaluates the new directions in generalisation prediction while arguing for the consideration of a broader setting than has been achieved currently. Importantly, note that although we started creating a generalisation data set, this is mostly left as future work. Therefore, this section presents a draft of the *proposed design*, not the details of a fully-built data set.

We propose to vary the hyperparameters very little, both because this has been thoroughly discussed in previous studies and because saving computation allows us to explore other important aspects of generalisation. Our focus is on having a large variety of learnt representations through modifying the training data. Thus, central to our work is the *relationship between data and generalisation*. As such, our proposed setting does not aim to replace existing studies but rather to complement them. Below, we detail and justify each design choice, positioning our future work with respect to prior art. In our vision, one could use our proposed setting as a way to filter out those estimators that do not reflect changes in the data well. The remaining estimators could subsequently be evaluated on the data set proposed by Jiang et al. (2021).

Assumptions About the Data and Hyperparameters Availability

Most a priori estimator evaluation studies assume the learnt model instance and the training data are provided. One of the exceptions is the work of Martin et al. (2021), who analyse pretrained models assuming that neither train nor test data is provided and that no information about the training process or hyperparameters used is made available. Martin et al. aim to replicate the case where the model user does not coincide with the model developer. This setting is interesting but we believe it is too stringent especially given how poorly generalisation is understood. As we have argued throughout the thesis, we believe integrating the data is a crucial step in furthering the understanding of generalisation and since there is no universally good solution, one must consider the problem of interest.

The estimators we have discussed do not make use of the choice of hyperparameters or initialisation state when predicting the generalisation performance of models. The exceptions are those which, in one way or another, account for the training dynamics. Although we have explicitly excluded these from our setting, there is a question related to the future scalability of generalisation data sets. Our take on this is that not including the initialisation state and training hyperparameters would allow the community to make use of a variety of publicly available pretrained models, significantly reducing the computational cost of building such a data set. In this sense, we have developed a script that makes models trained by Jiang et al. (2021) for the purpose of the Predicting Generalisation in Deep Learning competition compatible with PyTorch code, such that we can more easily access a wide pool of already pretrained models.

At the same time we appeal to the wider machine learning community to store details about their hyperparameter choices, training procedure, initialisation state, etc., and make them publicly available alongside the pretrained models. This would allow us to expand generalisation data sets in the future so as to account for dynamics-centred approaches as well, as these could play a role in solving the generalisation puzzle.

Types of Estimators Considered

The assumptions about what information is available implicitly restrict the type of predictors we can consider. Naturally, an initial study would exclude all estimators that need more information than the learnt model instance and the training data. The hope is that a collective community effort to gather complete information for a large number of models would allow us to expand the scope of the study.

No large-scale comparisons have yet been conducted on measures that move away from classical notions of generalisation. The closest such comparison is the Predicting Generalisation in Deep Learning competition, although it only includes those methods which were submitted to the competition and does not provide an in-depth follow-up analysis. Given both the extensive coverage of complexity-based measures as well as the reported success of the non-complexity approaches, we choose to primarily focus on creating a data set for evaluating the latter. Since they are generally less computationally intensive, the immediate implication is that we can extend the experiments to include architectures commonly used in practice.

Architectures

As highlighted in the overview, previous studies have mostly trained simplistic models, with no skip connections or reduced number of parameters (e.g. small-scale versions of the VGG network). This is mostly because at the time when the evaluations were

proposed, most empirical predictors could not be computed for the types of architectures that were achieving state-of-the-art results. Our proposal is to construct, in a data-centric way, two subsets of the generalisation data set. One that is compatible with early proposals, and a second one, which incorporates the most recent advances in machine learning architectures.

One exception among prior studies can be found in the work of Martin et al. who, like us, aim to also incorporate ResNets, full-sized VGGs, and DenseNets (Huang et al., 2017). To these, we also add BagNets, since the small receptive field restricts the model to learn more localised features. This brings further diversity in the types of learnt representations included in the study. Importantly, we believe more up-to-date architectures must be included, in particular transformer models, which are currently achieving state-of-the-art results in vision applications (e.g. Dosovitskiy et al., 2021; Yu et al., 2022) as well as large convolutional networks such as PyramidNet (Han et al., 2017*b*), GoogLeNet (Szegedy et al., 2015), large-kernel models (e.g. Ding et al., 2022).

We remind the reader that the intent is to provide an evaluation where the focus is on the relationship between data modification and generalisation performance. This would be followed up by the hyperparameter-centric evaluation of Jiang et al. (2021) for which we would use the already trained models made publicly available. To make the transition between the two scenarios less abrupt, we will choose a small number of architectures used in Jiang et al. (2021)'s study and train models under various data-centric scenarios. Note that we would generally fix a hyperparameter setting and only modify the data. For each architecture we plan to base the hyperparameter settings on values that are reported in the literature to lead to good generalisation. We will discuss hyperparameter settings later in this chapter.

We remind the reader that most prior studies trained their models for a very large number of epochs or until close to zero loss was achieved. Before discussing how we train the architectures we chose, we make an observation about the setting typically adopted by generalisation studies.

To What Extent Is Replicating ERM Relevant?

The context in which the zero-loss setting was proposed is necessary for understanding why it is restrictive when aiming to find estimators that are relevant in practice. The Empirical Risk Minimisation setting is fundamental for providing *guarantees* in classical statistical learning without making further assumptions about the data distribution. The early empirical estimators sought tight links with theoretical generalisation bounds since this is the literature they have evolved from. For this reason, the zero training loss regime was a natural setting to study. More recent approaches are not concerned with bounds anymore but rather, as we argue, with building an initial understanding

based on practical experimentation. We believe the accent must fall on verifying the estimators in scenarios that are likely to occur in practice. Although one can train to zero training loss, this is not necessary, especially in non-standardised data sets.

Real-world data sets are bound to have mislabelled data. Even the standard data sets used in computer vision are known to contain incorrect data samples. Learning the incorrect train data automatically leads to choosing suboptimal hypotheses. More specifically, it turns the study into an overfitting regime study. Thus, forcing all the models to achieve zero loss captures only a specific part of the generalisation spectrum, whereas we would like to be able to predict the generalisation performance of any model instance.

This is not the only limitation of the zero-loss setting. Jiang et al. (2020) find that data augmentation makes it difficult to reach zero loss and choose to not use data augmentation in their study. This is only one example of how training in this regime restricts the types of learning scenarios one can explore. We therefore do not aim to train our models in this regime.

Loss Function, Optimisers, Learning Rate, and Other Hyperparameters

As mentioned in the introduction of Section 4.2, there exists a generalisation data set with systematic parameter changes and therefore we focus instead on changes to the data. Nonetheless, the data set we intend to create as part of future work should also consider a variety of hyperparameters. Part of the variety will arise naturally since to reduce our carbon footprint we aim to include as many pretrained models made publicly available as possible. Such models are published by a variety of authors, each with their own training procedure, typically optimised for achieving generalisation on particular data sets.

Additionally, we aim to train new models specifically for constructing the generalisation data set. Since the objective is to couple our evaluation with that of Jiang et al. (2021), for those architectures and data sets that both them and us consider, we will replicate roughly half of their settings while the other half will be used to explore new parameter configurations. This is to ensure variety in data-centric scenarios both within similar settings and outside of them. The precise values used will be determined when building the data set.

Note that although we plan to vary the hyperparameters, we will not generate a large body of models where a single hyperparameter value differs between them. Our aim is to encompass a high variety of final learnt representations with a focus on variations caused by changes in the data. The long-term vision, as we will outline in Section 4.3, is to create a much broader study that can evaluate models which use hyperparameter information. Therefore, where known, we plan to document and store information about the hyperparameters for this initial phase as well.

Unlike prior studies, we also aim to train models with a variety of loss functions. Studies such as that of Kornblith et al. (2021) and Müller et al. (2019) show that the choice of loss function impacts the learnt representations. Therefore, apart from the cross-entropy loss, we aim to experiment with both classical losses such as the mean squared error or hinge loss, as well as techniques designed to improve calibration such as focal loss (Mukhoti et al., 2020) or label smoothing (Szegedy et al., 2016).

Estimators that depend on the learning trajectory are outside the scope of our study. However, it would be interesting to evaluate estimators on instances "sampled" along that trajectory. For example, when training a model for 250 epochs, we could save model instances at the 100th and 200th epoch as well. Note that this is different from varying the number of epochs since in that case one starts from another initial state with each experiment. We believe it could be interesting to see if estimators could differentiate between the generalisation ability of two points along the training path. It must be noted that in order to then understand what the estimator can capture in this case, we would need to incorporate information about the training process, such as changes in learning rate, to be used for the analysis step. Therefore, although the estimators we evaluate do not use hyperparameter details, the information could be useful for better understanding the results obtained on the generalisation data set.

As mentioned in the introduction of Section 4.2, one of our main proposals is to obtain a large number of models which vary in their learnt representations. Although of primary interest are generalising models, we believe that it is important to aim for variety in learnt representations for models belonging to a wider spectrum on the underfitting–overfitting scale. This brings to the discussion the problem of randomising labels so as to force models to learn spurious rules, a commonly considered practice, especially in the memorisation literature.

Label Randomisation

Jiang et al. (2020) criticise the use of random labels for evaluating generalisation performance claiming that this leads to conclusions that are not representative of what one observes in practice. We do not entirely agree with this argument since it is often the case that real-world data is not perfectly labelled. However, Jiang et al.'s claim is justified considering that the metrics they evaluate stem from the bonding literature, which is centred around ERM. As discussed earlier, the choice of loss function impacts the learnt representations, with cross-entropy being known to lead to overconfident predictions (Müller et al., 2019; Guo et al., 2017). Thus the setting automatically becomes one of being influenced by the peculiarities of the chosen loss function. In such a case, randomising labels only exacerbates the level of overfitting.

Thus, we advocate for including the label noise scenario outside the zero-loss regime. Although both introducing label noise and testing until zero loss is reached lead to an overfitting regime, the learning dynamics are different and, as a result, the learnt representations are different, and we believe both must be studied. Thus, to avoid extreme overfitting we only perturb small percentages of labels for models trained for a moderate number of epochs. Note that the estimator and generalisation performance will be computed on the *original* data so that there is no distribution shift when evaluating the estimator. As we will detail later on, our generalisation data set will contain a number of separate, more difficult scenarios for researchers to experiment with. Given the disputed nature of the label randomisation scenario, we believe estimator's performance on this task should also be evaluated separately.

Regularisers

In practical scenarios it is common to rely on implicit or explicit forms of regularisation. Prior comparative studies typically treat regularisers as part of the ablatable parameters, considering weight decay or Dropout most often. As we have seen in Section 3.4.1.1, the frequently used method of mixed augmentation increases the complexity of the data, *acting as a regulariser*. Viewing data modification as a form of regularisation is important for the scope of the study.

Going back to the core assumption of many generalisation studies that the data is i.i.d., one could argue that modifying the training data violates this assumption. However, the estimators are computed on the original training data. The modifications are used merely for the training phase and, therefore, can be seen as another form of regularisation. As previously mentioned, in this study we are interested in capturing generalisation in its wider sense. For this reason, we strive to include overfitting as well as underfitting models and regularisers can help achieve this.

The complexity of the problem can more generally be altered through multiple forms of data modification, not necessarily augmentation. We next discuss the data sets and alterations we consider.

Data Sets

As mentioned in the first part of the thesis, we are strictly interested in evaluating computer vision models. This is the most commonly considered type of task in empirical generalisation studies. We aim to incorporate a wider variety of data sets than previously considered by each individual study. In particular, we aim to look at all the publicly available data sets evaluated in the studies we review. Namely, we include CIFAR-10/20/100 (Krizhevsky et al., 2009), Fashion MNIST (Xiao et al., 2017), SVHN (Netzer

et al., 2011), Oxford Pets (Parkhi et al., 2012), Oxford Flowers (Nilsback and Zisserman, 2008), CINIC-10. To these, we add Tiny ImageNet (Karpathy et al., n.d.), Bengali grapheme classification (Alam et al., 2021). Details on these data sets can be found in Section F. The major modification we propose compared to previous studies is altering the data, which we introduce below.

Varying the Structure of the Data Set

In this thesis we look at two types of data alterations: we can either modify individual instances according to some rule, or we can modify the number of classes or instances available. We refer to the latter as modifying the data set "structure".

We aim to experiment with changing the number of classes of a problem. Similar to the experiment presented in Chapter 3 where we remove the class "Truck", we create problem instances where one or more classes are removed, modifying the complexity, or separability, of the problem. Note that we remove the class from both training and test data.

Similarly, one can experiment with modifying the number of training samples available. Dziugaite et al. (2020) also consider this in their setting, although only do so for the CIFAR-10 and SVHN data sets. For removing samples, they randomly select the images to be removed. We aim to randomly remove samples from all the data sets we consider, as well as to remove specific subclasses from the data. For example, for those data sets where the distinction between tail and typical data is made available by Feldman and Zhang (2020), we can experiment with either removing the tail data or a percentage of the typical data.

Modifying Samples

So far we have argued that modifying the data is important for its regularising effect. We next informally argue that integrating data modification can expose limitations of previously proposed estimators of generalisation.

A number of estimators from the Predicting Generalisation in Deep Learning competition have measured sensitivity to MixUp-like distortions. More precisely, a lack of robustness to this distortion was taken as an indicator of reduced generalisation performance. We emphasise that none of the models in the competition have been trained on distorted data and remind the reader that the models achieve near-zero training loss. In these circumstances, it is possible that a lack of robustness to structured noise (i.e. MixUp perturbation) could indicate a stronger overfit. But is robustness to MixUp perturbation necessarily indicating a better generalising model? Is MixUp perturbation a sufficiently good way of measuring robustness of representations?

We argue that MixUp robustness is neither necessary nor sufficient for a model to have good generalisation performance. The former is because there is no particular reason why a model that is more robust to MixUp distortion would necessarily be better than one that is more robust to, say, occlusion. A clear example here can be constructed using the models trained in Chapter 3. On CIFAR-10, when evaluating the FMix model on MixUp-distorted training samples, we obtain a lower accuracy ($87.18_{\pm 0.19}$) than for the MixUp model ($90.17_{\pm 0.09}$). However, the accuracy on undistorted test data for FMix ($95.51_{\pm 0.10}$) is higher than that of MixUp ($95.15_{\pm 0.12}$). Therefore, the FMix model has a better generalisation performance than MixUp, although it is less robust to MixUp distortion. Moreover, we have argued earlier that it is sometimes difficult for humans to identify the objects depicted in images on which MixUp distortions have been applied. It is unclear why a model that perceives inputs in this interpolative manner would have better generalisation capabilities.

To argue that robustness to MixUp is not sufficient for a model to have good generalisation performance, we construct an overfitting model that is robust to MixUp distortions. To do this, we randomised the labels of the CIFAR-10 training set and used it to train a model *with MixUp augmentation*. We refer to this model as "the randomised MixUp model". We then evaluate the model on the randomly labeled training data without applying the MixUp distortion at evaluation time. we find that the randomised MixUp model achieves perfect accuracy on the *on the randomly labeled training data*. In other words, the model is perfectly fitting the mislabeled data. Yet, it is more robust to the MixUp distortion than some generalising models. For example, the basic model trained on the original CIFAR-10 data, has a *lower* accuracy on MixUp-distorted training data ($86.83_{\pm 0.03}$) than that achieved by the randomised MixUp model ($90.58_{\pm 00.32}$). As such, a model with no better than random test performance can be more robust to MixUp distortion than a generalising model. Therefore robustness to MixUp distortion on the training data is *not necessarily an indicator of a generalising model*. Note that similar results can be obtained for robustness to occlusion, by training randomised models with mask-based distortions such as CutMix and FMix. More specifically, just as we did in Section 3.3.2, we can obtain a model that completely overfits the training data and yet is still robust to patch-based distortions. Therefore robustness on training data to any one specific type of distortion is not necessarily an indicator of generalisation performance.

The randomisation example above, however, has two limitations. In practice, MixUp distortion has been used for the task of *ranking* models, and in conjunction with a notion of separability or compression of learnt representations (e.g. Natekar and Sharma, 2020; Lassance et al., 2020). Note that by randomly shuffling labels for one model and using the original labels for the other, we are not providing a fair one-to-one comparison. Our intent was rather to show that even completely overfitting models can be robust to MixUp (and other) distortions. We can very easily construct an example for the ranking case, where two models are evaluated on the same problem. To also couple this with a

notion of separability, we use the measure proposed by the winning solution (Natekar and Sharma, 2020) of the Predicting Generalisation in Deep Learning competition.

Natekar and Sharma (2020)'s DB Index, introduced in Section 4.1.5, captures geometrical properties of learnt representations that are generally associated with good separability *of the training data at a certain layer*. How does one choose that layer, especially when comparing models with different architectures? Natekar and Sharma (2020) choose to compute the DB Index on the first convolutional layer of the network for all architectures. However, we argue that this is uninformative. It has been shown that models can decrease separability in early layers and only monotonically increase it after the feature extraction phase (Tsitsulin et al., 2020). This is also supported by Ansuini et al. (2019), which find that neural networks initially expand the manifold of representations and it is the deeper layers where the increasing compression takes place. It is thus unclear why a better DB Index on the early representations would be indicative of generalisation. In fact, we argue that the separability of intermediate representations is not reflective of generalisation performance in general. Nonetheless, we compute this quantity to expose the inability of Natekar and Sharma (2020)'s approach to distinguish the generalisation performance of models trained under mixed augmentation.

We evaluate PreAct ResNet-16 models trained on CIFAR-10 with FMix and reformulated MixUp augmentation (see Chapter 3 for a description of the augmentations). For reference, we also provide results for the basic model, trained with no mixed augmentation. We compute the DB Index at the first convolutional layer, as in Natekar and Sharma (2020), as well as on the last convolutional layer. In both cases, we find examples of models that would be misranked by Natekar and Sharma (2020)'s estimator, as we will discuss below. This experiment provides a limited and simplistic example. We believe, however, that a large-scale comparison that includes a bigger variety of learnt representations will expose more such cases.

In Table 4.2 we give the generalisation performance of the models, along with their DB Index and accuracy on MixUp-distorted training data. We first note that the DB Index for the first layer is the same for all three models, which empirically confirms that this is not an informative quantity. We next focus on comparing FMix and reformulated MixUp. The test accuracy of FMix is higher than that of reformulated MixUp. Yet, the accuracy on MixUp-distorted train data is comparable for the two models. Therefore, using the DB Index and robustness to MixUp, one would be unable to distinguish the generalisation performance of these two models.

We also computed the DB Index for the last layer. Although this quantity is smaller on average for the FMix model, therefore indicating better generalisation performance, it is within the margin of error of the DB Index computed for reformulated MixUp. We argue then that although computing the DB Index on the last convolutional layer could be more informative, it still does not correlate sufficiently well with generalisation.

TABLE 4.2: Test accuracy, accuracy on training data distorted with MixUp-like modifications and DB Index on first and last convolutional layers for ResNet-18 models trained on CIFAR-10. The test accuracy is taken to be an estimation of the generalisation performance. A lower DB Index is associated with better separability and therefore better generalisation performance. Although the FMix and reformulated MixUp models have the same level of MixUp accuracy and the same DB Index for the first layer, FMix has a better generalisation performance.

	basic	FMix	reformulated MixUp
test accuracy	$94.52_{\pm 0.10}$	$95.51_{\pm 0.10}$	$93.39_{\pm 0.52}$
MixUp accuracy	$86.83_{\pm 0.03}$	$87.18_{\pm 0.19}$	$87.53_{\pm 0.62}$
DB first layer	$4.26_{\pm 0.05}$	$4.25_{\pm 0.03}$	$4.27_{\pm 0.02}$
DB last layer	$3.44_{\pm 0.03}$	$2.98_{\pm 0.04}$	$3.05_{\pm 0.03}$

We conclude that robustness to MixUp distortion, even when coupled with notions of separability of learnt representations, is a limited indicator of generalisation performance. Although we have only presented one example, we believe creating a generalisation data set that is more diverse will further expose the weak correlation between robustness to MixUp distortion and generalisation performance.

We therefore aim to integrate a variety of sample modifications when creating the generalisation data set. Among the modifications we would consider are those used in Chapter 3: the mixed data augmentations (i.e. MixUp, CutMix, FMix), CutOut, Hide-and-seek and even patch-shuffling. To these we add commonly used augmentations such as Gaussian noise, flip, random crop, colour jitter, shear, rotation, etc.

In the case of mixed augmentations, to obtain a greater variety of representations we will experiment with different ways of choosing the mixing ratio. We will also experiment with mixing multiple augmentations during training as it is done in Harris et al. (2020), where augmentations are alternated between batches. Harris et al. claim that combining interpolating and masking augmentations gives improved generalisation performance, leveraging the benefits of both types of augmentations. We also plan to include models trained in the reformulated regime and models trained with inter-data set mixing.

Evaluation Criteria

Most commonly, models are evaluated in a comparative manner (e.g Dziugaite et al., 2020; Vakanski and Xian, 2021; Jiang et al., 2020). That is, models are *ranked* according to their generalisation gap. The previous studies use different variants of the Kendall ranking coefficient (Kendall, 1938), also known as Kendall's-τ, as evaluation criteria. When applied to the estimator evaluation problem, the standard Kendall ranking coefficient gives the correlation between the ranking of each pair of models and the ranking of their generalisation performance. Informally, when a model A has a higher generalisation performance than a model B, we would like the estimator to also predict a higher value for model A than for model B. This can be measured by the Kendall's-τ.

Jiang et al. (2020) bring to the reader's attention the scenario in which a measure can correctly capture the effect of a single hyperparameter change, but not perform so well when multiple hyperparameters are changed. To address such scenarios, they propose a variant of Kendall's coefficient, the *granulated* Kendall coefficient. They consider each hyperparameter at a time. Fixing a value for the considered hyperparameter, one computes the average Kendall coefficient for *all combinations of other hyperprameter values*. Averaging this for all possible values of the considered hyperparameter, one obtains the coefficient of that hyperparameter variable. Lastly, averaging the hyperparameter coefficients across all hyperparameters gives the granulated Kendall coefficient. Jiang et al. (2020) emphasise that this proposition cannot capture true causality. They propose conditional independence testing for those measures that exhibit a correlation with generalisation performance. Dziugaite et al. (2020) criticise Jiang et al. (2020)'s independence test, arguing it is limited for determining causality.

Instead, Dziugaite et al. (2020) propose a worst-case comparison of generalisation estimators across a fixed experimental setting (i.e. fixed values for each hyperparameter, a choice of architecture, dataset, optimiser, etc.), referred to as an *environment*. They draw inspiration from the field of distributional robustness, therefore calling their evaluation *robust ranking*. They are interested in finding the robust error of estimators, which is the supremum expected loss of the estimator over a family of environments. This is a more challenging setting and, in our opinion, a more relevant one for the task of identifying the true mechanism behind generalisation. Note that in their loss function Dziugaite et al. (2020) weight the ranking coefficient by the difference in generalisation performance between the ranked models.

Dziugaite et al. (2020) consider both the task of ranking two models and the task of predicting the test accuracy of a specific model. For the former, they use as a metric the above-mentioned robust ranking. For the latter, they once again take the supremum over a family of environments of the mean squared error. They only briefly present these results in their supplementary material.

Other Design Choices

Dziugaite et al. (2020) discard those pairs of models for which the generalisation performance does not differ more than a specified threshold. In a similar vein, Jiang et al. (2020) concluded that including models trained from the same initialisation is not computationally justifiable for such large-scale experiments. Jiang et al. (2021) then adopt the same view. Since we will include previously trained models to reduce the computational burden, we choose to have repeats in our data set where they exist. However, as this significantly increases the complexity of the generalisation estimation problem, we will provide the repeats as a separate task so that practitioners can choose whether or not they want to experiment with them.

Given the setting considered for empirical predictors of generalisation, a successful predictor should be able to estimate the performance of any model, regardless of how that model was obtained. Therefore, following the procedure employed in Chatterji et al. (2020) to test for network criticality, we could assess a predictor's ability to account for changes in the weights that are not caused by the learning algorithm. We would include a number of scenarios: modifying critical layers, modifying layers that are not critical, and marginally modifying random weights that do not necessary belong to the same layer.

Summary

The aim of the proposed study designed in this chapter is to capture an estimator's ability to reflect the generalisation performance of models trained in a variety of settings. The emphasis falls on obtaining a large pool of learnt representations and less on the precise impact of the optimisation hyperparameters. We aim to incur as low a computational and environmental cost as possible. For our measure to be relevant to practitioners, we aim to incorporate a larger number of real-world data sets. To further minimise our environmental impact we choose to reuse the models trained in previous chapters and couple these with trained models from other sources.

4.3 Future Work

As mentioned in the introduction of this chapter, our initial goal was to create a data set for a large-scale, data-centric evaluation of empirical estimators. However, to rigorously do this we must train a large number of models exploring combinations of the scenarios we have described in our proposed setting. This would allow us to apply the granulated Kendall coefficient to evaluate estimators, using Dziugaite et al. (2020)'s methodology.

Thus, an immediate next step is to create the data set; that is to train a variety of models. We have already created scripts for each individual scenario. We only need to integrate them in a single pipeline, and most importantly, we need the time to train the models. The intention is to make this data set publicly available and easy to use for future research. Once we have the data set, evaluation should be straightforward. For this, we would use the code we have adapted from Jiang et al. (2021)'s competition.

We would then do a case-wise analysis to identify the precise scenarios in which prior estimators fail to correlate well with generalisation performance. This will provide supporting empirical evidence for the limitations that we have highlighted in this chapter.

Once the limitations have been thoroughly explored and documented, the next step is to address them. In this regard, we have two possible initial directions that will hopefully be further refined by the insights we will gain from the initial analysis. Both of these directions are informally based on the Information Bottleneck Theory. We remind the reader that the issue with the Intrinsic Dimension approach proposed by Ansuini et al. (2019) is that it does not account for the mutual information between the learnt representations and the target variable. So, could we then devise a task-aware notion of Intrinsic Dimension? One idea we briefly explored was to project the learnt representations onto the class vectors. This, however, raises the problem of finding a principled way to learn this projection. We believe care must be taken in order to ensure we are measuring something meaningful.

The second direction is to design a more exhaustive notion of learnt representation robustness. Let us go back to our simple counter-example against the predictiveness of Intrinsic Dimension alone (see Section 4.1.1). We presented the hypothetical case of a binary classification problem where a collection artefact such as a timestamp is present on all training samples of a class. Such cases are known to occur in practice, and we have given the real-life incident with chest X-ray scans as an example. We argue that in such cases, the missing ingredient is a notion of robustness of learnt representations. Abusing terminology, we can think of notions of robustness as informal indicators of the mutual information between the learnt representations and the true target variable. In a sense, a model that is not robust to natural perturbations cannot have a satisfactory level of mutual information between its representations and the target variable, since it is relying on spurious information. Although related, note that the robustness to the

perturbations we are looking for is different from determining robustness against adversarial perturbations at large. This is because adversarial robustness is sometimes looking for "low probability pockets" in the data distribution. Like Novak et al. (2018), we aim for a study of the *expected generalisation performance* rather than the performance on a specific and possibly peculiar small set. This raises further questions: **Which perturbations are "natural" and diverse enough to capture a meaningful notion of robustness? Could our breakdown of distortion into local and global help us explore perturbation diversity in a more informative way?**

The main limitation of the directions we propose is that they are not straightforward; they often imply solving questions related to generalisation that are themselves known to be difficult. A more simplistic approach could be to try to build a linear "meta-model" to predict the generalisation performance. Note that this would turn the problem from a ranking one to a regression one. The meta-model could take estimates from a number of predictors associated with a model instance as input and output its generalisation performance. We could train this model to learn the relationship between various predictors and generalisation performance. The reason behind doing this is that it is possible that no predictor is indicative of generalisation but a *combination* of predictors could correlate better with generalisation. One data set design detail that will need to be established for this is the train-test split. Naturally, we would like the test data to be as challenging as possible while remaining realistic. We would therefore like to save a number of scenarios for the test data alone. This raises more questions: **What is a principled way of choosing the scenarios saved for test time? How do we ensure that newly proposed methods will not "overfit" to these scenarios?** These are very important issues for the validity of the analysis and we are as yet unsure what the right answer is. The hope is that once a good meta-model is created, we can create stronger proposals by looking for the most relevant features and investigating how existing notions of representation quality interact.

We also aim to expand our study, both in terms of the tasks and the setting we consider. With regard to the former, we would like to also include models trained on language or audio problems. In terms of the setting, we plan to consider estimators that require more information about the given problem than the learnt model instance and the training data alone. We will discuss adding more information about the data in the Future Work section of Chapter 5, as the answer is part of our larger vision on a data-centric future for generalisation studies. Apart from the data, we could consider incorporating the hyperparameters used during training and the initial state of the model.

Lastly, when defining the scope of our thesis we explicitly ruled out the learning process and focused on a learnt model instance. But learning dynamics also play an important part in understanding deep learning. **Can we get a better intuition of what are the important qualities of learnt representation by analysing their evolution**

throughout the learning process? More importantly, is the setting that discards learnability relevant in practice? In other words, must an estimator correctly predict the generalisation of peculiarly trained models that would never be learnt using standard training? Our personal answer to this question is that a generalisation estimator must be complete and account for all learners in the model class, even those that might not be learnt in conventional ways. However, it is possible that this is a naively optimistic goal whose validity remains an open question.

4.4 Conclusions

Despite the various attempts to capture generalisation, the field seems to be far from solving this fundamental puzzle. In this thesis we expressed the belief that stronger intuitions are needed in order to propose a theoretical framework able to explain generalisation in a way that is meaningful and informative. We believe that practical experimentation can help build the required intuitions. Therefore in this chapter we started building the basis for an empirical approach to furthering the community's understanding of generalisation.

To form a coherent picture of what the promising future directions are, we first reviewed and contextualised the broad evolution of ideas in generalisation studies. We noted that most studies fundamentally aim to capture separability and compression of learnt representations. The questions that remain to be addressed are centred around separability and compression. **How do we meaningfully capture these? If they can be captured, are they necessary and sufficient for generalisation?**

In our review we looked more closely at empirical predictors, which are part of a new and rising subfield of generalisation studies. The novelty of this subfield, however, comes with fairly limited standards of evaluation. In our opinion, this has caused approaches that are not well-founded, to appear successful. We motivated our claims at a conceptual level but also brought preliminary empirical evidence to support our view.

We argued the need for more extensive and rigorous evaluation and analysis of empirical estimators. Prior art focuses only on capturing those differences in model performance caused by changes in the hyperparameter values. Following the broad theme of this thesis, we advocated for the integration of a *data-centric set of experiments*. We therefore proposed and laid the initial design for a new generalisation data set on which to evaluate estimators. The emphasis falls on varying the data set structure and including data distortions as alternative ways of regularising models. Creating this data set remains part of future work.

Chapter 5

Closing Remarks and Future Directions

This chapter briefly recounts the high-level contributions of our work. We review the objectives of the thesis and how they were achieved. We then revisit our vision for the future of generalisation studies and acknowledge the factors which have shaped it. We discuss limitations and a broad range of future work. Lastly, we briefly reflect on the learning experience of writing this thesis.

As outlined in the Foreword, this thesis is a modern *exploration* of generalisation in machine learning. Undoubtedly, there are many more details of this complex problem that are worth mentioning than we have managed to cover. Our intent was to provide a high-level perspective on generalisation and then focus on the specifics of those topics that we see as most important.

One of the objectives of the thesis was to advance our current understanding of generalisation. We started from the perspective of classical theory. Engaging at a deep level with this theory, we slowly developed the belief that data plays a much more fundamental role than has been attributed before, and that the field lacks sufficient insight from practical experimentation.

These beliefs shaped the idea of constructing a data-centric approach to studying generalisation in the future. With this intent in mind, we focused on analysing data modification. We discovered that although data modification is extensively used, it is often poorly understood. We reflected on the regularising role of data modification and raised a number of questions around distribution shift. Although time constraints meant we did not go into further detail in this thesis, we believe that studying the effect of data modification plays an important role in understanding generalisation at large and we would like to study it in more depth in the future.

During our analysis of data modification, the idea of empirically predicting the generalisation performance in settings where the data is varied has started to form. Concurrently, hyperparameter-centric empirical studies started to appear in the literature. We therefore aimed to understand how other researchers also converged to this empirical direction. As a result, we started reconstructing the evolution of ideas in generalisation studies. We presented, classified, and contextualised the main lines of thought, with a focus on the most recent direction of empirically estimating generalisation. This has given us a sense of the recurring ideas in the different generalisation subfields and refined the beliefs we had prior to this extended literature review. More importantly, it has now enabled us to construct an informed vision for the future of generalisation studies.

Our data-centric perspective, coupled with insights gained from analysing data modification, has allowed us to identify limitations of the previously proposed estimators and their evaluation. This further motivated the need for the data-centric view we propose.

Therefore, throughout the thesis we explored a number of topics, each requiring different tools and research methodologies. For example, to show the importance of the data, in Chapter 2 we performed a theoretical analysis. This was followed by extensive empirical experimentation in Chapters 3 and parts of Chapter 4, where we carry out various quantitative evaluations. Lastly, Chapter 4 provides a rich review and a qualitative analysis of prior evaluation settings for empirical estimators. As such, our main concrete contributions are:

Chapter 2:

- Under the annealed approximation we propose the β-Risk model for classification (Section 2.2.1);
- We validate our model on learning with a perceptron (Section 2.3.1); and,
- Our calculations emphasise the need to account for the data (Section 2.2.1).

Chapter 3:

- We identified the presence of data interference when modifying data (Section 3.2);
- We introduced a quantity for detecting data interference (Section 3.2);
- We identified three limitations of measuring robustness through the most popular method currently used in the field (Section 3.3.2);
- We proposed iOcclusion, a method of measuring robustness to occlusion that addresses the limitations of prior art (Section 3.3);
- We designed and carried out a series of experiments to support our proposed method (Section 3.3.2); and,
- We showed that, contrary to what is currently assumed, an augmentation that better preserves the data distribution is not necessarily better (Section 3.4).

Chapter 4:

- We provide an overview of the main lines of thought in generalisation studies (Section 4.1.1);
- We give, to the best of our knowledge, the first literature review of empirical estimators of generalisation (Section 4.1.1);
- We argue that estimators previously considered successful are limited in their ability to capture generalisation (Sections 4.1.4 and 4.2.2);
- We overview prior settings for evaluating empirical estimators (Section 4.2.1); and,
- We design a new evaluation for empirical estimators (Section 4.2.2).

During the course of this work we also contributed to creating FMix (Harris et al., 2020), a new mixed sample data augmentation (MSDA). This work started as an attempt to understand the success of MixUp (Zhang et al., 2018a). At the time, our reasoning about generalisation was deeply rooted in the statistical learning paradigm. MixUp breaks two fundamental assumptions: that the samples are independent; and that the train and test samples are drawn from the same distribution. Yet, breaking these assumptions leads to *better generalisation*. We, therefore, started exploring augmentation. We studied the Vicinal Risk Minimisation framework, existing justifications for the success of MSDA, and other mixed augmentations such as CutMix. This has further shaped our data-centric vision for understanding the mechanism behind generalisation and further motivated us to take a practical approach.

We would lastly like to reflect on the inspiring experience of presenting our work at peer-reviewed venues, and how this shaped our research direction and ideas for future work. As we already discussed our contributions above, we will focus here on the reception of our work and fruitful discussions with fellow researchers.

Rethinking Generalisation at the NeurIPS 2019 Workshop on Machine Learning with Guarantees:

> Based on the work presented in Chapter 2, the paper proposed the β-Risk model for classification and the realisable perceptron. The central message of the paper was the importance of *attunement*. Our calculations contrasted with the bound-centric theme of the workshop. As explained throughout the thesis, we believe useful guarantees are not yet attainable with the field's current level of understanding. Although we discussed possible ways of starting to reason about attunement with the audience, a concrete direction had not caught form at the time. Nonetheless, we were encouraged to see that one year later, as part of the same conference, the Predicting Generalisation in Deep Learning competition was organised. This

signalled that the field was starting to take a more practical approach to capturing generalisation, which was in line with our belief.

On data-centric myths at the NeurIPS 2021 Workshop on Data-Centric AI:

This paper provides a different perspective to the results presented in Section 4.1.4 and Section 3.4. Motivated by the findings on data modification side-effects (Chapter 3) we wanted to find *qualities of the data and ways of quantifying them*. Since the literature on this topic is limited, we wanted to identify intuitions about data quality from various fields. We first looked at empirical predictors of generalisation and came across the idea that learning low-dimensional representations was associated with better generalisation (Ansuini et al., 2019). Reversing the perspective, then a data set that lies on a low-dimensional manifold should be desirable. Using the same reasoning as in Section 4.1.4, we show that this is not the case. Then, as in Section 3.4 we argue that the data distribution doesn't necessarily have to be preserved when modifying samples.

During an interesting discussion, we were introduced to a new perspective on learning which we would like to come back to in the future. Vieira et al. (2022) argue that the whole learning process is simply given by iterative manipulations of the data manifold and therefore analyse the learning dynamics from a vector field perspective. More precisely, they see data points as particles moved around until a configuration where they can be separated is reached. Although Vieira et al. (2022) focused on proposing a new architecture starting from this perspective, they believe that interesting insights on generalisation could be gained from this perspective. Similarly, we were encouraged to see that the authors of this work agreed with our vision that intrinsic dimension needs to be coupled with good notions of robustness in order to obtain meaningful results.

On pitfalls of measuring occlusion robustness through data distortion at the International Conference on Learning Representations Workshop on RobustML:

This paper shows the limitations of classical robustness evaluations and proposes a fairer alternative. We introduce the DI index (Section 3.2) and iOcclusion (Section 3.3) and construct a number of experiments to empirically demonstrate that our method addresses the limitations of CutOcclusion. The idea of extending our observations to audio data has emerged while presenting our work. We reflected along with our audience on what are the artefacts of occluding audio signals and how one would approach fairer alternatives in this case. We were enthusiastic about the impact of our work in other domains and would like to explore this more in the future.

On the effects of artificial data modification at the International Conference on Machine Learning:

> This paper presents a concise version of Chapter 3. We were encouraged to discuss with practitioners and theoreticians alike who found our work relevant. On the practical side, fellow researchers were particularly interested in insights about data augmentation as well as discovering which attributes of the data matter. On the theoretical side, researchers were interested in our vision for a data-centric future of generalisation studies. It was during these discussions that we understood the potential impact of writing an overview of ideas in the various subfields of generalisation.

During the creation of this thesis we have acquired many partial intuitions and insights. However, in our opinion, the most valuable outcome is shaping our future research direction: building a new, more flexible framework for reasoning about generalisation. As such, the overall understanding we have built is that practical experimentation has a great potential to solidify and correct our beliefs regarding the true mechanism behind generalisation. Researchers are approaching generalisation in various interesting ways. However, despite the paramount importance of the data, a data-centric view of generalisation still seems to be missing.

The vision that was shaped while creating this thesis is to build a new theoretical framework starting from data-centric practical experimentation. The present work is merely the birth and introduction of what is likely to be a years-long endeavour. Below we briefly sketch the ideas we would like to pursue in the future to achieve this goal.

5.1 Future Work

At the end of each chapter so far we have outlined a number of concrete directions that we would like to explore in future. We mostly dedicate this section to our perspective on building a data-centric future for generalisation studies.

From Practice to Theory

The plan outlined in Chapter 4 is to find an estimator or a combination of estimators that correlates well with generalisation. As previously motivated, a good estimator should also be able to capture variations in the learnt representations caused by changes in the data. Once the difficult task of finding a good estimator has been achieved, the next step would be to formalise the understanding that is gained. This will allow us

to more rigorously reason about generalisation. The hope is to be able to provide the practical guarantees required for determining trust in the machines used in sensitive applications. At this stage, it is difficult to draw a more precise set of steps for bridging theory and practice. However we believe it is important to keep this long-term goal in mind both because it provides a strong motivation, but also because it constantly steers our attention towards the core problem we are trying to address.

The Data

Our main criticism of prior approaches was their inability to account for the data. However, we believe there is a question on the extent to which our proposed direction could account for the data to a sufficient level. As mentioned in Chapters 1 and 3, it is often the case that real-world data is more problematic than the standard data sets in vision applications. These could range from relatively small deviations from the training data distribution to very significant ones. While the significant ones might be easier to spot in some cases, and therefore the problem is identified from the start as an "out-of-distribution" one, it is possible that milder cases could be treated as "in-distribution", just as in the chest X-Ray example from Chapter 1. Ideally, we would like an estimator to be able to address these cases. The question is whether it would be possible to do so in our original setting. Although peculiar cases, such as the chest X-Ray problem, could possibly be addressed by incorporating better notions of robustness, there is a question on the extent to which an estimator should be expected to perform even in "mild" out-of-distribution settings. Therefore, given that our aim is to create a framework that is *relevant in practice*, we might have to account for distribution differences.

In our view, there are two possible ways to deal with the very stringent assumption about the data. In the a priori view, one could integrate notions of data quality. In an a posteriori world, a solution would be to use measures of distribution similarity. We will discuss these next.

A Priori – Data Quality

A question that we have faced while researching prior art and creating our proposal for the generalisation data set is: **Can we even predict generalisation without notions of data quality? Could data quality be implicitly captured in the learned representations or do we need to explicitly account for it?** We do not know what the desirable attributes of a data set are. Nor do we how one would start reasoning about data quality in a formal manner. Intuitively, we would initially consider notions of problem complexity or separability. However, addressing these problems usually boils down to solving problems highly related to those of generalisation. For example, Zhang et al. (2017)'s label randomisation experiment has shown that even

ImageNet with randomised labels, a highly complex problem, can be separated. To address this, one could couple separability with the complexity of the function that is separating it or with its margin, therefore directly relating this problem to generalisation. However, we would be interested in researching alternatives that do not imply solving generalisation first.

A Posteriori – Distance Between Distributions

An alternative way of accounting for mild distribution changes could be to provide *unlabelled* test data and use the distance between the representations of the train and test data as a proxy for the *perceived* distribution distance. We do not know to which extent this method would allow us to detect the magnitude or even the presence of minor shifts. One possible avenue would be to measure the representation similarity using established methods such as canonical correlation analysis (Kornblith et al., 2019). The issue here, however, is that the perceived distance is dependent on the quality of the model. Modern-sized networks can make confident mispredictions (e.g. Guo et al., 2017). Therefore, it is possible that the perceived distance is not reflective of the expected generalisation gap for weak models.

Quantifying changes in the generalisation performance based on distributional distance has been attempted in digital pathology (Stacke et al., 2020). For each sample in the set, Stacke et al. compute the *mean value* of every feature map in a hidden layer. They then measure the distance between the distribution of those mean values on the train versus test data. However, Stacke et al. (2020) do not justify why they choose to look at the average values of the feature maps. They find that their notion of representation similarity can sometimes correlate with drops in the generalisation performance but their empirical evaluation is very limited. Although Stacke et al. (2020)'s notion could indicate broad distribution shifts, the problem of representation similarity is very challenging (Morcos et al., 2018) and subtle. We believe a more informative quantity would be needed to capture the finer-grained differences that are needed to predict the magnitude of the change in generalisation performance.

Note that there exist model-agnostic alternatives for measuring distributional distance, such as those proposed in the domain adaptation field (e.g. Alvarez-Melis and Fusi, 2020). However, these model-agnostic notions of distribution distance cannot necessarily inform on the impact the shift will have on the generalisation performance *of each model*, which is what we are interested in. Therefore, they do not represent a viable alternative in our case.

Widening the Perspective

Although the problem of generalising to out-of-distribution settings is currently treated as a separate field, we believe it is relevant to ask whether the insights gained from the in-distribution generalisation studies would be relevant to this related field and vice versa. There are two implications here, which we discuss separately below.

Can a theory for out-of-distribution generalisation solve in-distribution generalisation? Solving the former would theoretically solve the latter as well since in-distribution generalisation should be a special case of the more general scenario. However, the effort in the out-of-distribution literature is mostly invested in making models generalise across distribution shifts rather than capturing and formalising generalisation. In other words, the interest is mostly in achieved performance, rather than building a theoretical framework or understanding. Moreover, the tools used to advance out-of-distribution performance are rather different from the ones we discussed in this thesis, which are employed in generalisation studies. Therefore, bridging the two fields is not immediate.

We believe there is a subtle but important difference between in-distribution and out-of-distribution in terms of causality. In the in-distribution setting, one could benefit from leveraging contextual information. However, in an out-of-distribution setting, one would need to more closely identify true causality to ensure generalisation. Models that are highly robust to occlusion because they also learn contextual information, such as FMix and CutMix, are generalising better in an i.i.d. setting than the basic model, which uses less information overall (Chapter 3). Therefore, we would like an in-distribution notion of model quality to capture this increase in generalisation performance, while a good out-of-distribution notion would likely penalise models that employ non-causal relationships. It is, however, likely that there are many confounding factors in the example above. Therefore, the feasibility of bridging the two settings remains to be determined.

Can a theory for in-distribution generalisation solve out-of-distribution generalisation? We believe that scaling a theory for i.i.d. generalisation to an out-of-distribution one is more challenging than the reverse. However, we believe this direction has a significant head-start, with many active efforts invested in capturing and formalising generalisation. We believe drawing inspiration from the work done in the out-of-distribution field could be beneficial for building stronger intuitions. Importantly, it can also motivate the community to reason about and aim for a unified theory.

We therefore plan to familiarise ourselves with the literature on out-of-domain generalisation and try to establish if and how the two can be connected. Although making connections with the literature on out-of-distribution generalisation is the most immediate proposal, we believe in a wider cross-pollination of ideas across fields. In essence,

many of the studies proposing advances in machine learning provide some informal justification for their success. It would be interesting to collect and classify them to identify what is currently believed outside of the generalisation literature to be predictive of a model's performance.

Community Engagement

We would also like to raise the field's awareness with respect to the importance of better understanding generalisation. A wider engagement with the subject can ensure better scrutiny of the proposed methods and can accelerate the pace of advancement. Although the generalisation field is seeing an increase in interest, we believe this is insufficient given the importance of the problem. Pushing the state-of-the-art performance is challenging, but we argue that finding the fundamental reason for the achieved improvement has a much higher impact in the long term.

We believe that being able to properly understand and guarantee the performance of models could revolutionise the modern world, with applications in healthcare, transportation, renewable energies, and many more. A good understanding would also lead to much faster development and a reduced environmental cost for model training. Therefore it is important for the community to address this problem and we would like to motivate more researchers to join us in this quest.

In summary, studying generalisation remains a highly challenging problem. Although bounding has been the default way of approaching it, the paradigm is starting to change. There is a broad range of possible ways to approach generalisation estimation and we believe seeking inspiration outside of the established generalisation literature has great potential.

We are encouraged to see that the ideas we had while working on this thesis are starting to emerge independently of our work. They have given us the enthusiasm to pursue the directions we believe in. Specifically, we were motivated by the proposal of the Predicting Generalisation in Deep Learning competition and the distortion-centric methods which emerged as well as by the organisation of the first Data-Centric workshop, in which we participated. We were also encouraged by the reception of our work at conferences and workshops, and are excited and highly motivated by the large number of future directions that can be pursued.

5.2 Final Reflections

In this thesis we have, in turn, computed the risk in a classical theory-like framework, discussed model attributes evaluation, data augmentation, the evolution of ideas in generalisation studies, empirical predictors, and the evaluation of generalisation estimators.

We did not go into too much detail for any of these topics. This is due to our search for a direction that has the potential to provide *meaningful insights*.

Reflecting back on this wide search, we believe it has greatly helped us quickly adapt to new perspectives, ideas, and tools. We see this as an important learning outcome at this stage of our research journey. We believe it has given us the openness to embrace new research directions, the ability to critically analyse ideas from a high-level perspective and, overall, started shaping us as well-rounded researchers. We hope that the flexibility acquired will facilitate collaborations with experts from a variety of fields in the future, which is part of our vision.

The motivation of all directions we pursued remained deeply anchored in the original objective, which was to understand generalisation. We hope our work will help the community solve this complex and fascinating problem, whose solution we believe will reshape the modern world.

References

Achille, A. and Soatto, S. (2018), 'Emergence of invariance and disentanglement in deep representations', *Journal of Machine Learning Research* **19**(1), 1947–1980.

Adlam, B. and Pennington, J. (2020*a*), The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization, *in* 'International Conference on Machine Learning', PMLR, pp. 74–84.

Alaiz-Rodríguez, R. and Japkowicz, N. (2008), Assessing the impact of changing environments on classifier performance, *in* 'Conference of the Canadian Society for Computational Studies of Intelligence', Springer, pp. 13–24.

Alam, S., Reasat, T., Sushmit, A. S., Siddique, S. M., Rahman, F., Hasan, M. and Humayun, A. I. (2021), 'A large multi-target dataset of common Bengali handwritten graphemes', *International Conference on Document Analysis and Recognition* pp. 383–398.

Alquier, P. (2021), User-friendly introduction to PAC-Bayes bounds, *in* 'arXiv preprint arXiv:2110.11216'.

Alvarez-Melis, D. and Fusi, N. (2020), Geometric dataset distances via optimal transport, *in* 'Advances in Neural Information Processing Systems', pp. 21428–21439.

Amari, S., Ba, J., Grosse, R. B., Li, X., Nitanda, A., Suzuki, T., Wu, D. and Xu, J. (2021), When does preconditioning help or hurt generalization?, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=S724o4_WB3*

Amsaleg, L., Bailey, J., Barbe, D., Erfani, S., Houle, M. E., Nguyen, V. and Radovanović, M. (2017), The vulnerability of learning to adversarial perturbation increases with intrinsic dimensionality, *in* '2017 IEEE Workshop on Information Forensics and Security (WIFS)', IEEE, pp. 1–6.

Ansuini, A., Laio, A., Macke, J. H. and Zoccolan, D. (2019), Intrinsic dimension of data representations in deep neural networks, *in* 'Advances in Neural Information Processing Systems', p. 6111–6122.

Anthony, M. and Bartlett, P. L. (1999), *Neural Network Learning: Theoretical Foundations*, Cambridge University Press. ISBN 9780511624216.

Arora, S., Ge, R., Neyshabur, B. and Zhang, Y. (2018), Stronger generalization bounds for deep nets via a compression approach, *in* 'International Conference on Machine Learning', PMLR, pp. 254–263.

Arpit, D., Jastrzebski, S., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A., Bengio, Y. et al. (2017), A closer look at memorization in deep networks, *in* 'International Conference on Machine Learning', PMLR, pp. 233–242.

Ba, J. L., Kiros, J. R. and Hinton, G. E. (2016), Layer normalization, *in* 'arXiv preprint arXiv:1607.06450'.

Bahri, D. and Jiang, H. (2021), Locally adaptive label smoothing improves predictive churn, *in* 'International Conference on Machine Learning', PMLR, pp. 532–542.

Bai, Y., Yang, Y., Zhang, W. and Mei, T. (2022), Directional self-supervised learning for heavy image augmentations, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 16692–16701.

Balcan, M.-F. and Berlind, C. (2014), A new perspective on learning linear separators with large l_ql_p margins, *in* 'Artificial Intelligence and Statistics', PMLR, pp. 68–76.

Balestriero, R., Pesenti, J. and LeCun, Y. (2021), Learning in high dimension always amounts to extrapolation, *in* 'arXiv preprint arXiv:2110.09485'.

Banerjee, P. K. and Montúfar, G. (2021), Information complexity and generalization bounds, *in* '2021 IEEE International Symposium on Information Theory (ISIT)', IEEE, pp. 676–681.

Barratt, S. and Sharma, R. (2018), A note on the inception score, *in* 'arXiv preprint arXiv:1801.01973'.

Bartlett, P., Freund, Y., Lee, W. S. and Schapire, R. E. (1998a), 'Boosting the margin: A new explanation for the effectiveness of voting methods', *The Annals of Statistics* **26**(5), 1651–1686.

Bartlett, P. L. (1998b), 'The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network', *IEEE Transactions on Information Theory* **44**(2), 525–536.

Bartlett, P. L., Foster, D. J. and Telgarsky, M. J. (2017), Spectrally-normalized margin bounds for neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 6240–6249.

Bartlett, P. L. and Mendelson, S. (2002), 'Rademacher and gaussian complexities: Risk bounds and structural results', *Journal of Machine Learning Research* **3**(Nov), 463–482.

Bartlett, P., Maiorov, V. and Meir, R. (1998c), Almost linear VC dimension bounds for piecewise polynomial networks, *in* 'Advances in Neural Information Processing Systems', pp. 190–196.

Bartlett, P. and Shawe-Taylor, J. (1999), Generalization performance of support vector machines and other pattern classifiers, *in* 'Advances in Kernel Methods—Support Vector Learning', pp. 43–54.

Baum, E. B. and Haussler, D. (1989), What size net gives valid generalization?, *in* 'Advances in Neural Information Processing Systems', pp. 81–90.

Belcher, D., Marcu, A. and Prügel-Bennett, A. (2022), Generalisation and the risk–entropy curve, *in* 'arXiv preprint arXiv:2202.07350'.

Belharbi, S., Sarraf, A., Pedersoli, M., Ayed, I. B., McCaffrey, L. and Granger, E. (2021), F-CAM: Full resolution class activation maps via guided parametric upscaling, *in* 'arXiv preprint arXiv:2109.07069'.

Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019), 'Reconciling modern machine-learning practice and the classical bias–variance trade-off', *Proceedings of the National Academy of Sciences* **116**(32), 15849–15854.

Belkin, M., Ma, S. and Mandal, S. (2018), To understand deep learning we need to understand kernel learning, *in* 'International Conference on Machine Learning', PMLR, pp. 541–549.

Bietti, A. and Mairal, J. (2019), On the inductive bias of neural tangent kernels, *in* 'Advances in Neural Information Processing Systems', pp. 12893–12904.

Biggs, F. and Guedj, B. (2022), On margins and derandomisation in PAC-Bayes, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 3709–3731.

Bińkowski, M., Sutherland, D. J., Arbel, M. and Gretton, A. (2018), Demystifying MMD GANs, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=r1lUOzWCW*

Blumer, A., Ehrenfeucht, A., Haussler, D. and Warmuth, M. K. (1989), 'Learnability and the Vapnik-Chervonenkis dimension', *Journal of the ACM (JACM)* **36**(4), 929–965.

Boucheron, S., Bousquet, O. and Lugosi, G. (2005), 'Theory of classification: A survey of some recent advances', *ESAIM: probability and Statistics* **9**, 323–375.

Bousquet, O. and Elisseeff, A. (2002), 'Stability and generalization', *Journal of Machine Learning Research* **2**, 499–526.

Brand, M. (2002), Charting a manifold, *in* 'Advances in Neural Information Processing Systems', p. 985–992.

Breiman, L., Friedman, J., Stone, C. J. and Olshen, R. (1984), *Classification and Regression Trees*, CRC Press. ISBN 9781138469525.

Brendel, W. and Bethge, M. (2019), Approximating CNNs with bag-of-local-features models works surprisingly well on imagenet, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=SkfMWhAqYQ*

Carbonnelle, S. and De Vleeschouwer, C. (2020), Intraclass clustering: an implicit learning ability that regularizes dnns, *in* 'International Conference on Learning Representations'.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B. and Tommasi, T. (2019), Domain generalization by solving jigsaw puzzles, *in* 'CVPR'.

Carratino, L., Cissé, M., Jenatton, R. and Vert, J.-P. (2022), 'On mixup regularization', *Journal of Machine Learning Research* **23**(325), 1–31.
URL: *http://jmlr.org/papers/v23/20-1385.html*

Carrell, A., Mallinar, N., Lucas, J. and Nakkiran, P. (2022), The calibration generalization gap, *in* 'arXiv preprint arXiv:2210.01964'.

Chapelle, O., Weston, J., Bottou, L. and Vapnik, V. (2001), Vicinal risk minimization, *in* 'Advances in Neural Information Processing Systems', pp. 416–422.

Chatterji, N., Neyshabur, B. and Sedghi, H. (2020), The intriguing role of module criticality in the generalization of deep networks, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=S1e4jkSKvB*

Chattopadhay, A., Sarkar, A., Howlader, P. and Balasubramanian, V. N. (2018), Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks, *in* '2018 IEEE Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 839–847.

Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. (2020), A simple framework for contrastive learning of visual representations, *in* 'International Conference on Machine Learning', PMLR, pp. 1597–1607.

Chuang, C.-Y., Mroueh, Y., Greenewald, K., Torralba, A. and Jegelka, S. (2021), Measuring generalization with optimal transport, *in* 'Advances in Neural Information Processing Systems', pp. 8294–8306.

Chun, S., Oh, S. J., Yun, S., Han, D., Choe, J. and Yoo, Y. (2020), An empirical evaluation on robustness and uncertainty of regularization methods, *in* 'arXiv preprint arXiv:2003.03879'.

Cieslak, D. A. and Chawla, N. V. (2009), 'A framework for monitoring classifiers' performance: when and why failure occurs?', *Knowledge and Information Systems* **18**(1), 83–108.

Cortes, C. and Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Darlow, L. N., Crowley, E. J., Antoniou, A. and Storkey, A. J. (2018), CINIC-10 is not imagenet or CIFAR-10, *in* 'arXiv preprint arXiv:1810.03505'.

Dauphin, Y. N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. and Bengio, Y. (2014), Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, *in* 'Advances in Neural Information Processing Systems', pp. 2933–2941.

Davies, D. L. and Bouldin, D. W. (1979), 'A cluster separation measure', *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2), 224–227.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L. (2009), ImageNet: A Large-Scale Hierarchical Image Database, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition'.

Deng, W., Gould, S. and Zheng, L. (2021), What does rotation prediction tell us about classifier accuracy under varying testing environments?, *in* 'International Conference on Machine Learning', PMLR, pp. 2579–2589.

Deng, W. and Zheng, L. (2021), Are labels always necessary for classifier accuracy evaluation?, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 15069–15078.

Denti, F., Doimo, D., Laio, A. and Mira, A. (2021), Distributional results for model-based intrinsic dimension estimators, *in* 'arXiv preprint arXiv:2104.13832'.

DeVries, T. and Taylor, G. W. (2017), Improved regularization of convolutional neural networks with cutout, *in* 'arXiv preprint arXiv:1708.04552'.

Dherin, B., Munn, M. and Barrett, D. G. (2021), The geometric occam's razor implicit in deep learning, *in* 'arXiv preprint arXiv:2111.15090'.

Ding, X., Zhang, X., Han, J. and Ding, G. (2022), Scaling up your kernels to 31x31: Revisiting large kernel design in cnns, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 11963–11975.

Dinh, L., Pascanu, R., Bengio, S. and Bengio, Y. (2017), Sharp minima can generalize for deep nets, *in* 'International Conference on Machine Learning', PMLR, pp. 1019–1028.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. and Houlsby, N. (2021), An image is worth 16x16 words: Transformers for image recognition at scale, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=YicbFdNTTy*

Duan, L. L. and Dunson, D. B. (2021), 'Bayesian distance clustering', *Journal of Machine Learning Research* **22**(224), 1–27.
URL: *http://jmlr.org/papers/v22/20-688.html*

Dwivedi, R., Singh, C., Yu, B. and Wainwright, M. J. (2020), Revisiting complexity and the bias-variance tradeoff, *in* 'arXiv preprint arXiv:2006.10189'.

Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I. and Roy, D. M. (2020), In search of robust measures of generalization, *in* 'Advances in Neural Information Processing Systems', pp. 11723–11733.

Dziugaite, G. K. and Roy, D. M. (2017), Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data, *in* 'arXiv preprint arXiv:1703.11008'.

Elsayed, G., Krishnan, D., Mobahi, H., Regan, K. and Bengio, S. (2018), Large margin deep networks for classification, *in* 'Advances in Neural Information Processing Systems', pp. 850–860.

Engel, A. and den Broeck, V. (2001), *Statistical Mechanics of Learning*, Cambridge University Press. ISBN 9780521774796.

Engstrom, L., Tran, B., Tsipras, D., Schmidt, L. and Madry, A. (2019), Exploring the landscape of spatial robustness, *in* 'International Conference on Machine Learning', PMLR, pp. 1802–1811.

Evgeniou, T., Pontil, M. and Poggio, T. (2000), 'Regularization networks and support vector machines', *Advances in Computational Mathematics* **13**(1), 1–50.

Facco, E., d'Errico, M., Rodriguez, A. and Laio, A. (2017), 'Estimating the intrinsic dimension of datasets by a minimal neighborhood information', *Scientific Reports* **7**(1), 1–8.

Fawzi, A. and Frossard, P. (2016), Measuring the effect of nuisance variables on classifiers, *in* E. R. H. Richard C. Wilson and W. A. P. Smith, eds, 'Proceedings of the British Machine Vision Conference (BMVC)', BMVA Press, pp. 137.1–137.12. ISBN 1-901725-59-6.
URL: *https://dx.doi.org/10.5244/C.30.137*

Feldman, V. and Zhang, C. (2020), What neural networks memorize and why: Discovering the long tail via influence estimation, *in* 'Advances in Neural Information Processing Systems', p. 2881–2891.

Foster, D. J., Greenberg, S., Kale, S., Luo, H., Mohri, M. and Sridharan, K. (2019), Hypothesis set stability and generalization, *in* 'Advances in Neural Information Processing Systems', pp. 6729–6739.

Fyodorov, Y. V. and Williams, I. (2007), 'Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity', *Journal of Statistical Physics* **129**(5), 1081–1116.

Gardner, E. (1988), 'The space of interactions in neural network models', *Journal of Physics A: Mathematical and General* **21**(1), 257.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. and Brendel, W. (2019), Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness., *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=Bygh9j09KX*

Goldfeld, Z., Van Den Berg, E., Greenewald, K., Melnyk, I., Nguyen, N., Kingsbury, B. and Polyanskiy, Y. (2019), Estimating information flow in deep neural networks, *in* 'International Conference on Machine Learning', PMLR, pp. 2299–2308.

Golowich, N., Rakhlin, A. and Shamir, O. (2018), Size-independent sample complexity of neural networks, *in* 'Conference On Learning Theory', PMLR, pp. 297–299.

Gontijo-Lopes, R., Smullin, S., Cubuk, E. D. and Dyer, E. (2021), Tradeoffs in data augmentation: An empirical study, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=ZcKPWuhG6wy*

Goodfellow, I. J., Shlens, J. and Szegedy, C. (2015), Explaining and harnessing adversarial examples, *in* Y. Bengio and Y. LeCun, eds, 'International Conference on Learning Representations'.
URL: *http://arxiv.org/abs/1412.6572*

Granata, D. and Carnevale, V. (2016), 'Accurate estimation of the intrinsic dimension using graph distances: Unraveling the geometric complexity of datasets', *Scientific Reports* **6**(1), 1–12.

Guo, C., Pleiss, G., Sun, Y. and Weinberger, K. Q. (2017), On calibration of modern neural networks, *in* 'International Conference on Machine Learning', PMLR, pp. 1321–1330.

Gutiérrez-Fandiño, A., Pérez-Fernández, D., Armengol-Estapé, J. and Villegas, M. (2021), Persistent homology captures the generalization of neural networks without a validation set, *in* 'arXiv preprint arXiv:2106.00012'.

Han, D., Kim, J. and Kim, J. (2017*b*), Deep pyramidal residual networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5927–5935.

Hardt, M., Recht, B. and Singer, Y. (2016), Train faster, generalize better: Stability of stochastic gradient descent, *in* 'International Conference on Machine Learning', PMLR, pp. 1225–1234.

Harris, E., Marcu, A., Painter, M., Niranjan, M., Prügel-Bennett, A. and Hare, J. (2020), Understanding and enhancing mixed sample data augmentation, *in* 'arXiv preprint arXiv:2002.12047'.

Harvey, N., Liaw, C. and Mehrabian, A. (2017), Nearly-tight VC-dimension bounds for piecewise linear neural networks, *in* 'Conference on Learning Theory', PMLR, pp. 1064–1068.

Hastie, T., Tibshirani, R., Friedman, J. H. and Friedman, J. H. (2009), *The elements of statistical learning: data mining, inference, and prediction*, Vol. 2, Springer. ISBN 9781282126749.

Haussler, D. (1992), 'Decision theoretic generalizations of the PAC model for neural net and other learning applications', *Information and Computation* **100**(1), 78–150.

He, F. and Tao, D. (2020), Recent advances in deep learning theory, *in* 'arXiv preprint arXiv:2012.10931'.

He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. (2020), Momentum contrast for unsupervised visual representation learning, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 9729–9738.

He, K., Zhang, X., Ren, S. and Sun, J. (2016*b*), Identity mappings in deep residual networks, *in* 'European Conference on Computer Vision', Springer, pp. 630–645.

Herbrich, R. and Graepel, T. (2002), 'A PAC-Bayesian margin bound for linear classifiers', *IEEE Transactions on Information Theory* **48**(12), 3140–3150.

Hermann, K. and Lampinen, A. (2020), What shapes feature representations? exploring datasets, architectures, and training, *in* 'Advances in Neural Information Processing Systems', pp. 9995–10006.

Hernández-García, A. and König, P. (2018), Data augmentation instead of explicit regularization, *in* 'arXiv preprint arXiv:1806.03852'.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B. and Hochreiter, S. (2017), Gans trained by a two time-scale update rule converge to a local nash equilibrium, *in* 'Advances in Neural Information Processing Systems', pp. 6626–6637.

Hochreiter, S. and Schmidhuber, J. (1997), 'Flat minima', *Neural Computation* **9**(1), 1–42.

Hooker, S., Erhan, D., Kindermans, P.-J. and Kim, B. (2019), A benchmark for interpretability methods in deep neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 9737–9748.

Hu, X., Chu, L., Pei, J., Liu, W. and Bian, J. (2021), 'Model complexity of deep learning: A survey', *Knowledge and Information Systems* **63**(10), 2585–2619.

Huang, G., Liu, Z., van der Maaten, L. and Weinberger, K. Q. (2017), Densely connected convolutional networks, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 4700–4708.

Huszár, F. (2017), 'mixup: Data-dependent data augmentation'.
URL: *http://www.inference.vc/mixup-data-dependent-data-augmentation/*

Inoue, H. (2018), Data augmentation by pairing samples for images classification, *in* 'arXiv preprint arXiv:1801.02929'.

Ioffe, S. and Szegedy, C. (2015), Batch normalization: Accelerating deep network training by reducing internal covariate shift, *in* 'International Conference on Machine Learning', PMLR, pp. 448–456.

Jacot, A., Gabriel, F. and Hongler, C. (2018), Neural tangent kernel: Convergence and generalization in neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 8571–8580.

Jiang, Y., Nagarajan, V., Baek, C. and Kolter, J. Z. (2022), Assessing generalization of SGD via disagreement, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=WvOGCEAQhxl*

Jiang, Y., Natekar, P., Sharma, M., Aithal, S. K., Kashyap, D., Subramanyam, N., Lassance, C., Roy, D. M., Dziugaite, G. K., Gunasekar, S. et al. (2021), Methods and analysis of the first competition in predicting generalization of deep learning, *in* 'NeurIPS 2020 Competition and Demonstration Track', PMLR, pp. 170–190.

Jiang, Y., Neyshabur, B., Mobahi, H., Krishnan, D. and Bengio, S. (2020), Fantastic generalization measures and where to find them, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=SJgIPJBFvH*

Kakade, S. M., Sridharan, K. and Tewari, A. (2008), On the complexity of linear prediction: Risk bounds, margin bounds, and regularization, *in* 'Advances in Neural Information Processing Systems', p. 793–800.

Kalimeris, D., Kaplun, G., Nakkiran, P., Edelman, B., Yang, T., Barak, B. and Zhang, H. (2019), SGD on neural networks learns functions of increasing complexity, p. 3496–3506.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D. (2020), Scaling laws for neural language models, *in* 'arXiv preprint arXiv:2001.08361'.

Karpathy, A., Johnson, J. and Li, F.-f. (n.d.), 'Tiny ImageNet data set (from the CS 231N course at Stanford)'.
URL: *https://www.kaggle.com/c/tiny-imagenet*

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R. and Fei-Fei, L. (2014), Large-scale video classification with convolutional neural networks, *in* 'Proceedings of the IEEE conference on Computer Vision and Pattern Recognition'.

Kashyap, D., Subramanyam, N. et al. (2021), Robustness to augmentations as a generalization metric, *in* 'arXiv preprint arXiv:2101.06459'.

Kendall, M. G. (1938), 'A new measure of rank correlation', *Biometrika* **30**(1/2), 81–93.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P. T. P. (2016), On large-batch training for deep learning: Generalization gap and sharp minima, *in* 'arXiv preprint arXiv:1609.04836'.

Kirsch, A. and Gal, Y. (2022), 'A note on "assessing generalization of SGD via disagreement"', *Transactions on Machine Learning Research* .
URL: *https://openreview.net/forum?id=oRP8urZ8Fx*

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S. et al. (2020), Captum: A unified and generic model interpretability library for PyTorch, *in* 'arXiv preprint arXiv:2009.07896'.

Koltchinskii, V. and Panchenko, D. (2002), 'Empirical margin distributions and bounding the generalization error of combined classifiers', *The Annals of Statistics* **30**(1), 1–50.

Kornblith, S., Chen, T., Lee, H. and Norouzi, M. (2021), Why do better loss functions lead to less transferable features?, *in* 'Advances in Neural Information Processing Systems', pp. 28648–28662.

Kornblith, S., Norouzi, M., Lee, H. and Hinton, G. (2019), Similarity of neural network representations revisited, *in* 'International Conference on Machine Learning', PMLR, pp. 3519–3529.

Krizhevsky, A., Hinton, G. et al. (2009), 'Learning multiple layers of features from tiny images'.
URL: *http://www.cs.utoronto.ca/ kriz/learning-features-2009-TR.pdf*

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012), ImageNet classification with deep convolutional neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 1097–1105.

Kuzborskij, I. and Lampert, C. (2018), Data-dependent stability of stochastic gradient descent, *in* 'International Conference on Machine Learning', PMLR, pp. 2815–2824.

Kuznetsov, V., Mohri, M. and Syed, U. (2015), Rademacher complexity margin bounds for learning with a large number of classes, *in* 'ICML Workshop on Extreme Classification: Learning with a Very Large Number of Labels', Vol. 2.

Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T. and Lehtinen, J. (2022), The role of ImageNet classes in Fréchet Inception Distance, *in* 'arXiv preprint arXiv:2203.06026'.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J. and Aila, T. (2019), Improved precision and recall metric for assessing generative models, *in* 'Advances in Neural Information Processing Systems', pp. 3927–3936.

Langford, J. and Caruana, R. (2002), (not) bounding the true error, *in* 'Advances in Neural Information Processing Systems', pp. 809–816.

Langford, J. and Shawe-Taylor, J. (2002), PAC-Bayes & margins, *in* 'Advances in Neural Information Processing Systems', p. 439–446.

Lassance, C., Béthune, L., Bontonou, M., Hamidouche, M. and Gripon, V. (2020), Ranking deep learning generalization using label variation in latent geometry graphs, *in* 'arXiv preprint arXiv:2011.12737'.

LeCun, Y. and Cortes, C. (2010), 'MNIST handwritten digit database'.
URL: *http://yann.lecun.com/exdb/mnist/*

Lei, Y. and Ying, Y. (2020), Fine-grained analysis of stability and generalization for stochastic gradient descent, *in* 'International Conference on Machine Learning', PMLR, pp. 5809–5819.

Liang, D., Yang, F., Zhang, T. and Yang, P. (2018), 'Understanding mixup training methods', *IEEE Access* **6**, 58774–58783.

Liang, T., Poggio, T., Rakhlin, A. and Stokes, J. (2019), Fisher-Rao metric, geometry, and complexity of neural networks, *in* 'International Conference on Artificial Intelligence and Statistics', PMLR, pp. 888–896.

Luo, T., Cai, T., Zhang, M., Chen, S., He, D. and Wang, L. (2019), Defective convolutional layers learn robust CNNs, *in* 'arXiv preprint arXiv:1911.08432'.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Houle, M. E., Song, D. and Bailey, J. (2018), Characterizing adversarial subspaces using local intrinsic dimensionality, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=B1gJ1L2aW*

Maddox, W. J., Benton, G. and Wilson, A. G. (2020), Rethinking parameter counting in deep models: Effective dimensionality revisited, *in* 'arXiv preprint arXiv:2003.02139'.

Marcu, A. and Prügel-Bennett, A. (2017), On data-centric myths, *in* 'arXiv preprint arXiv:2111.11514'.

Martin, C. H. and Mahoney, M. W. (2021), Post-mortem on a deep learning contest: a simpson's paradox and the complementary roles of scale metrics versus shape metrics, *in* 'arXiv preprint arXiv:2106.00734'.

Martin, C. H., Peng, T. S. and Mahoney, M. W. (2021), 'Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data', *Nature Communications* **12**(1), 1–13.

Mason, L., Bartlett, P. L. and Baxter, J. (2000), 'Improved generalization through explicit optimization of margins', *Machine Learning* **38**(3), 243–255.

McAllester, D. A. (1999), 'Some PAC-Bayesian theorems', *Machine Learning* **37**(3), 355–363.

Mežnar, S. and Škrlj, B. (2020), Predicting Generalization in Deep Learning via Metric Learning–PGDL shared task, *in* 'arXiv preprint arXiv:2012.09117'.

Miller, G. A. (1995), 'Wordnet: a lexical database for english', *Communications of the ACM* **38**(11), 39–41.

Miyato, T., Maeda, S.-i., Koyama, M. and Ishii, S. (2018), 'Virtual adversarial training: a regularization method for supervised and semi-supervised learning', *IEEE transactions on Pattern Analysis and Machine Intelligence* **41**(8), 1979–1993.

Mohri, M., Rostamizadeh, A. and Talwalkar, A. (2018), *Foundations of machine learning*, MIT Press. ISBN 9780262018258.

Montufar, G. F., Pascanu, R., Cho, K. and Bengio, Y. (2014), On the number of linear regions of deep neural networks, *in* 'Advances in neural information processing systems', pp. 2924–2932.

Morcos, A., Raghu, M. and Bengio, S. (2018), Insights on representational similarity in neural networks with canonical correlation, *in* 'Advances in Neural Information Processing Systems', p. 5732–5741.

Morwani, D., Vashisht, R. and Ramaswamy, H. G. (2020), Using noise resilience for ranking generalization of deep neural networks, *in* 'arXiv preprint arXiv:2012.08854'.

Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. and Dokania, P. (2020), Calibrating deep neural networks using focal loss, *in* 'Advances in Neural Information Processing Systems', pp. 15288–15299.

Müller, R., Kornblith, S. and Hinton, G. E. (2019), When does label smoothing help?, pp. 4694–4703.

Mummadi, C. K., Subramaniam, R., Hutmacher, R., Vitay, J., Fischer, V. and Metzen, J. H. (2021), Does enhanced shape bias improve neural network robustness to common corruptions?, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=yUxUNaj2Sl*

Nagarajan, V. and Kolter, J. Z. (2019*a*), Generalization in deep networks: The role of distance from initialization, *in* 'arXiv preprint arXiv:1901.01672'.

Nagarajan, V. and Kolter, J. Z. (2019*b*), Uniform convergence may be unable to explain generalization in deep learning, *in* 'Advances in Neural Information Processing Systems', pp. 11615–11626.

Nagarajan, V. and Kolter, Z. (2019*c*), Deterministic PAC-bayesian generalization bounds for deep networks via generalizing noise-resilience, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=Hygn2o0qKX*

Nakkiran, P. and Bansal, Y. (2020), Distributional generalization: A new kind of generalization, *in* 'arXiv preprint arXiv:2009.08092'.

Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B. and Sutskever, I. (2021), Deep double descent: Where bigger models and more data hurt, *in* 'Journal of Statistical Mechanics: Theory and Experiment', IOP Publishing, p. 124003.

Natekar, P. and Sharma, M. (2020), Representation based complexity measures for predicting generalization in deep learning, *in* 'arXiv preprint arXiv:2012.02775'.

Neill, J. O. (2020), An overview of neural network compression, *in* 'arXiv preprint arXiv:2006.03669'.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B. and Ng, A. Y. (2011), 'Reading digits in natural images with unsupervised feature learning'.
URL: *http://ufldl.stanford.edu/housenumbers*

Neyshabur, B., Bhojanapalli, S., McAllester, D. and Srebro, N. (2017), Exploring generalization in deep learning, *in* 'Advances in Neural Information Processing Systems', pp. 5947–5956.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y. and Srebro, N. (2019), The role of over-parametrization in generalization of neural networks, *in* 'International Conference on Learning Representations'.

Neyshabur, B., Tomioka, R. and Srebro, N. (2014), In search of the real inductive bias: On the role of implicit regularization in deep learning, *in* 'arXiv preprint arXiv:1412.6614'.

Neyshabur, B., Tomioka, R. and Srebro, N. (2015), Norm-based capacity control in neural networks, *in* 'Conference on Learning Theory', PMLR, pp. 1376–1401.

Ng, A. Y. (2004), Feature selection, L1 vs. L2 regularization, and rotational invariance, *in* 'International Conference on Machine learning', PMLR, pp. 78–86.

Niculescu-Mizil, A. and Caruana, R. (2005), Predicting good probabilities with supervised learning, *in* 'International Conference on Machine Learning', PMLR, pp. 625–632.

Nilsback, M.-E. and Zisserman, A. (2008), Automated flower classification over a large number of classes, *in* 'Sixth Indian Conference on Computer Vision, Graphics & Image Processing', IEEE, pp. 722–729.

Novak, R., Bahri, Y., Abolafia, D. A., Pennington, J. and Sohl-Dickstein, J. (2018), Sensitivity and generalization in neural networks: an empirical study, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=HJC2SzZCW*

Omeiza, D., Speakman, S., Cintas, C. and Weldermariam, K. (2019), Smooth Grad-CAM++: An enhanced inference level visualization technique for deep convolutional neural network models, *in* 'arXiv preprint arXiv:1908.01224'.

Opper, M. (1995), 'Statistical mechanics of learning: Generalization', *The Handbook of Brain Theory and Neural Networks* pp. 922–925.

Opper, M. and Urbanczik, R. (2001), 'Universal learning curves of support vector machines', *Phys. Rev. Lett.* **86**, 4410–4413.
URL: *https://link.aps.org/doi/10.1103/PhysRevLett.86.4410*

Osherov, E. and Lindenbaum, M. (2017), Increasing CNN robustness to occlusions by reducing filter support, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 550–561.

Palma, J. G. (2011), 'Homogeneous middles vs. heterogeneous tails, and the end of the 'inverted-u': It's all about the share of the rich', *development and Change* **42**(1), 87–153.

Parkhi, O. M., Vedaldi, A., Zisserman, A. and Jawahar, C. (2012), Cats and dogs, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', IEEE, pp. 3498–3505.

Pearson, K. (1901), 'Liii. on lines and planes of closest fit to systems of points in space', *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**(11), 559–572.

Philipp, G. and Carbonell, J. G. (2018), The nonlinearity coefficient-predicting generalization in deep neural networks, *in* 'arXiv preprint arXiv:1806.00179'.

Pleiss, G., Zhang, T., Elenberg, E. and Weinberger, K. Q. (2020), Identifying mislabeled data using the area under the margin ranking, *in* 'Advances in Neural Information Processing Systems', pp. 17044–17056.

Poggio, T., Kawaguchi, K., Liao, Q., Miranda, B., Rosasco, L., Boix, X., Hidary, J. and Mhaskar, H. (2017), Theory of deep learning iii: explaining the non-overfitting puzzle, *in* 'arXiv preprint arXiv:1801.00173'.

Power, A., Burda, Y., Edwards, H., Babuschkin, I. and Misra, V. (2022), Grokking: Generalization beyond overfitting on small algorithmic datasets, *in* 'arXiv preprint arXiv:2201.02177'.

Raghu, M., Poole, B., Kleinberg, J., Ganguli, S. and Sohl-Dickstein, J. (2017), On the expressive power of deep neural networks, *in* 'International Conference on Machine Learning', PMLR, pp. 2847–2854.

Rahimi, A. and Recht, B. (2007), Random features for large-scale kernel machines, *in* 'Advances in Neural Information Processing Systems', p. 1177–1184.

Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R. and Khaligh-Razavi, S.-M. (2019), 'Beyond core object recognition: Recurrent processes account for object recognition under occlusion', *PLoS computational Biology* **15**(5), e1007001.

Rueckel, J., Trappmann, L., Schachtner, B., Wesp, P., Hoppe, B. F., Fink, N., Ricke, J., Dinkel, J., Ingrisch, M. and Sabel, B. O. (2020), 'Impact of confounding thoracic tubes and pleural dehiscence extent on artificial intelligence pneumothorax detection in chest radiographs', *Investigative Radiology* **55**(12), 792–798.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C. and Fei-Fei, L. (2015), 'ImageNet Large Scale Visual Recognition Challenge', *International Journal of Computer Vision (IJCV)* **115**(3), 211–252.

Sagawa, S., Raghunathan, A., Koh, P. W. and Liang, P. (2020), An investigation of why overparameterization exacerbates spurious correlations, *in* 'International Conference on Machine Learning', PMLR, pp. 8346–8356.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A. and Chen, X. (2016), Improved techniques for training gans, *in* 'Advances in Neural Information Processing Systems', pp. 2234–2242.

Scheffer, T. and Joachims, T. (1999), Expected error analysis for model selection, *in* 'International Conference on Machine Learning', PMLR, p. 361–370.

Schiff, Y., Quanz, B., Das, P. and Chen, P.-Y. (2021), Predicting deep neural network generalization with perturbation response curves, *in* 'Advances in Neural Information Processing Systems', pp. 21176–21188.

Seeger, M., Langford, J. and Megiddo, N. (2001), An improved predictive accuracy bound for averaging classifiers, *in* 'International Conference on Machine Learning', PMLR, pp. 290–297.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. (2017), Grad-CAM: Visual explanations from deep networks via gradient-based localization, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 618–626.

Shah, H., Tamuly, K., Raghunathan, A., Jain, P. and Netrapalli, P. (2020), The pitfalls of simplicity bias in neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 9573–9585.

Shalev-Shwartz, S. and Ben-David, S. (2014), *Understanding machine learning: From theory to algorithms*, Cambridge University Press. ISBN 9781107057135.

Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C. and Anthony, M. (1998*a*), 'Structural risk minimization over data-dependent hierarchies', *IEEE transactions on Information Theory* **44**(5), 1926–1940.

Shi, B., Zhang, D., Dai, Q., Zhu, Z., Mu, Y. and Wang, J. (2020), Informative dropout for robust representation learning: A shape-bias perspective, *in* 'International Conference on Machine Learning', PMLR, pp. 8828–8839.

Simonyan, K. and Zisserman, A. (2015), Very deep convolutional networks for large-scale image recognition, *in* 'International Conference on Learning Representations'.

Singh, K. K., Yu, H., Sarmasi, A., Pradeep, G. and Lee, Y. J. (2018), Hide-and-seek: A data augmentation technique for weakly-supervised localization and beyond, *in* 'arXiv preprint arXiv:1811.02545'.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S. and Srebro, N. (2018), 'The implicit bias of gradient descent on separable data', *Journal of Machine Learning Research* **19**(1), 2822–2878.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R. (2014), 'Dropout: a simple way to prevent neural networks from overfitting', *Journal of Machine Learning Research* **15**(1), 1929–1958.

Stacke, K., Eilertsen, G., Unger, J. and Lundström, C. (2020), 'Measuring domain shift for deep learning in histopathology', *IEEE Journal of Biomedical and Health Informatics* **25**(2), 325–336.

Summers, C. and Dinneen, M. J. (2019), Improved mixed-example data augmentation, *in* '2019 IEEE Winter Conference on Applications of Computer Vision (WACV)', IEEE, pp. 1262–1270. ISBN 9781475732641.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015), Going deeper with convolutions, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. and Wojna, Z. (2016), Rethinking the inception architecture for computer vision, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2818–2826.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J. and Fergus, R. (2014), Intriguing properties of neural networks, *in* Y. Bengio and Y. LeCun, eds, 'International Conference on Learning Representations'.
URL: *http://arxiv.org/abs/1312.6199*

Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., Hardesty, W., Cox, D. and Kreiman, G. (2018), 'Recurrent computations for visual pattern completion', *Proceedings of the National Academy of Sciences* **115**(35), 8835–8840.

Tishby, N., Pereira, F. C. and Bialek, W. (2000), The information bottleneck method.

Tishby, N. and Zaslavsky, N. (2015), Deep learning and the information bottleneck principle, *in* '2015 IEEE Information Theory Workshop (ITW)', IEEE, pp. 1–5.

Tsitsulin, A., Munkhoeva, M., Mottin, D., Karras, P., Bronstein, A., Oseledets, I. and Mueller, E. (2020), The shape of data: Intrinsic distance for data distributions, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=HyebplHYwB*

Vakanski, A. and Xian, M. (2021), Evaluation of complexity measures for deep learning generalization in medical image analysis, *in* '2021 IEEE 31st International Workshop on Machine Learning for Signal Processing (MLSP)', pp. 1–6.

Valiant, L. (1984*a*), 'A Theory of the Learnable', *Communications of the ACM* **27**(11), 1134–1142.

Valle-Perez, G., Camargo, C. Q. and Louis, A. A. (2019*a*), Deep learning generalizes because the parameter-function map is biased towards simple functions, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=rye4g3AqFm*

Vapnik, V. (1992), Principles of risk minimization for learning theory, *in* 'Advances in Neural Information Processing Systems', pp. 831–838.

Vapnik, V. (1999*a*), *The nature of statistical learning theory*, Springer Science & Business Media.

Vapnik, V. and Chervonenkis, A. Y. (1971), 'On the uniform convergence of relative frequencies of events to their probabilities', *Theory of Probability and its Applications* **16**(2), 264.

Vieira, D., Rangel, F. M., de Faria, F. F. and Paixão, J. (2022), Vector field based neural networks, *in* 'arXiv preprint arXiv:1802.08235'.

Wei, A., Hu, W. and Steinhardt, J. (2022), More than a toy: Random matrix models predict how real-world neural representations generalize, *in* 'International Conference on Machine Learning', PMLR, pp. 23549–23588.

Wilson, A. G. and Izmailov, P. (2020), Bayesian deep learning and a probabilistic perspective of generalization, *in* 'Advances in Neural Information Processing Systems', pp. 4697–4708.

Xiao, H., Rasul, K. and Vollgraf, R. (2017), Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms, *in* 'arXiv preprint arXiv:1708.07747'.

Xu, H. and Mannor, S. (2012), 'Robustness and generalization', *Machine learning* **86**(3), 391–423.

Yang, R., Mao, J. and Chaudhari, P. (2022), Does the data induce capacity control in deep learning?, *in* 'International Conference on Machine Learning', PMLR, pp. 25166–25197.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M. and Wu, Y. (2022), 'Coca: Contrastive captioners are image-text foundation models', *Transactions on Machine Learning Research* .
URL: *https://openreview.net/forum?id=Ee277P3AYC*

Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J. and Yoo, Y. (2019), CutMix: Regularization strategy to train strong classifiers with localizable features, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 6023–6032.

Zahavy, T., Kang, B., Sivak, A., Feng, J., Xu, H. and Mannor, S. (2016), Ensemble robustness and generalization of stochastic deep learning algorithms, *in* 'arXiv preprint arXiv:1602.02389'.

Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O. (2017), Understanding deep learning requires rethinking generalization, *in* 'International Conference on Learning Representations'.

Zhang, C., Bengio, S. and Singer, Y. (2022), 'Are all layers created equal?', *Journal of Machine Learning Research* **23**(67), 1–28.
URL: *http://jmlr.org/papers/v23/20-069.html*

Zhang, H., Cisse, M., N. Dauphin, Y. and Lopez-Paz, D. (2018a), mixup: Beyond empirical risk minimization, *in* 'International Conference on Learning Representations'.
URL: *https://openreview.net/forum?id=r1Ddp1-Rb*

Zhang, T. (2002), 'Covering number bounds of certain regularized linear function classes', *Journal of Machine Learning Research* **2**(Mar), 527–550.

Zhang, T. and Zhu, Z. (2019), Interpreting adversarially trained convolutional neural networks, *in* 'International Conference on Machine Learning', PMLR, pp. 7502–7511.

Zhong, Z., Zheng, L., Kang, G., Li, S. and Yang, Y. (2020*b*), Random erasing data augmentation, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', number 34(07), pp. 13001–13008.

Zhou, Y., Liang, Y. and Zhang, H. (2018), Generalization error bounds with probabilistic guarantee for SGD in nonconvex optimization, *in* 'arXiv preprint arXiv:1802.06903'.

Zhu, H., Tang, P., Park, J., Park, S. and Yuille, A. (2019), Robustness of object recognition under extreme occlusion in humans and computational models, *in* 'arXiv preprint arXiv:1905.04598'.

Supplementary Material

A Supplementary Material for Directions in Generalisation: a Short Introduction

In this section we give the function definitions, inequalities, and theorems that we use in the calculations presented in Directions in Generalisation: a Short Introduction, as well as other quantities mentioned throughout the thesis, such as the Rademacher complexity and the covering number.

For every positive integer n, the **Gamma function** is given by

$$\Gamma(a) = (a-1)! \, .$$

The integral form of the Gamma function for complex numbers with positive real part can be written as

$$\Gamma(a) = \int_0^\infty \tau^{a-1} \, e^{-\tau} d\tau \, ,$$

while the **probability density function** of the Gamma distribution is given by

$$f(\tau \,|\, a, b) = \frac{b^a \, e^{-b\,\tau} \, \tau^{a-1}}{\Gamma(a)} \, .$$

The **Beta function** for complex numbers with positive real parts is given by

$$\text{Beta}(a,\, b) = \int_0^1 r^{a-1} \, (1-r)^{b-1} dr \, .$$

The Beta function can also be written using the Gamma function as such

$$\text{Beta}(a,\, b) = \frac{\Gamma(a)\,\Gamma(b)}{\Gamma(a,\, b)}\ .$$

The mean of the beta distribution is given by $\frac{a}{a+b}$ while the variance is $\frac{a\,b}{(a+b)^2\,(a+b+1)}$.

The probability mass function of the **Binomial distribution** is given by

$$\text{Binom}(\ell\,|\,m, R) = \binom{m}{\ell} R^\ell\,(1-R)^{m-\ell}\ ,$$

with mean $m\,R$ and variance $m\,R\,(1-R)$.

The Normal distribution with a mean of 0 and standard deviation of 1 has a probability density function of

$$\Phi(x) = \frac{1}{\sqrt{2\,\pi}}\,e^{-\frac{x^2}{2}}\ .$$

For a p-dimensional sphere, the **surface area** is given by

$$S_p(r) = \frac{2\,\pi^{\frac{p+1}{2}}}{\Gamma(\frac{p+1}{2})}\,r^p\ ,$$

where r is the radius of the hypersphere.

For defining the Rademacher complexity and the covering number, we follow the notation of Shalev-Shwartz and Ben-David (2014). Given a set of vectors $A \in \mathbb{R}^m$ and i.i.d. random variables $\sigma_1, \ldots, \sigma_m$ such that $\mathbb{P}\Big(\sigma_i = 1\Big) = \mathbb{P}\Big(\sigma_i = -1\Big) = 1/2$, the **Rademacher complexity** is given by

$$R(A) = \frac{1}{m}\,\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\mathbf{a}\in A}\sum_{i=1}^{m} a_i\,\sigma_i\right]\ .$$

Conversely, the **covering number** of the set A is given by the cardinality of the smallest set A' that covers the set A with balls of radius r. We say that the set A is covered by a set A' if $\forall \mathbf{a} \in A$, $\exists \mathbf{a}' \in A'$ such that $\|\mathbf{a} - \mathbf{a}'\| \leq r$.

Jensen's inequality states that given a concave function f, and a random variable X,

$$\mathbb{E}\Big[f(X)\Big] \leq f\Big(\mathbb{E}\Big[X\Big]\Big)\ .$$

B Supplementary Material for The Theoretical Approach: the Importance of the Data

B.1 Asymptotic Generalisation Performance

In this section we show that generalisation performance is driven by the power-law growth in the distribution of risks. Consider the case of a realisable problem with an infinite hypothesis space such that a randomly chosen hypothesis has a risk, R_h, distributed according to

$$\rho(r) = r^{a-1} \sum_{i=0}^{\infty} c_i \, r^i \; .$$

In this scenario a hypothesis $h \in \mathcal{H}_{\text{ERM}}$ will be distributed according to

$$f_{R_{\text{ERM}}}(r) = \frac{(1-r)^m \, \rho(r)}{\displaystyle\int_0^1 (1-r')^m \, \rho(r') \, \mathrm{d}r'} \; .$$

The expected generalisation performance is thus given by

$$\mathbb{E}\left[R_{\text{ERM}}\right] = \frac{\displaystyle\sum_{i=0}^{\infty} c_i B(a+1+i, m+1)}{\displaystyle\sum_{i=0}^{\infty} c_i B(a+i, m+1)} = \frac{c_0 \, B(a+1, m+1) + c_1 \, B(a+2, m+1) + \cdots}{c_0 \, B(a+1, m+1) + c_1 \, B(a+2, m+1) + \cdots}$$

$$= \frac{\frac{B(a+1,m+1)}{B(a,m+1)} + \frac{c_1 \, B(a+2,m+1)}{c_0 \, B(a,m+1)} + \cdots}{1 + \frac{c_1 \, B(a+1,m+1)}{c_0 \, B(a,m+1)} + \cdots} = \frac{a}{m} + O\left(\frac{1}{m^2}\right) \; .$$

where we have used

$$\frac{B(a+i, m+1)}{B(a, m+1)} = \frac{\Gamma(a+i)\,\Gamma(a+m+1)}{\Gamma(a)\,\Gamma(a+i+m+1)}$$

$$= \frac{a\,(a+1)\,\cdots\,(a+i-1)}{(a+m+1)\,(a+m+2)\,\cdots\,(a+m+i)} \; .$$

Thus, the generalisation error in the limit of large m depends only on the exponents describing the polynomial growth in the distribution of risk. As mentioned in Section 2.2, this observation will make our results more general.

C The Distribution of Risks: Case Study

In this section we compute the distribution of risks for two problems. The first one is a hypothetical scenario where we have a hypothesis space that includes all binary functions. We use this example to reason about the label randomisation experiment from the perspective of model attunement. Subsequently, we compute $\rho(r)$ for an unrealisable perceptron.

C.1 All Binary Functions

Let E denote the number of errors made by a randomly chosen hypothesis. If the hypothesis space, $\mathcal{H}$, consists of all Boolean functions, $f : \mathcal{X} \to \{T, F\}$, where $\mathcal{X}$ is the set of all possible inputs, then the probability distribution of the risks for a randomly chosen hypothesis is given by

$$\rho(r) = \mathbb{P}\left(E = r\,N\right) = \text{Binom}\left(E \,\middle|\, 2^{|\mathcal{X}|}, \tfrac{1}{2}\right) = \frac{1}{2^{2^{|\mathcal{X}|}}}\binom{2^{|\mathcal{X}|}}{E}\,. \tag{1}$$

In most machine learning applications $|\mathcal{X}|$ is exponential in the number of features. For example, for binary strings of length n, $|\mathcal{X}| = 2^n$. This distribution is very sharply concentrated around the mean $\mathbb{E}\left[R\right] = 1/2$, having a variance of $|\mathcal{X}|/4$. We can approximate this distribution with a Beta distribution where $a = b = |\mathcal{X}|/2$, which has the same mean and almost identical variance as the binomial distribution.[1] The expected ERM error for the Beta distribution approximation is $|\mathcal{X}|/(2\,|\mathcal{X}| + m)$. We therefore require m to be of order $|\mathcal{X}|$ before we obtain any generalisation performance. In this case, the lack of generalisation is a result of the *huge value of the attunement* parameter rather than the *size of the hypothesis space*.

Zhang et al. (2017) trained models on randomly labeled CIFAR-10 training data to near-zero training loss. The CIFAR-10 training data consists of 50 000 samples belonging to 10 different classes. This suggests a hypothesis space consisting of at least $10^{50\,000}$ hypotheses. However, this is much smaller than $2^{|\mathcal{X}|}$, which for colour images with 32×32 pixels taking 256 possible values, is $2^{256^{3072}}$. Provided $|\mathcal{H}|$ is substantially smaller than $2^{256^{3072}}$ we can still achieve a relatively high degree of attunement (i.e. small value of a). The simple problem of learning all binary functions illustrates a case of poor attunement, which leads to no generalisation.

[1] Recall that for a Beta distribution, $\text{Beta}(r|a,b)$, the mean is $a/(a+b)$ and the variance is equal to $a\,b/\left((a+b)^2(a+b+1)\right)$. So for $a = b$ the mean is $\frac{1}{2}$ and the variance is $1/(8a+4)$.

C.2 Unrealisable Perceptron

We now consider using a perceptron with a different distribution of data. Consider a two-class problem with data $(\boldsymbol{x}, y)$ where $y \in \{-1, 1\}$ and $\boldsymbol{x}$ is

$$f_X(\boldsymbol{x}|y) = \mathcal{N}\big(\boldsymbol{x}\big|\Delta\, y\, \boldsymbol{w}^*, \mathbf{I}\big) \ ,$$

where $\boldsymbol{w}^*$ is some arbitrary unit norm vector. The parameter Δ determines the separation between the means of the two classes. The Bayes optimal classifier corresponds to a hyperplane orthogonal to $\boldsymbol{w}^*$. We consider learning a perceptron defined by the unit variance weight vector $\boldsymbol{w}$. Defining $\boldsymbol{\eta} = y\, \boldsymbol{w}^\mathsf{T}(\boldsymbol{x} - y\, \Delta, \boldsymbol{s}^*) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\cos(\theta) = \boldsymbol{w}^\mathsf{T}\boldsymbol{w}^*$ we can write vector $\boldsymbol{w} = cos(\theta)\, \boldsymbol{w}^* + sin(\theta)\, \boldsymbol{x}^\mathsf{T}$, such that

$$\begin{aligned}
\boldsymbol{x}^\mathsf{T}\, \boldsymbol{w} &= (y\, \Delta\, \boldsymbol{w}^* + \boldsymbol{\eta})^\mathsf{T}\, (cos(\theta)\, \boldsymbol{w}^* + sin(\theta)\, \boldsymbol{w}^\mathsf{T}) \\
&= y\, \Delta\, cos(\theta) + \boldsymbol{\eta}^\mathsf{T}\, (cos(\theta)\, \boldsymbol{w}^* + sin(\theta)\, \boldsymbol{w}^\mathsf{T}) \\
&= y\, \Delta\, cos(\theta) + \boldsymbol{\eta}^\mathsf{T}\, \boldsymbol{w} \ .
\end{aligned}$$

We denote $\xi = \boldsymbol{\eta}^\mathsf{T}\, \boldsymbol{w}$. As a sum of independent normal components, ξ will also be normally distributed. And since the expected value of $\boldsymbol{\eta}$ is 0, ξ will also have 0 mean, while $\mathbb{E}\left[\xi^2\right] = \boldsymbol{w}^\mathsf{T}\, \mathbb{E}\left[\boldsymbol{\eta}\, \boldsymbol{\eta}^\mathsf{T}\right]\, \boldsymbol{w} = 1$. Thus, $\xi \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $y\, \boldsymbol{x}^\mathsf{T}\, \boldsymbol{w} = \Delta\, cos(\theta) + \xi$. The expected risk is

$$R_{\boldsymbol{w}} = \mathbb{P}\left(y\, \boldsymbol{x}^\mathsf{T}\, \boldsymbol{w} < 0\right) = \mathbb{P}\left(\Delta\, \cos(\theta) < -\xi\right) = \Phi(-\Delta\, \cos(\theta)) \ ,$$

where $\Phi(z)$ is the cumulative probability distribution for a zero mean, unit variance normally distributed random variable. The distribution of weight vectors at an angle θ to $\boldsymbol{w}^*$ is the same as that for the realisable perceptron (Equation (2.10)). The distribution of risks is given by $\rho(r) = f_\Theta(\theta(r))/\frac{\mathrm{d}r}{\mathrm{d}\theta}$, where $r = \Phi(-\Delta\, \cos(\theta))$ or $\theta(r) = \arccos(\Phi^{-1}(r)/\Delta)$. Noting that

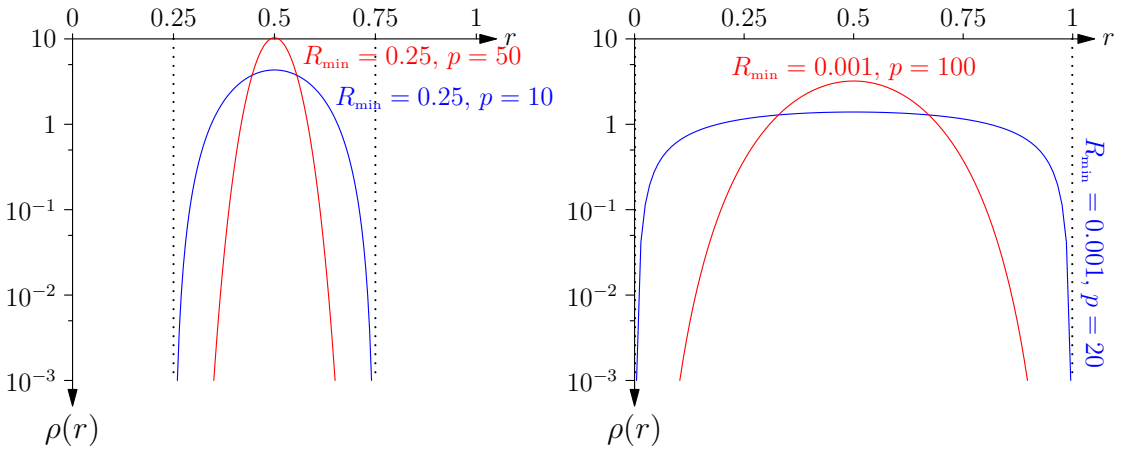
$$\frac{\mathrm{d}r}{\mathrm{d}\theta} = \Delta\, \sin(\theta)\, \frac{\mathrm{e}^{\,-\Delta^2\, \cos^2(\theta)/2}}{\sqrt{2\,\pi}}$$

and writing

$$\sin^{p-3}(\theta) = \left(1 - \cos^2(\theta)\right)^{\frac{p-3}{2}} = \left(1 - \left(\frac{\Phi^{-1}(r)}{\Delta}\right)^2\right)^{\frac{p-3}{2}} \ ,$$

we get

$$\rho(r) = \frac{\sqrt{2\,\pi}}{\Delta\, B(\frac{1}{2}, \frac{p-1}{2})} \left(1 - \left(\frac{\Phi^{-1}(r)}{\Delta}\right)^2\right)^{\frac{p-3}{2}}\, \mathrm{e}^{\,(\Phi^{-1}(r))^2/2} \ .$$



(A) $R_{\min} = 0.25$ (B) $R_{\min} = 0.001$

FIGURE C.1: Probability density, $f_R(r)$, plotted on a logarithmic scale against the risk, r, for $R_{\min} = 0.25$ (left) so that $\Delta \approx -0.674$ with $p = 10$ and $p = 50$, and $R_{\min} = 0.001$ (right) so that $\Delta \approx -3.090$ with $p = 20$ and $p = 100$. The vertical dotted lines show the maximum and minimum risks in $\mathcal{H}$.

To help understand this equation, in Figure C.1 we depict the probability density, $\rho(r)$, plotted against the risk, r, on a logarithmic scale for two different levels of class separability which correspond to $R_{\min} = 0.25$ (Figure C.1a) and $R_{\min} = 0.001$ (Figure C.1b). For each level of separability, we look at varying the number of features. Just as in the case of the realisable perceptron, a reduction in the number of features changes the distribution of risks and therefore directly influences the attunement.

We note that for unrealisable models the distribution of risks, $\rho(r)$, will be 0 for $r < R_{\min}$. When $\mathbb{E}\left[R_{\text{ERM}}\right]$ is substantially greater than $R_{\min}$, then the generalisation behaviour will be similar to a realisable model with the same attunement. As m increases, $\mathbb{E}\left[R_{\text{ERM}}\right]$ will converge to $R_{\min}$. The two quantities that characterise the asymptotic behaviour in the unrealisable case are $R_{\min}$, and the power-law growth of $\rho(r)$ as we increase r from $R_{\min}$.

A More Detailed Analysis. Revisiting Assumptions

In our analysis, we assumed the independence of the losses. We also replaced the expectation of a ratio by the ratio of expectations. This is clearly only an accurate approximation if the values are heavily concentrated around their mean. In this section, we treat these assumptions and approximations more carefully.

Corrections due to Fluctuations

For any independently chosen finite data set, $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu) \mid \mu = 1, 2, \ldots, m\}$, there will be chance fluctuations between the features vectors, $\boldsymbol{x}^\mu$, that lead to variations in the generalisation performance, which in turn lead to a change in the mean behaviour. In

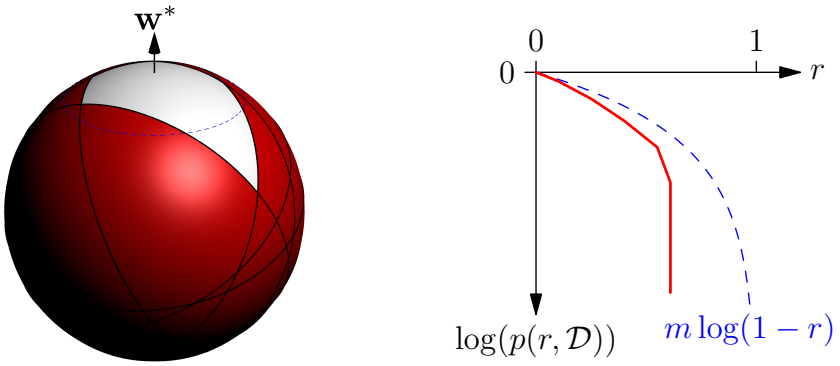


FIGURE C.2: Schematic illustration of the proportion, $p(r, \mathcal{D})$ of parameter space with risk r that correctly classifies the training examples for a realisable perceptron. (a) Shows the weight space for a perceptron with three inputs. The white area represents the parameters that correctly classify the examples. The vector $\boldsymbol{w}^*$ represents the correct concept. The proportion $p(r, \mathcal{D})$ would corresponds to the fraction of weight space at a given value of r that correctly classifies the inputs. For the perceptron, constant risk would correspond to line of constant latitude as illustrated by the dashed blue line. (b) Shows an illustrate of $\log(p(r, \mathcal{D}))$ together with $\log(\mathbb{E}\left[p(r, \mathcal{D})\right])$ (dashed curve).

this section we derive an approximation to these corrections which depend on the detail of the learning machine beyond the distribution of the risks, so cannot be computed in general. We derive these corrections for the realisable perceptron. For any realisable model we have shown that

$$\mathbb{E}\left[R_{\text{ERM}}|\mathcal{D}\right] = \frac{M_1}{M_0} = \frac{\mathbb{E}_{r\sim\rho(r)}\Big[r\, p(r, \mathcal{D})\Big]}{\mathbb{E}_{r\sim\rho(r)}\Big[p(r, \mathcal{D})\Big]} \,,$$

where $p(r, \mathcal{D})$ is the proportion of hypotheses with risk r that correctly classify all training examples (i.e. $\forall\,(\boldsymbol{x}, y) \in \mathcal{D},\ h(\boldsymbol{x}) = y$, where $h(\boldsymbol{x})$ denotes the prediction of hypothesis h given a feature vector y). In Figure C.2 we illustrate schematically what $p(r, \mathcal{D})$ might look like for the realisable perceptron.

Denoting the set of hypotheses with risk r that correctly classify the first k training examples by

$$\mathcal{H}_r^k = \left\{h \in \mathcal{H}\middle| R_h = r \wedge \forall\,\mu = 1,\, 2,\, \ldots,\, k,\ h(\boldsymbol{x}^\mu) = y^\mu\right\} \,,$$

then $p(r, \mathcal{D}) = |\mathcal{H}_r^m|/|\mathcal{H}_r^0|$, where $\mathcal{H}_r^0$ is the set of hypotheses with risk r. We note that we can also write $p(r, \mathcal{D})$ as

$$p(r, \mathcal{D}) = \prod_{k=1}^{m} \frac{|\mathcal{H}_r^k|}{|\mathcal{H}_r^{k-1}|} = \frac{|\mathcal{H}_r^m|}{|\mathcal{H}_r^{m-1}|}\frac{|\mathcal{H}_r^{m-1}|}{|\mathcal{H}_r^{m-2}|}\cdots\frac{|\mathcal{H}_r^1|}{|\mathcal{H}_r^0|} = \frac{|\mathcal{H}_r^m|}{|\mathcal{H}_r^0|}.$$

Defining $p_r^k = |\mathcal{H}_r^k|/|\mathcal{H}_r^{k-1}|$ then

$$p(r, \mathcal{D}) = \prod_{k=1}^{m} p_r^k \,.$$

The quantity p_r^k is the proportion of hypotheses in $\mathcal{H}_r^{k-1}$ that correctly classify the k^{th} data point. By the definition of risk, $\mathbb{E}\left[p_r^k\right] = 1 - r$. However, due to chance correlations between training examples, p_r^k, will fluctuate. As the training examples are drawn independently, p_r^k and p_r^j will be independent random variables when $k \neq j$. Now,

$$\ln\Big(p(r,\mathcal{D})\Big) = \sum_{k=1}^{m} \ln\Big(p_r^k\Big)$$

is a sum of independent random variables and by the central limit theorem this sum will converge towards a normal distribution [2]. As a consequence, $p(r,\mathcal{D})$ will be close to a log-normal distribution and its median value will typically be smaller than its mean. The typical value of $\mathbb{E}\left[R_{\text{ERM}}|\mathcal{D}\right]$ is going be given when $p(r,\mathcal{D})$ takes its most likely value or equivalently by the median of $\ln\Big(p(r,\mathcal{D})\Big)$. Since $\ln\Big(p(r,\mathcal{D})\Big)$ is normally distributed, its median, mode and mean are all the same. Thus to compute the typical value of $\mathbb{E}\left[R_{\text{ERM}}|\mathcal{D}\right]$ we can use the most likely value of $p(r,\mathcal{D})$ which will be

$$\hat{p}(r,\mathcal{D}) = \exp\Big(\mathbb{E}\left[\ln\Big(p(r,\mathcal{D})\Big)\right]\Big)$$

where

$$\mathbb{E}\left[\ln\Big(p(r,\mathcal{D})\Big)\right] = \sum_{k=1}^{m} \mathbb{E}\left[\ln\Big(p_r^k\Big)\right] = \sum_{k=1}^{m} \mathbb{E}\left[\frac{|\mathcal{H}_r^k|}{|\mathcal{H}_r^{k-1}|}\right] .$$

By Jensen's inequality $\mathbb{E}\left[\ln\Big(p(r,\mathcal{D})\Big)\right] \leq \ln\Big(\mathbb{E}\left[p(r,\mathcal{D})\right]\Big)$. This does not tell us whether the fluctuations improve or worsen the generalisation performance (which depends on the gradient of $\ln\Big(p(r,\mathcal{D})\Big)$). However, for $r = 0$ we know that $p(r,\mathcal{D}) = 1$ so that the fluctuations can only increase the gradient of $\mathbb{E}\left[\ln\Big(p(r,\mathcal{D})\Big)\right]$ around $r = 0$. As this gradient determines the asymptotic generalisation performance (what we have termed the attunement) we see that the 'annealed approximation' will be an upper bound on the asymptotic generalisation performance (i.e. it will be overly conservative).

To get an understanding of the quantitative corrections we need to model the fluctuations that we are likely to get in p_r^k. As p_r^k is a random variable that lies in the range from 0 to 1 it is reasonable to approximate its distribution by a beta distribution, $p_r^k \sim \text{Beta}(A_r^k, B_r^k)$. This distribution is unrelated to that used in the β-Risk model — we use a beta distribution as in both cases we are modelling a random variable that lies in the range 0 to 1. If p_r^k is beta distributed then

$$\mathbb{E}\left[\ln\Big(p(r,\mathcal{D})\Big)\right] = \sum_{k=1}^{m} \Big(\psi(A_r^k) - \psi(A_r^k + B_r^k)\Big) .$$

[2] m is usually sufficiently large that the distribution of $\ln\big(p(r,\mathcal{D})\big)$ will be very closely approximated by a normal distribution

Using the fact that $\mathbb{E}\left[p_r^k\right] = 1 - r$ and denoting the variance of p_r^k by v_r^k we get

$$1 - r = \frac{A_r^k}{A_r^k + B_r^k}, \qquad v_r^k = \frac{A_r^k\, B_r^k}{(A_r^k + b_r^k)^2\,(A_r^k + B_r^k + 1)}\ ,$$

from which we find

$$A_r^k = \frac{1-r}{\Delta_r^k}, \qquad B_r^k = \frac{r}{\Delta_r^k}, \qquad \Delta_r^k = \frac{v_r^k}{r(1-r) - v_r^k}\ .$$

By definition

$$v_r^k = \mathbb{E}\left[(p_r^k)^2\right] - (1-r)^2\ ,$$

where

$$\begin{aligned}\mathbb{E}\left[(p_r^k)^2\right] &= \mathbb{E}\left[\frac{|\mathcal{H}_r^k|^2}{|\mathcal{H}_r^{k-1}|^2}\right] \\ &= \mathbb{E}\left[\frac{1}{|\mathcal{H}_r^{k-1}|^2}\sum_{h\in\mathcal{H}_r^{k-1}}\sum_{h'\in\mathcal{H}_r^{k-1}}\mathbb{E}_{(\boldsymbol{x}^k,y^k)}\Big[\left[\!\left[h(\boldsymbol{x}^k)=y^k\right]\!\right]\ \left[\!\left[h'(\boldsymbol{x}^k)=y^k\right]\!\right]\Big]\right]\ .\end{aligned}$$

We observe that the fluctuations depend on the expected correlation between hypotheses of a given risk. Denoting the joint probability of a pair of hypotheses making a particular prediction for a randomly sampled data-point, $(\boldsymbol{x}, y)$, by

$$\mathbb{P}\left(\left[\!\left[h(\boldsymbol{x})=y\right]\!\right]=w,\ \left[\!\left[h'(\boldsymbol{x})=y\right]\!\right]=z\right) = p_{wz}(h,h')$$

(with $w, z \in \{0,1\}$) then

$$\mathbb{E}\left[(p_r^k)^2\right] = \frac{1}{|\mathcal{H}_r^{k-1}|^2}\sum_{h\in\mathcal{H}_r^{k-1}}\sum_{h'\in\mathcal{H}_r^{k-1}} p_{11}(h,h')\ .$$

We note that $p_{11}(h,h') + p_{01}(h,h') = \mathbb{P}\Big(h(\boldsymbol{x})=y\Big) = 1-r$. Also for randomly selected hypotheses, by symmetry, $p_{10}(h,h') = p_{01}(h,h')$, while for any pair of hypotheses $p_{10}(h,h') + p_{01}(h,h') = \mathbb{P}\Big(h(\boldsymbol{x})\neq h'(\boldsymbol{x})\Big)$. From this we find $p_{11}(h,h') = 1 - r - \mathbb{P}\Big(h(\boldsymbol{x})\neq h'(\boldsymbol{x})\Big)/2$ and the variance in $p(r,\mathcal{D})$ is given by

$$v_r^k = r(1-r) - \frac{1}{2\,|\mathcal{H}_r^{k-1}|^2}\sum_{h\in\mathcal{H}_r^{k-1}}\sum_{h'\in\mathcal{H}_r^{k-1}}\mathbb{P}\Big(h(\boldsymbol{x})\neq h'(\boldsymbol{x})\Big)\ .$$

Up to now the only information we have required about the problem was the distribution of risks. However, to compute $\mathbb{P}\Big(h(\boldsymbol{x})\neq h'(\boldsymbol{x})\Big)$ we need to know more about the

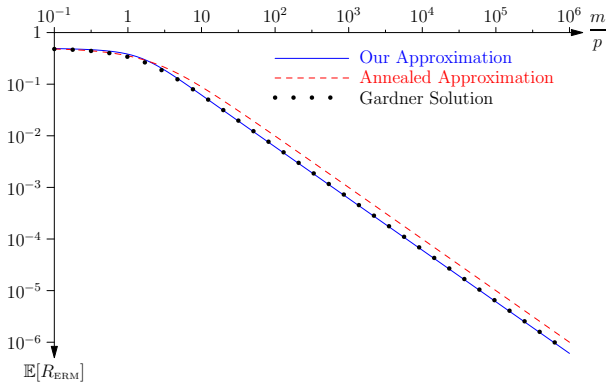


FIGURE C.3: Plot of the expected ERM error versus the ratio m/p for the realisable perceptron in the limit $p \to \infty$ plotted on a log-log scale. The solid blue line shows the approximation developed in this paper. The black dotted line shows the Gardner solution while the red dashed line shows the annealed solution.

learning algorithm. We consider the realisable perceptron where

$$\mathbb{P}\left(h(\boldsymbol{x}) \neq h'(\boldsymbol{x})\right) = \frac{\theta_{hh'}}{\pi} \ ,$$

where $\theta_{hh'} = \arccos(\boldsymbol{w}_h^{\mathsf{T}} \boldsymbol{w}_{h'})$ is the angle between the weight vectors corresponding to hypotheses h and h'. For any hypothesis, h, with risk r, the weight vector can be written as

$$\boldsymbol{w}_h = \boldsymbol{w}^* \cos(\pi\, r) + \boldsymbol{w}_h^{\perp} \sin(\pi\, r) \ ,$$

where $\boldsymbol{w}^*$ is a unit vector in the direction of the perfect perceptron and $\boldsymbol{w}_h^{\perp}$ is some unit orthogonal to $\boldsymbol{w}^*$. For two hypotheses with risk r

$$\theta_{hh'} = \arccos(\cos^2(\pi\, r) + \boldsymbol{w}_h^{\perp \mathsf{T}} \boldsymbol{w}_{h'}^{\perp} \sin^2(\pi\, r)) \ .$$

If there are a large number of features $\boldsymbol{w}_h^{\perp \mathsf{T}} \boldsymbol{w}_{h'}^{\perp} \approx 0$ for the vast majority of hypothesis pairs so that $\theta_{hh'} \approx \arccos(\cos^2(\pi\, r))$. Ignoring other correlations

$$v_r^k = r(1-r) - \frac{1}{2\,\pi} \arccos\big(\cos^2(\pi\, r)\big)$$

and

$$\Delta_r^k = \frac{2\,\pi\, r\,(1-r)}{\arccos(\cos^2(\pi\, r))} - 1 \ .$$

In Figure C.3 we show the expected ERM risk versus m/p (recall p is the number of features in the perceptron) in the limit when $p \to \infty$. For comparison, the annealed approximation is also shown in Figure C.3. Finally, we also show the Gardner solution, which is only defined in this limit (Gardner, 1988; Engel and den Broeck, 2001). As we can see, our approximation is very close to the Gardner solution in the limit when m/p becomes large. There are discrepancies for smaller values of m/p due to ignoring other

fluctuations. There will be fluctuations because pairs of training examples $\boldsymbol{x}^\mu$ and $\boldsymbol{x}^\nu$ will typically have small chance correlations. These are of order $1/p$, but because there are $m-1$ other training examples, the fluctuations will grow.

Although the Gardner solution strictly requires us to take the limit $p \to \infty$ it has been shown that it provides a reasonable approximation to Gibb's learning for perceptrons with a smaller number of features (see Figure 1.4 in Engel and den Broeck (2001)). Gibb's learning for the perceptron can be well approximated by the perceptron learning algorithm with some added noise to ensure that different parts of H_{ERM} are explored (Engel and den Broeck, 2001, Section 3.2). The Gardner approach has been used to examine other learning rules, noisy training sets, etc. See Engel and den Broeck (2001) for a review of the literature. The approach has also been extended to SVMs see, for example, Opper and Urbanczik (2001). These calculations are very involved and model specific. In this section, we have proposed understanding generalisation behaviour more generally by considering $\rho(r)$. However, to obtain results applicable to any learning machine we use the annealed approximation.

D Supplementary Material for Steps Towards the Empirical Approach: Understanding by Distorting

D.1 Experimental details

Throughout Chapters 3 and 4, we use PreAct-ResNet18 (He et al., 2016*b*) models, trained for 200 epochs with a batch size of 128. For the MSDA parameters we use the same values as Harris et al. (2020). All models are augmented with random crops and horizontal flips and results are averaged across 5 runs. We train using SGD with a momentum of 0.9, learning rate of 0.1 up until epoch 100 and 0.001 for the rest of the training. This is due to an incompatibility with newer versions of the PyTorch library of the official implementation of Harris et al. (2020), which we use as a starting point for model training. However, the difference in learning rate schedule between our work and prior art does not affect our findings since we are not introducing a new method to be applied at training time. In our case, it is sufficient to show that the bias exists in at least one configuration.

For the analysis we also used adapted code from Carlucci et al. (2019) for patch-shuffling. The models were trained on either one of the following GPUs: Titan X Pascal, GeForce GTX 1080ti or Tesla V100. For the analyses, a GeForce GTX 1050 was also used. The average training time was less than two hours, with the exception of models trained on Tiny-ImageNet, which took around 10 hours to run.

TABLE D.1: DI index for alternative grid sizes. Gaps can be identified for various grid sizes, with more pronounced differences for finer-grained grids.

		basic	MixUp	FMix	CutMix
CIFAR-10	2×2	$0.61_{\pm 0.24}$	$0.56_{\pm 0.33}$	$0.19_{\pm 0.14}$	$0.12_{\pm 0.06}$
	8×8	$6.41_{\pm 0.55}$	$6.95_{\pm 1.96}$	$2.75_{\pm 1.46}$	$1.41_{\pm 1.15}$
CIFAR-100	2×2	$1.03_{\pm 0.29}$	$0.46_{\pm 0.14}$	$0.21_{\pm 0.14}$	$0.12_{\pm 0.07}$
	8×8	$9.16_{\pm 6.15}$	$3.10_{\pm 4.59}$	$1.62_{\pm 0.89}$	$0.65_{\pm 0.50}$
Tiny ImageNet	8×8	$5.76_{\pm 6.61}$	$5.73_{\pm 3.82}$	$2.49_{\pm 1.38}$	$0.60_{\pm 0.69}$
	16×16	$44.01_{\pm 36.47}$	$14.06_{\pm 14.63}$	$11.94_{\pm 17.79}$	$1.86_{\pm 1.98}$
ImageNet	16×16	0.72	1.38	0.55	–
	64×64	4.89	41.16	12.77	–

Training models

The code for model training is largely based on the open-source official implementation of FMix, which also includes those of MixUp, CutOut, and CutMix. For the experiment where we use the reformulated objective to combine data sets, instead of mixing with a permutation of the batch, as it is done in the original implementation of the mixed-augmentations, we draw a batch from the desired data set. To ensure a fair comparison, for the basic we also perform inter-batch mixing.

Evaluating robustness

For the CutOcclusion measurement, we modify open-source code to restrict the occluding patch to lie withing the the margins of the image to be occluded. This is to ensure that the mixing factor λ matches the true proportion of the occlusion. For iOcclusion, the implementation of Grad-CAM is again adapted from publicly available code. With both methods, we evaluate 5 instances of the same model and average over the results obtained. The added computation time of iOcclusion over the regular CutOcclusion for a fixed occlusion fraction is that of performing Grad-CAM on train and test data, as well as evaluating on the latter. With a batch size of 128, this takes under half an hour.

D.2 Varying the grid size

Table D.1 gives the results obtained when varying the number of image tiles to be randomly rearranged. We observe that data interference appears for different grid sizes. Note that the considered grid sizes are chosen according to the size of the images. For example, for samples of 32×32 pixels we consider 2×2 and 8×8 grids in addition to the 4×4 used in the main body of the thesis. Conversely, for 224×224 images, we use larger grid sizes (16×16 and 64×64).

TABLE D.2: DI index for occluding with images from another data set. Again, a gap can be identified.

	basic	MixUp	FMix	CutMix
CIFAR-10	$1.71_{\pm 0.15}$	$1.05_{\pm 0.22}$	$0.14_{\pm 0.03}$	$0.16_{\pm 0.05}$
CIFAR-100	$0.48_{\pm 0.09}$	$0.61_{\pm 0.27}$	$0.90_{\pm 0.15}$	$1.25_{\pm 0.25}$
Fashion MNIST	$3.40_{\pm 0.29}$	$3.06_{\pm 1.07}$	$1.81_{\pm 0.55}$	$2.61_{\pm 0.80}$
Tiny ImageNet	$0.25_{\pm 0.12}$	$0.17_{\pm 0.04}$	$0.06_{\pm 0.03}$	$0.12_{\pm 0.04}$

TABLE D.3: DI index for patching using masks sampled from Fourier space. Just as in the case of rectangular masks, a gap can be identified.

	basic	MixUp	FMix	CutMix
CIFAR-10	$2.08_{\pm 1.13}$	$1.79_{\pm 1.09}$	$1.32_{\pm 0.99}$	$4.21_{\pm 1.23}$
CIFAR-100	$4.06_{\pm 01.47}$	$3.11_{\pm 02.29}$	$9.90_{\pm 14.32}$	$2.89_{\pm 05.36}$
Fashion MNIST	$49.55_{\pm 20.45}$	$40.69_{\pm 21.63}$	$27.87_{\pm 17.57}$	$61.04_{\pm 17.92}$
Tiny ImageNet	$4.37_{\pm 0.85}$	$6.95_{\pm 1.84}$	$3.60_{\pm 1.73}$	$5.92_{\pm 4.38}$
ImageNet	3.27	2.24	6.08	–

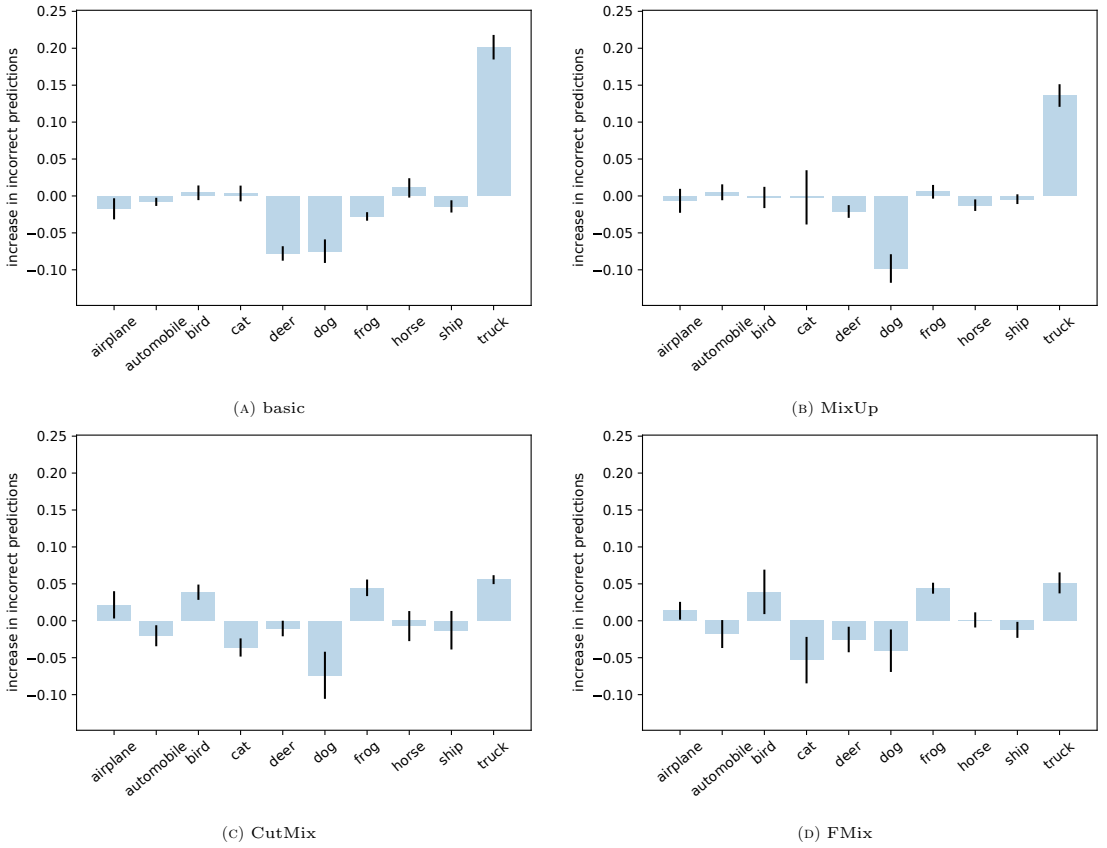
D.3 Patch-shuffling

We look at the classes which have the highest increase in incorrect predictions and note that their shapes are characterised by strong horizontal and vertical edges. For example, on CIFAR-100, varying the grid size between 2×2, 4×4 and 8×8 gives "Lamp", "Bus" and "Table" as dominant c_{max} classes, while the model trained on Fashion MNIST with the standard procedure tends to predict grid-shuffled images as "Bag". On ImageNet, the basic model tends to wrongly identify the patch-shuffled images as belonging to the 550th class "Espresso Maker".

D.4 CutOcclusion

In this section we experiment with alternative masking methods when computing CutOcclusion. We note that the bias exists when occluding with patches taken from images belonging to different data sets (Table D.2). Figure D.1 gives a visual account of the results obtained for CIFAR-10 when mix-patching. Note that for Fashion MNIST we use MNIST, for Tiny ImageNet we use ImageNet, while for CIFAR-10 we mix with CIFAR-100 and vice versa. Since ImageNet images are significantly larger than those of the other data sets, mixing would imply padding large areas, which would give results very similar to uniform patching. We also experiment with VGG models, where on CIFAR-10 the basic has a DI index of $0.80_{\pm 0.40}$ compared to $0.18_{\pm 0.11}$ of MixUp.

We then use masks sampled from Fourier space (Table D.3) and note that even for these irregularly shaped distortions, we can identify a gap in most cases. The only exception is in the case of Fashion MNIST. It must be stressed that although all the models we experimented with presented Data Interference for this problem, this does not exclude



(A) basic (B) MixUp

(C) CutMix (D) FMix

FIGURE D.1: Difference between wrongly predicted classes when testing on original data versus CutMix-distorted images. The evaluated models from left to right, top to bottom are trained on CIFAR-10 with: no mixed-data augmentation (basic), MixUp, CutMix, and FMix.

the possibility of constructing a different model that is insensitive to this distortion. For example, we identify a gap for this problem when mix-masking (DI index of $4.09_{\pm 1.74}$ for the basic model as opposed to $1.87_{\pm 0.27}$ for a model trained on images that were masked out using FMix-like masks). Thus, when occluding with a particular shape we implicitly disfavour models in which learnt representations are related to the features introduced by that shape.

D.5 Alternative CutOcclusion

Table D.4 gives the DI index when forcing the occluding patch to lie within image boundaries for patch sizes sampled uniformly from [0.1, 1]. Note that in the case of Tiny ImageNet the bias is more visibly present for larger occluders. As such, uniformly sampling the patch size from the interval [0.3, 1] results in a DI index of $13.46_{\pm 5.74}$ for the basic model, while the level of data interference from MixUp is only $4.75_{\pm 1.93}$. Similarly, for Fashion MNIST, when we increase the size of the occluder we obtain $0.65_{\pm 0.20}$ for Mixup as opposed to $0.07_{\pm 0.11}$ for the basic model. However, this does not change the conclusions of our experiments since, as mentioned in the main paper, robustness studies are usually carried out with large occluder sizes.

TABLE D.4: DI index for sampling occluder size from a uniform distribution when the patch is restricted to lying within image boundaries and the size is sampled from [0.1, 1] uniformly.

	basic	MixUp	FMix	CutMix
CIFAR-10	$5.74_{\pm 1.86}$	$0.75_{\pm 0.69}$	$1.25_{\pm 1.24}$	$3.61_{\pm 3.60}$
CIFAR-100	$28.63_{\pm 9.85}$	$6.31_{\pm 7.03}$	$5.86_{\pm 5.86}$	$12.63_{\pm 24.69}$
Fashion MNIST	$1.88_{\pm 3.36}$	$3.54_{\pm 1.33}$	$1.91_{\pm 3.33}$	$1.76_{\pm 2.91}$
ImageNet	0.25	0.48	0.14	–

D.6 Data Interference across Architectures

To verify if a shape bias evaluation based on patch-shuffling would give unfair results when comparing across architectures, we compute the *DI* index for a number of models trained with the basic approach (without mixed data augmentation) on CIFAR-10. The DI index of models such as WideResNet and ResNet is high ($0.95_{\pm 0.20}$ and $1.27_{\pm 0.39}$ respectively), while for PyramidNet, BagNet17 and BagNet9 it is small (0.26, $0.53_{\pm 13}$, $54_{\pm 0.16}$). DenseNet ($0.66_{\pm 0.48}$) and VGG ($0.60_{\pm 0.15}$) have comparable DI indices.

Thus, we find that intensity with which distortions interfere with learnt representations is different for different architectures. Comparing robustness to occlusion using CutOcclusion would give biased results when comparing architectures trained under the same conditions.

E Sensitivity to the Patch Shape

In Section 3.3.1 we created a new augmentation method, RM, which samples 3 random masks from Fourier space and uses only those for the whole training. This was used to obtain a model that learns *some* robustness to occlusion but does not see sufficient variety in the masks so as to learn invariance to the strong edges. As a result, it will have a higher *DI* index compared to the FMix model while having a robustness level close to that of mask-trained models.

In Figure E.1 we provide the results for a range of occlusion fractions. When evaluated using CutOcclusion, the robustness of CutMix models has a high standard deviation. Nonetheless, we can see that the RM model appears to be closer to MixUp than to FMix or CutMix. On the other hand, when robustness is evaluated using iOcclusion, the curve obtained for RM is closer to the mask-trained models. Therefore, our method provides a fairer evaluation for models which, like RM, are affected by data interference.

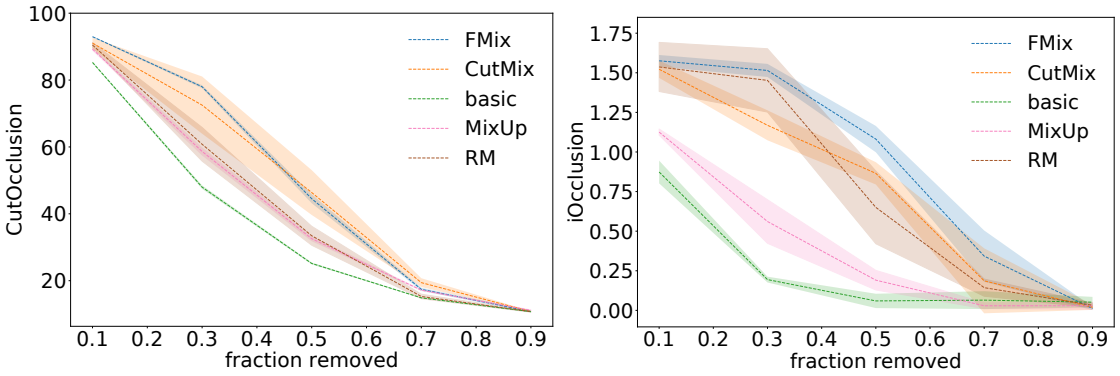


FIGURE E.1: Robustness to occlusion as measured by CutOcclusion (left) and iOcclusion (right). The occluding patches are non-uniform in this case. The robustness curve for RM very closely follows that of the MixUp model when evaluated with CutOcclusion. However, iOcclusion captures the robustness that RM gains with mask-training, situating it closer to the curves of FMix and CutMix.

F Description of Data Sets Used

Table F.1 provides the essential information describing the data sets used in the thesis.

TABLE F.1: Basic information regarding the main data sets referenced in the thesis. Note that for SVHN, Oxford Pets, Oxford Flowers, Bengali we have rounded up the number of available training and test samples.

Name	Short description	#Classes	Image size	#Training samples	#Test samples
CIFAR-10	images of animals and vehicles	10	32×32	50×10^3	10×10^3
CIFAR-100	images of animals, man-made items, food, etc.	100	32×32	50×10^3	10×10^3
MNIST	grayscale images of handwritten digits	10	28×28	60×10^3	10×10^3
Fashion-MNIST	grayscale images of garments and shoes	10	28×28	50×10^3	10×10^3
SVHN	images of digits in real-world scenes	10	32×32	610×10^3	27000
Tiny Imagenet	images of animals, man-made items, food, etc.	200	64×64	110×10^3	10×10^3
Bengali.AI	Bengali Handwritten Graphemes	168	64×64	200×10^3	200×10^3
Oxford Pets	images of breeds of cats and dogs	37	32×32	9×10^3	1800
Oxford Flowers	images of flower species	102	32×32	7×10^3	10^3
CINIC-10	images of animals and vehicles	10	32×32	180×10^3	90×10^3