University of Southampton

Faculty of Engineering and the Physical Sciences
School of Electronics and Computer Science

Human Processes in Artificial Vision

by

Ethan William Albert Harris

ORCiD: 0000-0002-1825-0097

A thesis for the degree of
Doctor of Philosophy

September 2022

University of Southampton

Abstract

Faculty of Engineering and the Physical Sciences
School of Electronics and Computer Science

Doctor of Philosophy

Human Processes in Artificial Vision

by Ethan William Albert Harris

From the earliest experiments with artificial neurons to the development of convolutional neural networks and beyond, nature has proven to be an inexhaustible source of inspiration. Not only can the mechanisms of vision in nature provide a template for the construction of new artificial models, but the analyses and tools developed over centuries of biological and psychological study can help to illuminate the nature of learned functions. In this work, we start by studying the extent to which convolutional networks make use of shape and colour cues when classifying images by analysing collections of synthetic images that maximise the models prediction for a target class (super-stimuli). Next, we explore the impact of convolutional neural network architecture on learned colour processing via the concept of opponency from the vision science literature. We show that the distribution of opponent cells and an analysis of cell tuning can provide valuable insight regarding the nature of colour processing. Following on from these findings, we consider the impact of opponency on adversarial robustness, finding that an increase in the percentage of opponent cells is not sufficient to consistently improve robustness to a suite of attacks. In the penultimate chapter, we introduce a convolutional model of foveation. We show that the addition of this foveated convolution improves the localisation performance of a visual attention model. We study the distribution of opponent cells and stimulus preference in these foveated layers as a function of eccentricity and find strong connections with findings from the biology literature regarding the nature of peripheral colour vision. In our final chapter, we construct a model of visual attention with the capacity to sketch its input, inspired by psychological concepts surrounding the production of visual memories. We demonstrate several interesting properties of this model, particularly regarding the drawing policies it learns and their connection to a notion of object.

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:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. 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;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. 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;
5. I have acknowledged all main sources of help;
6. 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;
7. Parts of this work have been published as: Harris et al. (2019a), Harris et al. (2020b), Harris et al. (2020a), and Harris et al. (2019b).

Signed:..... Date:.....

Acknowledgements

This has been the hardest part to write. I think it's because it feels silly to express thanks for something that I've found so challenging, painful even, at times. It also feels self-aggrandising; as if by thanking others I'm acknowledging that there is an achievement somewhere to be thankful for. So why write anything at all? Because at this moment I'm acutely aware of how selfish the whole thing is. I'm aware that for all my struggles I have stood to gain where so many have struggled along with me and sought nothing in return.

To my extraordinary wife, Jess, who was there through it all and managed to give me time, space, and encouragement to write at the same time as caring for our two baby daughters. To my supervisor, Jon, who spent many energetic days and long nights at the coalface and from who I have learnt a great deal in my time both as a PhD student and previously as an undergrad. To my parents, who gave me the confidence to push myself and resist the easier, less fulfilling, alternatives. Lastly, to my daughters Morgan and Lexie. In case you ever read this, know that I was smart once, regardless of how old and stupid I must seem to you now. To all of you, I'm afraid I have very little to offer in return, but whatever contributions lie in my ramblings are now irrevocably yours.

Chapter 1

Introduction

The history of artificial neural networks and what we now refer to as deep learning is steeped in biological inspiration. The early neural network models of Rosenblatt (1962) drew direct inspiration from Hebb's postulate (Hebb, 1949) and other contemporaneous theories of neuronal adaptation. Later, the work of Oja (1982) showed that a normalised variant of Hebb's rule can be used to find principle components. Further developments in the field of self-organisation, such as Kohonen networks (Kohonen, 1982), trace their roots back to the neuronal models of Rosenblatt (1962) and the work of Turing (1952) on morphogenesis (the biological process of tissue growth and pattern formation).

Not only are the roots of this scientific tree planted firmly in the discovery and elucidation of the biological neuron, but many of its branches and leaves draw on specific insights from the fields of psychology and neuroscience. An example of this can be found in the development of Convolutional Neural Networks (CNNs) (Le Cun et al., 1990). In primate vision, cells fire according to the presence or absence of specific visual features in their receptive field. The relative positions of features in the visual field are preserved in what is known as a retinotopic map and the types of features that are extracted get more abstract with depth. A single convolutional neuron is a feature extractor that is tiled over the input to produce a feature map. In a convolutional layer, many convolutional neurons are executed in parallel to produce a set of feature maps. In a CNN, many convolutional layers are executed in series to produce layers of 'retinotopic' feature extraction that become more abstract with depth, mirroring their biological counterpart.

Since the foundation of our contemporary view has so abundantly benefited from the rich world of neuroscience, it seems reasonable to suggest that future developments may benefit also. In a world so busied by the ceaseless march toward the state of the art we may find inspiration in nature's most complex creation. The goal of this thesis

is to explore this belief, not just as a source of ideas but of those much rarer, tricker things: good ideas.

Of the many important challenges currently facing deep learning community, two stand out in their relevance to this work: interpretation, and generalisation. The key axis which separates traditional machine learning from modern 'deep' methods is one of complexity. Not complexity of the ideas necessarily, but complexity of the resultant learned function. It is from this simple reality that the challenges of interpretation and generalisation arise.

1.1 Interpretation

The interpretation challenge derives from the fact that the learned functions of models with millions (and even billions) of parameters cannot be easily understood. This of course depends on the level of understanding required. One view by which we might gauge our level of understanding is that of Feynman, "What I cannot create, I do not understand.". The seminal work of Olah et al. (2017) gave an enticing glimpse at would could be achieved with gradient ascent techniques for feature visualisation. Briefly, these approaches try to synthesise stimuli which elicit an extreme response from a particular functional unit (such as a cell or a layer) of interest. These early approaches suffer a few limitations. First, they do not allow for any automatic analysis. Second, they describe only individual cells or layers, rather than interactions between them. Finally, they characterise only the functional extremes (the most or least excitatory stimulus) rather than providing a more nuanced understanding. The later works of Olah et al. (2018, 2020a) and Cammarata et al. (2020) sought to address the latter two limitations, ultimately leading to the recent work of Cammarata et al. (2021) which directly targets Feynman's view by trying to reverse engineer part of a trained CNN. They test the effectiveness of their approach by ablating the reverse engineered part and replacing it with an equivalent subnetwork that is constructed manually to have the same functional characteristics. The results of this experiment show that performance of the ablated network is improved by the addition of the hard-coded subnetwork. These recent developments draw on long-standing ideas from neuroscience relating to the characterisation and classification of function.

Of the three limitations, one remains: these analyses are still performed manually and cannot be automated over a population of trained models. This is arguably the most important limitation since it prevents us from realising many of the grander visions around model interpretation. For example, one area of interest is to articulate precisely how trained models differ from each other in order to understand (and ideally capitalise on) why one model outperforms another in a particular setting. The first key

objective of this thesis is to find automated methods for model interpretability inspired by approaches from the neuroscience literature.

1.2 Generalisation

The generalisation challenge is one of performance. This includes finding more performant models, but also broadening the scope of such models (that is, the range of possible applications for which these approaches can be expected to perform). Several recent works have looked to nature for inspiration with the goal of improving generalisation. In Geirhos et al. (2019), the authors find that ImageNet trained CNNs are biased towards texture which, the authors argue, is in contrast to the strong shape bias of human vision. The authors subsequently construct a 'Stylized-ImageNet' data set and show that models trained on this variant have improved shape bias. These shape-biased models exhibit improved robustness to a wide range of adversarial distortions. In Dapello et al. (2020), the authors use a hand-crafted set of features designed to simulate the primary visual cortex (V1) as the first layer of a CNN. They find that the addition of this 'VOneBlock' dramatically improves adversarial robustness. These approaches work either by manipulating the environment or by removing the learned component altogether. It remains to be seen whether approaches to promoting similarity between CNNs and human vision that retain the data distribution and don't hard-code features can still provide these benefits. The work of Deza and Konkle (2020a) explores the impact of foveation on learned representations and finds evidence for improved robustness in foveated models.

In general, the majority of recent efforts in this space have focused on the contribution made by the visual cortex, and V1 in particular, on the generalisation and performance of human vision. In many cases, the impact of earlier visual processes such as foveation and retinal feature extraction has been assumed to be negligible. The second objective of this thesis is to consider whether these components of the visual pathway may have been too readily dismissed and may yet lend a hand to the generalisation objective.

1.3 Outline and List of Contributions

This section gives an outline and lists the contributions of each chapter.

1.3.1 Chapter 2: From Photons to Visual Phenomenology

This chapter gives a brief overview of the key concepts from vision science that are relevant to the work presented here.

1.3.2 Chapter 3: Colour and Classification

This chapter considers the extent to which colour is used by CNNs when classifying images. The contributions are:

- A review and reimplementation of key techniques for the production of super-stimuli.
- The introduction of the concept of a trait group; a set of classes that all are characterised by a particular trait.
- Specification of two trait groups for the ImageNet data set. The first, 'fruits', are well characterised by a particular colour. The second, 'animals' are characterised jointly by particular colours and patterns.
- Demonstration that super-stimuli for these trait groups effectively capture the characteristic colours, suggesting that CNNs do rely on colour for their classifications.
- Further demonstration that the super-stimuli depict characteristic shape cues for their target classes, presenting new evidence in support of a representation of shape in deep convolutional networks.

1.3.3 Chapter 4: Architecture and Opponency

The contributions made in this chapter have been published as Harris et al. (2019a) in the NeurIPS 2019 workshop on Shared Visual Representations in Human and Machine Intelligence and as Harris et al. (2020b) in Neural Computation. The main contributions are:

- A review of the physiology and psychophysics of early colour vision.
- A review of opponency in artificial vision systems.
- A method for the analysis of spatial, colour, and double opponency in deep CNNs.
- A method for determining the distribution of stimuli that elicit the most excitatory and inhibitory responses in convolutional neurons.
- A method for inferring the hue sensitivity curve of a cell or layer in a network using automatic differentiation.
- A colour variant of the RetinaNet model from Lindsey et al. (2019), trained over a comprehensive range of hyper-parameters and assessed for it's accordance with the original model.

- An in depth analysis of a single convolutional cell and discussion of how these analyses relate to our proposed automated methods.
- Presentation and discussion of the results of our proposed methods on the trained networks.
- A comprehensive set of control experiments across a range of different conditions.
- A discussion of how these results provide insight to the effect of convolutional network architecture on the nature of learned opponency and colour tuning.

1.3.4 Chapter 5: Opponency and Robustness

This chapter details contributions made in our work presented as Harris et al. (2020a) in the NeurIPS 2020 workshop on Shared Visual Representations in Human and Machine Intelligence. The contributions are:

- An anatomically constrained ResNet variant.
- Reproduction of the experiments from Harris et al. (2020b) on the constrained ResNet trained on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) data set.
- Extension of the methodology from Harris et al. (2020b) to further classify cells as luminance opponent.
- Analysis and visualisation of receptive fields of cells in both constrained and unconstrained networks.
- An application of our previously proposed technique for determining whether networks capture colour information through the construction of super-stimuli for a trait group, showing that even with a bottleneck size of one our model still uses colour information.
- A comprehensive analysis of the impact of a bottleneck on the adversarial robustness of our ResNet variant.

1.3.5 Chapter 6: Foveation and Function

Parts of this chapter were presented as Harris et al. (2019b) at the NeurIPS 2019 workshop on Shared Visual Representations in Human and Machine Intelligence. The contributions are:

- Development of a convolutional model of foveation (the foveated convolution).

- Demonstration that the addition of foveation to Spatial Transformer Networks (STNs) (Jaderberg et al., 2015) improves localisation and classification performance.
- Demonstration that the learned attention policies of a simple network with a foveated convolution can be used as an effective preprocessing step enabling much larger networks to generalise to highly translation variant settings.
- Analysis of receptive fields in foveated and non-foveated networks showing evidence for a multi-modal representation that supports both classification and localisation objectives.
- Demonstration that the natural phenomenon of a reduction in opponency in peripheral vision also emerges in classification networks equipped with a foveated layer.
- A method for determining whether cells prefer Red-Green or Blue-Yellow opponent stimuli.
- Results showing that Red-Green preference is highest in the centre of the visual field and reduces in the periphery for foveated networks, again matching observations from biology.

1.3.6 Chapter 7: Attention and Memory

This chapter presents unpublished efforts to construct a deep model based on a simplified psychological model of vision, incorporating concepts from the psychology literature such as a working memory and a visual sketchpad. The contributions are:

- A deep model of pose-adjusted visual attention where the attention network is based on a Spatial Transformer Network (STN).
- A method for constructing an inverted glimpse that enables the network to *sketch* on a canvas in the glimpse region, producing an auto-encoding model.
- A demonstration that the glimpse *resolution* controls the sketching mode and can induce a parts-based representation.
- A series of experiments showing that this parts-based representation is an effective self-supervised feature.

1.3.7 Chapter 8: Conclusions and Future Work

This chapter summarises and provides suggestions for the further development of the ideas and contributions presented in this thesis.

Chapter 2

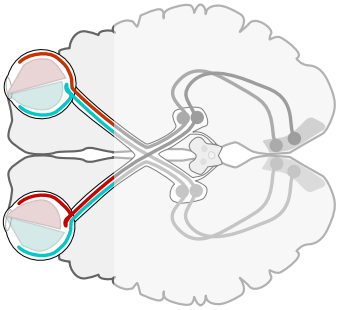
From Photons to Visual Phenomenology

In this chapter, we review at a high level the findings from vision science that are of relevance to this work. Our review charts a course through the visual system, tracking light entering the eye as it gives rise to perception. At a high level, we divide the visual system into three functional units: the eye, the optic chiasm and the thalamus, and the primary visual cortex and the two streams.

Throughout this chapter we take a purely functional view of the visual system. That is, we are concerned with enumerating precisely those visual components that could be modelled with machine learning methods. Where relevant, we will review such modelling attempts from the literature. We also ignore the role of dynamics; taking the view that integrated spike trains (the number of activations in a given time frame) are sufficient to explain cognition, or visual cognition at least.

2.1 The Eye

We begin our journey with the retina. First, light hits the back of the eye causing photoreceptors (Rod or Cone cells) to activate. These activations then propagate through several layers of cells, finally reaching retinal ganglion cells whose axons project down the optic nerve. The distribution of photoreceptors and ganglion cells is foveated with high density, and thus accuity, at the fovea and low density in the periphery. We now discuss each of these functional aspects of the eye in more detail.



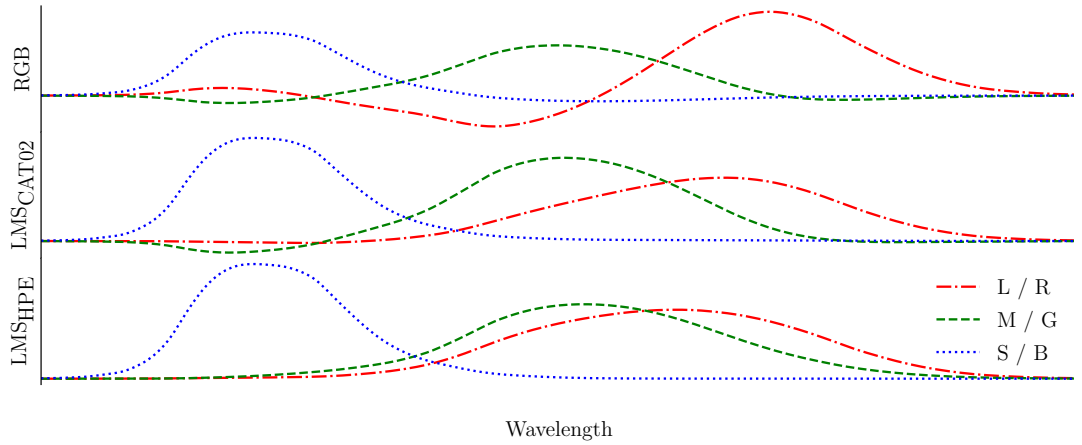
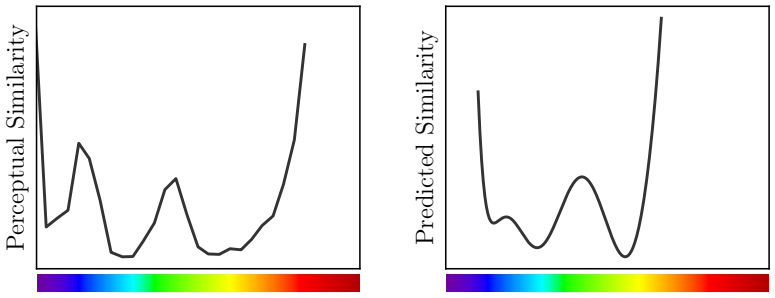


FIGURE 2.1: Channel response curves against wavelength for the three colour spaces used in our experiments: RGB, LMS_{CAT02}, and LMS_{HPE}. CIE XYZ tristimulus values are computed by transformation from wavelength (between 380 and 700) with the CIE 1931 2° standard observer colour matching function using the colour science Python library (Mansencal et al.). Tristimulus values for each colour space are subsequently computed by application of the relevant transformation matrix.

2.1.1 Photoreceptors

The two types of photoreceptor (cells which activate in response to light stimulus) present in the eye are Rods and Cones. The key difference between Rods and Cones is that Rod cells respond only to the amount of light present whereas Cone cells are specialised for different wavelengths and thus encode colour. It is a common misconception to think of Rod activations as providing brightness information, influencing normal daylight vision. In fact, Rods are saturated in even moderately lit environments and are responsible for our ability to see (without colour) in low-light conditions. This is a useful addition to our visual tool-belt, but not of any great relevance when modelling typical daylight vision.

Cone cells, on the other hand, can be thought of as the first transformation of visual information in the visual system. In the trichromatic vision of Humans, there are three Cone types, L, M, and S which are tuned to light of long, medium, and short wavelengths respectively. These tunings are typically approximated as Red, Green, and Blue, although this is a point of occasional controversy. Figure 2.1 shows the tuning curves with respect to wavelength of RGB and two biologically inspired cone colour systems: LMS_{CAT02}, and LMS_{HPE}. These cone colour spaces are obtained by transformation from CIEXYZ space with the CAT02 (Moroney et al., 2002) and Hunt-Pointer-Estévez (Hunt and Pointer, 1985; Estévez, 1981) transformation matrices respectively. From the figure, we can see that the LMS colour spaces exhibit a more pronounced S response and that the L and M channels tend to be directly proportional. This contrasts with the inverse relationship between the R and G channels in RGB.



(A) Human Observer (B) Natural Scenes

FIGURE 2.2: (a) Wavelength change needed to elicit a just-noticeable difference in hue for a human observer from Bedford and Wyszecki (1958). (b) Predicted similarity derived from images of natural scenes from Long et al. (2006).

The ability to distinguish between light of different wavelengths in Humans is a function of wavelength. Figure 2.2a shows this function, obtained by Bedford and Wyszecki (1958) via the just-noticeable difference method. This is the change in wavelength between two light sources required to elicit a just-noticeable difference in perceived Hue. In Long et al. (2006) the authors further suggest that the reason for this non-uniform spectral sensitivity derives from the statistics of natural scenes, showing that the curve predicted from a data set of natural images (reproduced in Figure 2.2b) bares a strong resemblance to that obtained for a human observer. Although no component of early vision is necessarily wholly responsible for the characteristic curve of wavelength discrimination, it is important to note that one way to explain the discrimination curve is in terms of the cone responses (Zhaoping et al., 2011). Indeed, Figure 2.1 serves to demonstrate that the choice of colour space acts as the first feature in a visual pipeline, accentuating or masking changes in colour (and impacting wavelength discrimination). This is a more direct explanation of the phenomenon since scene statistics can be seen as indirectly controlling wavelength discrimination through evolutionary modifications of cone properties. In Chapter 4 we take inspiration from these experiments to propose a novel analysis of the colour sensitivity of deep networks.

2.1.2 Properties of Ganglion Cell Receptive Fields

Following Adrian and Matthews (1928), Hartline (1938, 1940) discovered evidence for different types of cellular behaviour to stimuli, and in particular found that inhibitory interactions were sometimes revealed when multiple receptors were excited (Hartline et al., 1952). Kuffler (1953) and Barlow (1953) investigated this finding further, and discovered cells with spatial receptive fields that are opponent to each other. These early results, obtained by presenting spots of light to different parts of the receptive field, showed an antagonism (opponency) between an inner centre and outer surround.

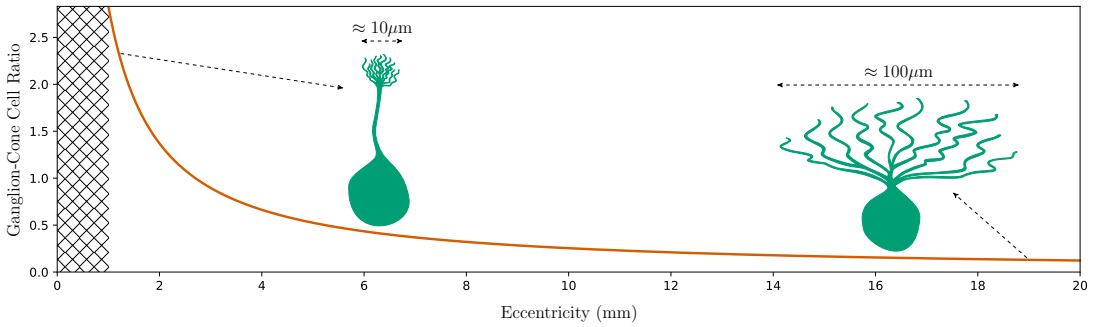


FIGURE 2.3: Ratio of ganglion to cone cell density as a function of eccentricity from the fovea. Approximate ganglion and cone cell densities taken from Curcio and Allen (1990) and Curcio et al. (1990) respectively. Dendritic field size is based on that of midget ganglion cells described by Dacey and Petersen (1992). The hatched region represents the foveal center which we do not address in this work.

Nowadays, it is widely accepted that such 'centre-surround' cells can be found in the retina and LGN (Hubel and Wiesel, 2004).

There is a connection between anatomy and the relative presence of linear and non-linear cells in the retina. For example, midget cells, which are well approximated by a linear model (Smith et al., 1992), are the most prevalent ganglion cell type in the human retina (Dacey, 1993). In contrast, the most prevalent ganglion cell type in the mouse retina is a non-linear feature detector that is thought to act as an overhead predator detection mechanism (Zhang et al., 2012), not dissimilar to the 'fly detectors' and 'bug perceivers' observed by Barlow (1953) and Lettvin et al. (1959) respectively. In their experiments with CNNs, Lindsey et al. (2019) suggest that the contrast between the anatomy of the primate and mouse visual systems can be considered in terms of network depth. The authors subsequently present evidence that the natural differences in function derive from these associated differences in visual system anatomy. In particular, deeper networks learn linear features in early layers, whereas shallower networks learn more complex, non-linear, functions similar to those found in mice retina.

2.1.3 Foveation

We can describe visual acuity in the retina in terms of the relative densities of cone and ganglion cells at varying degrees of eccentricity from the fovea. We will also need to consider the dendritic spread of ganglion cells in order to establish some understanding of their field of view. An approximation of these statistics with a visual demonstration of the dendritic spread is given in Figure 2.3. From the figure we can see that oversampling occurs near the fovea with a ratio of two or more ganglion cells for each cone cell. This falls off rapidly until, at the periphery, there are nearly ten cone cells for each ganglion cell. At the same time, dendritic spread increases by a factor of around ten.

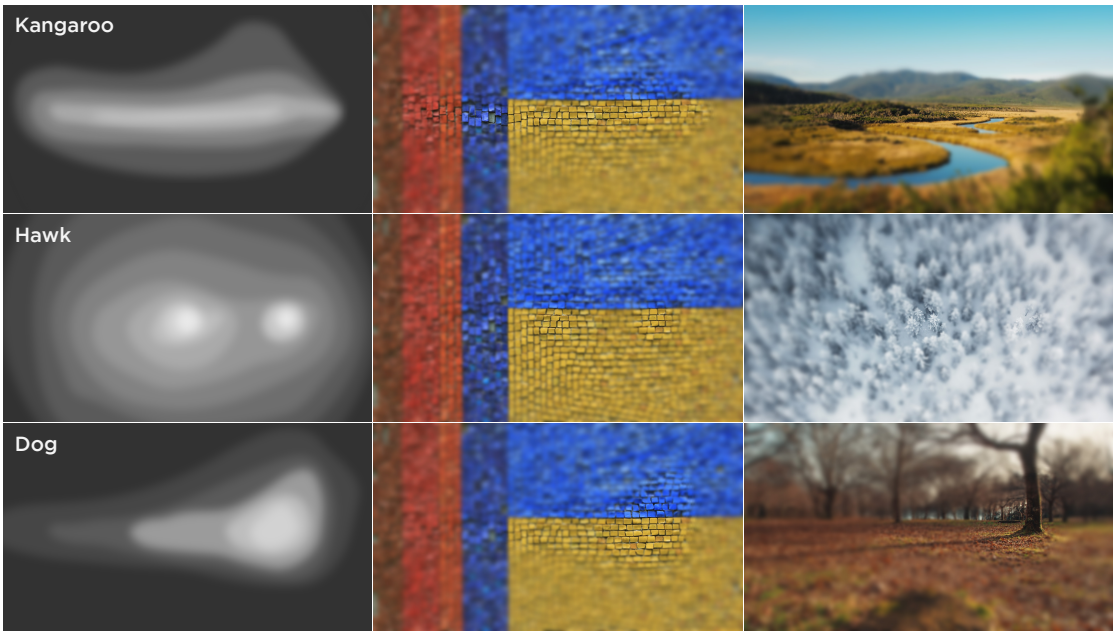


FIGURE 2.4: Foveated images based on ganglion cell distributions for different species, reproduced from (Malkin et al., 2020).

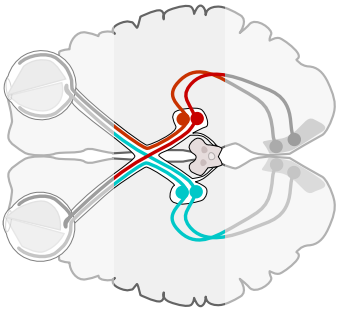
Foveal mechanisms have attracted a following within deep learning for a range of tasks such as gaze prediction (Zhang et al., 2017), object detection (Akbas and Eckstein, 2017), video processing (Wu et al., 2018) and the automated discovery of discriminative visual elements (Matzen and Snavely, 2015). A common approach to modelling foveation is image foveation. This is the process of resampling images with a foveated sampling grid and has an established use in traditional image processing for the purpose of source coding (Wang and Bovik, 2001). Some approaches to image foveation, such as the approach from (Malkin et al., 2020) shown in Figure 2.4, allow for the sampling grid to be adjusted based on ganglion cell distributions for different species. In Deza and Konkle (2020b) the authors find evidence for increased robustness to occlusions in networks trained on foveated images.

Another approach for modelling foveation is to introduce architectural modifications. In Cheung et al. (2017), the authors show that an attention mechanism equipped with a learnable sampling grid exhibits emergent foveation. Specifically, the grid nodes learn to concentrate and reduce in size towards the centre of the field of view. The type of sampling this mechanism learns to perform is analogous to foveation. The fact that this emerges in a learnable attention mechanism is a strong validation that this type of representation is desirable. Early deep visual attention mechanisms used a simple model of foveation which involves extracting the chosen region of the image at several scales before concatenating them together and passing them forward to subsequent layers (Ba et al., 2014; Mnih et al., 2014). This improves localisation ability and also enables convolutional layers to infer information regarding the size of an object by

observing its presentation over the scale space. In Chapter 6 we introduce and study a convolutional model of foveation that can be inserted into a deep network.

2.2 The Optic Chiasm and the Thalamus

The next functional unit of the visual system that we address contains the optic chiasm and the thalamus. The thalamus is comprised of densely packed cell bodies referred to as the thalamic nuclei. Two nuclei whose functions are of interest here are the Lateral Geniculate Nucleus (LGN) and the Thalamic Reticular Nucleus (TRN).



At the optic chiasm, nerve fibers are re-organised and grouped according to the side of the visual field that they are stimulated by. Following this, the fibres in the left hemisphere represent only the right visual field and vice versa. This organisation serves to support stereoscopic vision. Throughout this thesis we consider the eyes to act as one perceptive unit since stereopsis has only a limited affect (Purves and Lotto, 2003). The truth of this statement can be easily verified should the reader close one eye and note that depth perception continues largely unhindered.

Following the optic chiasm, the optic nerve innervates the LGN. The LGN has many roles, but for our purposes we focus on just two: feature extraction, and visual attention. Regarding feature extraction, the LGN performs a feature transform resembling that of retinal ganglion cells. Regarding visual attention, it has been shown that attention modulates neuronal activity in Macaque TRN and LGN (McAlonan et al., 2008). This finding constitutes empirical evidence for the speculation of Crick (1984) regarding the role of the TRN in selective visual attention.

2.2.1 Feature Extraction in the LGN

With respect to colour vision, the first major function of the LGN relates to the discovery of two broad classes of cell that respond to colour: those that exhibit opponent spectral sensitivity, and those (non-opponent) cells that do not. Experiments by De Valois et al. (1966) discovered 'spectrally opponent' cells in the LGN of a trichromatic primate which are excited by particular single-wavelength stimuli and inhibited by others. Additionally, De Valois et al. (1966) discovered that broadly speaking the cells could be grouped into those that were excited by red and inhibited by green (and vice-versa), and cells that were excited by blue and inhibited by yellow (and vice-versa). Indeed, these cells would appear to align with Hering's unique hues (red, green, blue, and yellow) (Hering, 1920), which are unique in the sense that none

of them can be viewed as a combination of the others. However, the experiments from Derrington et al. (1984) reveal that the cardinal axes of the chromatic response in the macaque LGN are not aligned to Hering's unique hues but to cone responses. The consequence of this finding is that spectrally opponent cells in early primate vision are best described as 'cone opponent'. It has similarly been argued that so called red / green opponency is better described as magenta / cyan and that these should be viewed as complementary colours, rather than opponent (Pridmore, 2005, 2011). For a more in-depth exposition of the contention between the physiological and psychophysical understanding of spectral opponency see Shevell and Martin (2017). Cells that are 'spectrally non-opponent' have also been observed in primate LGN; these are cells which are not sensitive to specific wavelengths but respond to broad range of wavelengths in the same way (either inhibitory or excitatory) (De Valois et al., 1958a; Jacobs, 1964).

Following De Valois et al.'s initial findings, there has been a realisation that cells responsive to colour could be further grouped into 'single opponent' and 'double opponent' cells. The defining characteristic of double opponent cells is that they respond strongly to colour patterns but are non-responsive or weakly responsive to full-field colour stimuli (e.g. solid colour across the receptive field, slow gradients or low frequency changes in colour) (Shapley and Hawken, 2011). In the retina, double opponency presents as spectrally opponent cells with centre-surround organisation (Troy and Shou, 2002). In the primary visual cortex, there are both the spectrally opponent cells with oriented receptive fields mentioned above and non-oriented double opponent cells in the cytochrome oxidase rich blobs (Livingstone and Hubel, 1984). Note that one interpretation is that double opponent cells are both spatially and spectrally opponent.

Lehky and Sejnowski (1999) use a four layer neural network to map cone responses to a population of Gaussian tuning curves in CIE colour space and demonstrate colour opponent neurons in the hidden layers. Wang et al. (2015) use Recursive ICA to automatically learn visual features that accord with those found in the early visual cortex. Wang et al. (2015) demonstrate that the features in the first ICA layer, trained on natural images, are oriented-edges with the colour opponent characteristics typical of V1 neurons (dark-light, yellow-blue, red-green).

In Chapter 4, we consider opponency in deep CNNs, for which some early approaches used variants of ICA to learn the filters (Le et al., 2011). Modern CNNs are trained using the back-propagation algorithm, similar to the work of Lehky and Sejnowski (1988, 1999), such that the features learned are dependent on the objective function of the model. In addition, CNNs are typically constructed with many more layers of non-linear feature extraction than the one or two layers used in ICA. As a result, CNNs permit a notion of functional organisation: 'what happens where' rather than just 'what happens'. Due to the connections between CNNs and ICA, one might

reasonably expect CNNs to exhibit emergent opponency. This is indeed the case, with multiple works pointing out that learned filters in early layers appear to be spatially and colour selective (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Lindsey et al., 2019; Rafegas and Vanrell, 2018; Olah et al., 2020a).

Rafegas and Vanrell (2018) propose an automated measurement of the spectral selectivity of convolutional neurons. For their approach, the authors find image patches which maximally excite each neuron and construct an index with high values when these patches are consistent in colour. The authors further suggest that a neuron is double opponent if it is selective to two distinct colours that are roughly opposite in hue. Note that these definitions of opponency are not direct correlates of the previously discussed definition. The key difference is that the electrophysiological definition requires an understanding of the stimuli which inhibit cells in addition to the stimuli which excite them. This is important since although cells which are excited by two colours may be projecting the input on an opponent axis, they may also just be activating for both colours indiscernibly. In Chapter 4 we introduce an automated assessment of colour opponency that alleviates this issue. The double colour selective neurons found by Rafegas and Vanrell (2018) are typically red-cyan, blue-yellow and magenta-green. These do not closely reflect the opponent axes of the primate LGN. This is to be expected since cone opponency observed in nature translates to channel opponency in a convolutional model, and so we can reasonably expect the opponent axes to be aligned with extreme RGB values rather than cone responses or Hering's unique hues (although note that these are a subset of the RGB extrema).

2.2.2 Visual Attention

Put simply, visual attention is the act of choosing what to look at. It is widely held that our visual attention is the result of two processes: bottom-up attention, and top-down attention. Bottom-up describes real-time attention driven by stimulus whereas top-down describes attention that is goal oriented and driven by context. Of these two processes, deep models of visual attention generally seek to model the former. However, deep models of visual attention may still be divided into two groups: recurrent, and non-recurrent. Although it is tempting to describe these as dynamic and static attention models, both types are typically only applied to static images and thus do not truly meet any definition of dynamic attention.

The archetypal deep recurrent visual attention model is the Deep Recurrent Attention Model (DRAM) from Ba et al. (2014) shown in Figure 2.5. It consists of two Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997): the emission network and the glimpse network. The emission network defines a glimpse policy (a series of consecutive glimpses) over the input image. The glimpse network integrates visual features from this glimpse sequence in a hidden state that can then

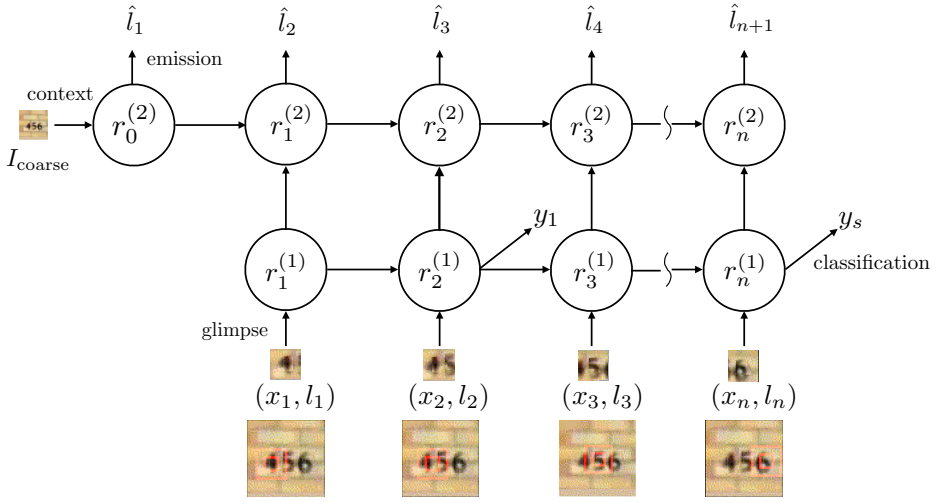


FIGURE 2.5: Deep Recurrent Attention Model (DRAM), reproduced from (Ba et al., 2014)

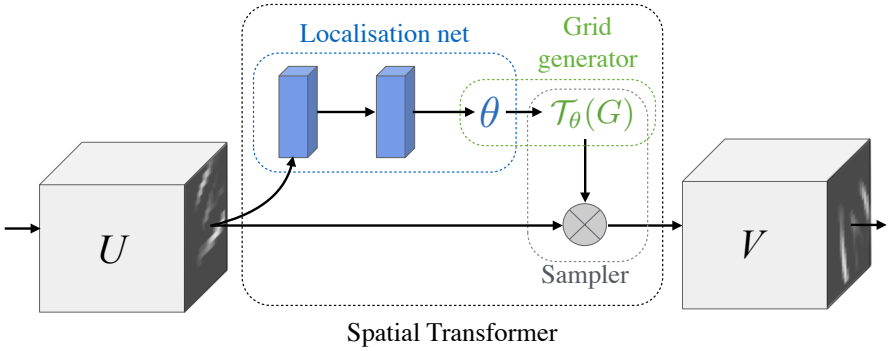


FIGURE 2.6: Spatial Transformer Network (STN), reproduced from (Jaderberg et al., 2015)

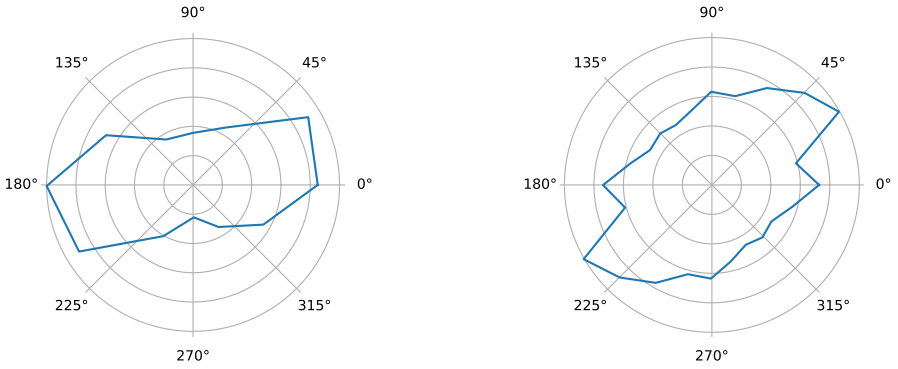
both inform future policy (by providing input to the emission network) and be used for downstream tasks such as classification. The initial input to the emission network is a context feature derived from a downsampled version of the image which could be seen as facilitating a form of top-down attention.

A non-recurrent model of visual attention that we will build on in Chapters 6 and 7 is the Spatial Transformer Network (STN) from Jaderberg et al. (2015) shown in Figure 2.6. This is a differentiable layer that regresses the co-ordinates of an affine transform which is then applied to it's input. The STN can be broken down into three stages: regression of the attention co-ordinates (localisation net), sampling of the interpolation grid (grid generator), and application of the interpolation grid to the layer input (sampler). The regression network is typically a fully connected network, although any architecture may be used, with an output for each of the six affine

co-ordinates. From these six co-ordinates an interpolation grid is then generated. This grid with two values for each output pixel which describe the co-ordinates in the input (as floating point values) that should be sampled to give the output. In the final step, these co-ordinates are then sampled from the input with bilinear interpolation. In this way, each output pixel has a gradient with respect to multiple input pixels. This layer can be inserted at any point in a network to enable it to apply a pose-normalising transform. It is assumed that this enables the network to become invariant to affine transforms of the input. However, we will later show that there are several failure modes of this approach that do not give the desired behaviour.

Yet another bisection one might make is between soft and hard attention. In hard attention a non-differentiable step is performed to extract the desired pixels from the image to be used in later operations. Conversely, in soft attention, differentiable interpolation is used. The training mechanism differs between the two approaches. Early hard attention models such as the Recurrent Attention Model (RAM) and the Deep Recurrent Attention Model (DRAM) used the REINFORCE algorithm to learn an attention policy over non-differentiable glimpses (Mnih et al., 2014; Ba et al., 2014). More recent architectures such as Spatial Transformer Networks (STNs) (Jaderberg et al., 2015), Recurrent STNs (Sønderby et al., 2015) and the Enriched DRAM (EDRAM) (Ablavatski et al., 2017) use soft attention and are trained end to end with backpropagation. The DRAM model and its derivatives use a two layer LSTM to learn the attention policy. The first layer is intended to aggregate information from the glimpses and the second layer to observe this information in the context of the whole image and decide where to look next. The attention models described thus far predominantly focus on single and multi-object classification on the MNIST and Street View House Numbers (Goodfellow et al., 2013) datasets respectively. Conversely, the DRAW network from Gregor et al. (2015) is a fully differentiable spatial attention model that can make a sequence of updates to a canvas in order to draw the input image. Here, the canvas acts as a type of working memory which is constrained to also be the output of the network.

Visual attention is intimately related to foveation. One might argue that we attend to visual scenes because our vision is foveated for efficiency, to which another might counter that our vision is foveated because it improves performance when attending to visual scenes. We explore this relationship between performance and efficiency in deep networks in Chapter 6.

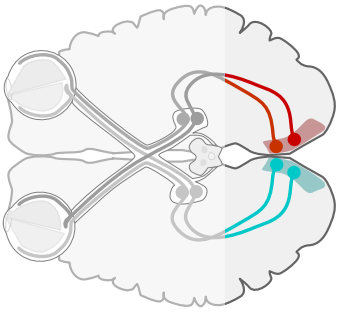


(A) Drifting Gratings (B) Flashing Bar

FIGURE 2.7: Spatial tuning curves for cells in the Mouse Lateral Geniculate Nucleus (LGN) from Zhao et al. (2013).

2.3 The Primary Visual Cortex and the Two Streams

The final functional unit of the visual system that we address contains the primary visual cortex and the ventral and dorsal visual streams (the two streams). Combined, these structures are responsible for higher order visual processing such as: the formation of a notion of object, spatial awareness and scene composition, and the construction and maintenance of visual memory.



2.3.1 Feature Extraction in the Visual Cortex

In contrast to the retina and LGN, the majority of cells in primate V1 are orientation tuned (Livingstone and Hubel, 1984). One approach to analysing this spatial selectivity involves the presentation of drifting high contrast sinusoidal gratings (Levick and Thibos, 1982; De Valois et al., 1982; Lennie et al., 1990; Johnson et al., 2001, 2008; Zhao et al., 2013). For example, one can characterise orientation selectivity through presentation of gratings with fixed frequency and contrast at a range of orientations (Levick and Thibos, 1982; Lennie et al., 1990; Johnson et al., 2008; Zhao et al., 2013). Similarly, a spatial frequency tuning curve can be obtained through the use of a fixed orientation and contrast (De Valois et al., 1982; Johnson et al., 2001). Figure 2.7 shows examples of spatial tuning curves obtained for Mouse LGN cells with two groups of stimuli, reproduced from Zhao et al. (2013). These analyses again grant a notion of spatial antagonism (spatial opponency herein) in the cortex, where there exists a grating configuration that excites the cell and an opponent grating configuration which inhibits the cell (Shapley and Hawken, 2011). Note that, although non-typical, presentation of grating stimuli have also been used to detect

centre-surround organisation in the retina (Bilotta and Abramov, 1989, e.g.) since these are cells which are highly tuned to frequency but not orientation selective.

The notion of a spatially opponent receptive field has a long history in computer vision. Notably, the Marr-Hildreth algorithm for edge detection (Marr and Hildreth, 1980) performs a Laplacian of Gaussian (often approximated by a Difference of Gaussian (DoG)) which resembles the function performed by centre-surround ganglion cells in the retina. Oriented-edge receptive fields were also modelled in early approaches to visual recognition. In particular, edge orientation histograms (McConnell, 1986; Freeman and Roth, 1995) and later histograms of oriented gradients (Dalal and Triggs, 2005) are similar in principle to a layer of neurons with oriented-edge receptive fields with different rotation, frequency, and phase. DoG and edge orientation assignment are also integral components of the well-known Scale Invariant Feature Transform (SIFT) descriptor (Lowe, 1999).

In addition to approaches which directly model opponent receptive fields, several studies have shown emergent spatial opponency in learning machines. For example, Lehky and Sejnowski (1988) found evidence for orientation selectivity in a neural network trained with back-propagation to determine the curvature of simple surfaces in procedurally generated images. Olshausen and Field (1996) demonstrated the emergence of basis functions which resemble oriented receptive fields when learning an efficient sparse linear code for a set of images. Similar results are presented by Bell and Sejnowski (1997) who show that a nonlinear 'infomax' network which performs Independent Component Analysis (ICA), trained on images of natural scenes, produces sets of visual filters that show orientation and spatial selectivity. The second layer filters in the model of Wang et al. (2015) are sensitive to edges of different frequency and orientation, reminiscent of cells in V1.

In V1, it has been suggested that cells described as selective to orientation but not colour by Livingstone and Hubel (1984) are in fact colour opponent but with unbalanced cone inputs such that they respond to general changes in luminance (Lennie et al., 1990; Johnson et al., 2001). More recently, techniques such as functional Magnetic Resonance Imaging (fMRI) have been used to explore population coding of vision and colour related processes (e.g. Engel et al., 1997; Seymour et al., 2015; Boynton, 2002; Wade et al., 2008). In particular, studies have shown strong responses in V1 to stimuli that are preferred by spectrally opponent cells (Kleinschmidt et al., 1996; Engel et al., 1997; Schluppeck and Engel, 2002). The work of Wade et al. (2008) validates that the early visual system of the macaque (where many of the single-cell measurements of colour vision have been taken) correlates strongly with humans in terms of overall population responses to chromatic contrast; this is important to our work in Chapter 4 since we seek functional archetypes that are of general efficacy in visual intelligence.

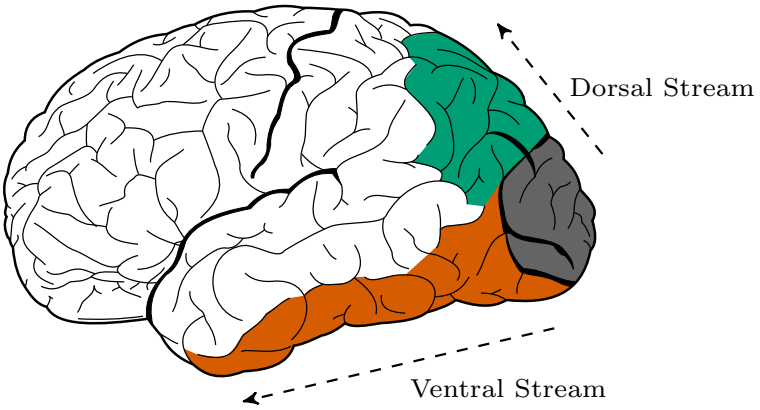


FIGURE 2.8: The two streams hypothesised by Goodale and Milner (1992). The ventral stream, the 'what' pathway, starts in the primary visual cortex and flows into the temporal lobe (located on the ventral surface of the brain). The dorsal stream, the 'where' pathway, again starts in the primary visual cortex but flows into the parietal lobe (located on the dorsal surface of the brain).

2.3.2 What and Where

The two streams hypothesis suggests that in human vision there exists a 'what' pathway and a 'where' pathway (Goodale and Milner, 1992). The 'what' pathway is concerned with recognition and classification tasks and the 'where' pathway is concerned with spatial attention and awareness. Figure 2.8 depicts the location of two streams in the brain.

In computer vision, the vast majority of recent efforts have focused solely on object recognition and as such the ventral visual stream is typically cited as the basis of any biological inspiration. In contrast, in the DRAM model discussed in Section 2.2.2, a multiplicative interaction between the pose and feature information, first proposed by Larochelle and Hinton (2010), is suggested to emulate the role of the two streams. Humans are endowed with two methods of spatial understanding. The first is the ability to infer pose from visual cues and the second is the knowledge we have of our own pose (Gibson, 1950). The pose features in the DRAM model derive from the location of the glimpse and thus act as an analogue of the latter. It is expected that the glimpse CNN (which derives visual features from each glimpse) can account for the former. In Chapter 6 we will study the ability of CNNs to infer pose and show that it can be improved dramatically by foveation.

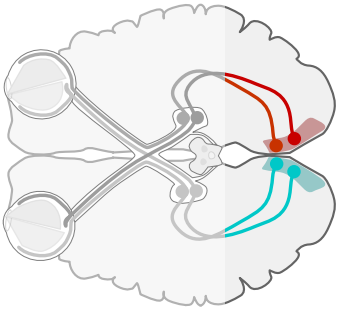
2.3.3 Memory

Memory is generally split into three temporal categories, immediate memory, short term (working) memory and long term memory (Purves et al., 1997). Weights learned through error correction procedures in a neural network are analogous to a long term

memory in that they change gradually over the course of the models existence. We can go further to suggest that the activation captured in the hidden state of a recurrent network corresponds to an immediate memory. There is, however, a missing component in current deep architectures, the working memory. One model of working memory is that of Baddeley and Hitch (1974) which includes a visual sketchpad that allows individuals to momentarily create and revisit a mental image that is pertinent to the task at hand. This sketchpad requires the integration of 'what' and 'where' information from the two streams. In Chapter 7 we study a deep network that can reconstruct it's input through a series of sketches inspired by the notion of a visual sketchpad.

Chapter 3

Colour and Classification



Providing a convolutional network with colour images instead of greyscale gives only a mild improvement in classification performance. The reason for this seems simple: colour information is not particularly important to the classification task. However, several works have explored colour processing in deep convolutional networks (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Lindsey et al., 2019; Rafegas and Vanrell, 2018; Olah et al., 2020a). If colour information is not of value to the classification task, then that would call in to question the validity of analysing colour representations in classification networks. This raises the important question of how we can determine if a network relies on colour to inform it's decisions.

Visualising learned features is well established in deep learning. The work of Erhan et al. (2009) used gradient ascent to construct an input (a super-stimulus) that maximised the networks prediction logit for a particular class. The super-stimuli produced by these early works are noisy images that rarely resemble the target class. In fact, such unconstrained gradient ascent was also used in early demonstrations of the phenomenon of adversarial examples (Szegedy et al., 2013). Many subsequent visualisation works, particularly the works of Olah et al. (2017) and Mordvintsev et al. (2018), have built on this to propose mechanisms through which more visually appealing super-stimuli may be constructed. However, we believe that the full extent of the value that these approaches can provide has not yet been demonstrated.

In this chapter we contend that super-stimuli may hold the key to understanding whether ImageNet-trained networks use colour to make their decisions. We propose generating super-stimuli for a set of classes that are all characterised by a particular colour and assessing whether these colours are present. We start by reviewing approaches for generating super-stimuli. Next, we define a set of fruit classes, chosen

for their characteristic colours, and generate super-stimuli for each using the reviewed methods for a range of network architectures. Our results show that the characteristic colours are recovered in the generated super-stimuli for some architectures and not others. This presents a concrete demonstration that convolutional networks do sometimes use colour information for their classifications. Anecdotally, our results suggest that the characteristic colours are recovered more readily in the super-stimuli of more performant architectures. We further note that the super-stimuli, particularly those produced by a Compositional Pattern Producing Network (CPPN) following the approach of Mordvintsev et al. (2018), capture the characteristic shape of the target class in many cases. This perhaps disagrees with the conventional wisdom and key results such as the findings of Geirhos et al. (2019), although it may also be considered as complementary to this work since the authors only argue that CNNs are biased towards texture rather than wholly consumed by it. We subsequently extend our approach to propose an additional set of classes that are characterised jointly by their colour and pattern: animals. We again find good evidence for the use of colour information when classifying images as one of these classes. Regarding shape, the super-stimuli produced for the animal set arguably capture the characteristic shapes of their target classes even more closely. This is again most evident for stimuli generated with a CPPN.

3.1 Feature Visualisation

In this section, we review approaches for constructing super-stimuli, particularly following the work of Olah et al. (2017). In essence this constitutes a collection of tricks and approaches that each improve the quality of generated visualisations. We go on to review an approach to using Compositional Pattern Producing Networks (CPPNs) from Stanley (2007) applied to the construction of super-stimuli by Mordvintsev et al. (2018).

3.1.1 Gradient Ascent With Raw Pixels

To establish a baseline, we first show the result of gradient ascent in pixel space without any amendments in Figure 3.1. The image was produced by optimizing the input of a ResNet-18 to maximise the output for the hummingbird class (94). Although some mild texture is present, the image is noisy and bears no resemblance to a hummingbird. As previously mentioned, the lack of resemblance relates to the well known failure of CNNs to generalise out of distribution and there susceptibility to tiny adjustments of the input that is also responsible for the phenomenon of adversarial examples (Szegedy et al., 2013). To produce more interpretable super-stimuli, we must therefore look to constrain the input to the space of natural images.

FIGURE 3.1: Gradient ascent in pixel space to maximise the output for the hummingbird class (94) of an ImageNet-trained ResNet-18.

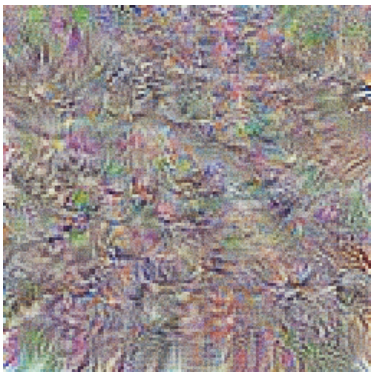


FIGURE 3.2: Gradient ascent in colour-corrected pixel space to maximise the output for the hummingbird class (94) of an ImageNet-trained ResNet-18.

3.1.2 Colour Correlation

The first approach taken by Olah et al. (2017) to improve the quality of generated super-stimuli is to perform gradient ascent in a de-correlated pixel space. The motivation for this is that gradient ascent treats the channels of the input as independent. However, in reality, the R, G, and B values of natural images are not independent and are instead correlated. As a result, gradient ascent on the pixel values can produce colours that would not be found in natural images. To address this, the authors propose to correlate the colour channels of the input according to statistics taken from the ImageNet data set before passing it to the model so that gradient ascent is performed in a de-correlated colour space. Denoting the de-correlated channels as XYZ, we obtain the RGB model input with

$$\begin{vmatrix} 0.5628 & 0.1948 & 0.0433 \\ 0.5845 & 0.0000 & -0.1082 \\ 0.5845 & -0.1948 & 0.0649 \end{vmatrix} \begin{vmatrix} \mathrm{X} \\ \mathrm{Y} \\ \mathrm{Z} \end{vmatrix} = \begin{vmatrix} \mathrm{R} \\ \mathrm{G} \\ \mathrm{B} \end{vmatrix}.$$

Figure 3.2 shows the result of the hummingbird ascent in the de-correlated colour space. The resultant super-stimulus now contains much more natural colours in

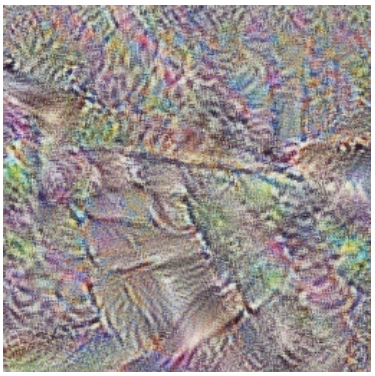


FIGURE 3.3: Gradient ascent in colour-corrected pixel space with random spatial shifts, scales, and rotations to maximise the output for the hummingbird class (94) of an ImageNet-trained ResNet-18.

comparison to the result from Figure 3.1.

3.1.3 Augmentation

The second technique that can be employed to improve the quality of generated super-stimuli is data augmentation. This regularisation technique is commonly used elsewhere in the deep learning literature and consists of applying random transforms to the input at each step. We can use this technique in gradient ascent with one caveat: the augmentations must be differentiable. Many different augmentation techniques can be employed and the exact recipe can be tuned to obtain the best results, however, for our purposes we have found a combination of random scale, random rotation, and spatial jitter (where the input is shifted vertically and horizontally by a random number of pixels) introduced by Mordvintsev et al. (2015) to work well. By employing this technique, we force the super-stimulus to represent the target in a way that is robust to spatial shifts, changes in scale, and rotations. Figure 3.3 shows the result of the hummingbird ascent with the addition of data augmentation. The visual artefacts present in this super-stimulus are much larger and more defined than in the previous attempts and it can now be said to resemble in some way a hummingbird.

3.1.4 Fourier Space

We have seen how results can be improved by performing ascent in a decorrelated colour space. Another form of correlation prevalent in natural images is spatial correlation. The final innovation of Olah et al. (2017) is to perform gradient ascent on the parameters of an image in Fourier space. The motivation for this is that a spatially consistent (that is, equal in all spatial positions) correlation can be modelled by independent variables in Fourier space. An example given by the authors to illustrate this face is to consider that a convolution can express a spatially consistent correlation.

FIGURE 3.4: Gradient ascent in colour-corrected Fourier space with random spatial shifts, scales, and rotations to maximise the output for the hummingbird class (94) of an ImageNet-trained ResNet-18.

FIGURE 3.5: Gradient ascent on the parameters of a colour-corrected CPPN with random spatial shifts, scales, and rotations to maximise the output for the hummingbird class (94) of an ImageNet-trained ResNet-18.

In Fourier space, by the convolution theorem, such a convolution becomes a pointwise multiplication. To implement this technique, we differentiably apply the inverse Fourier transform to the image parameters before then colour correcting and augmenting the image. Figure 3.4 shows the super-stimulus generated with this technique. Two observations can be made regarding this result. First, it resembles a natural image much more closely than our previous attempts. Second, the super-stimulus now recovers textures, patterns, and colours that we would tend to associate with a hummingbird. This result shows that although there is a space of super-stimuli that do not resemble the target class according to a human observer, there is a also a space of equivalent super-stimuli that do. This approach is of most utility when we wish to study super-stimuli that are constrained to the realm of natural images rather than out-of-distribution artefacts.

3.1.5 Compositional Pattern Producing Networks (CPPNs)

An alternative image parametrisation in the same vein as gradient ascending in Fourier space is the Compositional Pattern Producing Network (CPPN) (Stanley, 2007). A CPPN is a network which maps coordinates x and y to RGB values. In the context of the production of super-stimuli, Mordvintsev et al. (2018) perform gradient ascent to

learn the weights of a CPPN. The CPPN is implemented as a multi-layer CNN with 1×1 filters whose input is a grid of coordinates. The result of this parametrisation for the hummingbird ascent (again using colour correction and augmentation) is given in Figure 3.5. The figure appears glassy and with bold colours and shapes, a key characteristic of super-stimuli generated with this approach. This example arguably resembles a hummingbird most closely out of all the techniques demonstrated here, capturing its vibrant colours and shape characteristics.

3.2 Methods

In this section we introduce our approach to characterising how classifiers use colour through the production of super-stimuli. We introduce the notion of a trait group; a collection of classes that are characterised by a particular property such as a characteristic colour or pattern. We subsequently specify two such trait groups that will be used in our experiments.

Humans naturally form a strong association between object classes and particular attributes or traits. For example, it is unambiguously true that we associate oranges with being orange. Indeed, if shown a picture of an orange and asked to explain how you know it is an orange, while your may be extraordinarily verbose it would almost surely include a statement of this simple fact.

We take the view that super-stimuli represent a collection of the visual features that are required in order for the network to classify an image as belonging to the target class. As such, an effective test of whether CNN classifiers use colour information in a manner that is at all sensible is to produce super-stimuli for classes that are well characterised by their colour. For example, if a network does not effectively require that an image classified as containing an orange is orange in colour then, we argue, it is likely that the network does not make meaningful use of colour anywhere. It would, however, be flippant to form such a judgement on the basis of super-stimuli for a single class. We therefore suggest using a group of classes that are all well characterised by a common trait.

For our experiments, the first trait group we define is that of fruits (with the corresponding ImageNet class index given in brackets)

$$\begin{aligned}\text{fruits} = \{&\text{strawberry}(949),\\ &\text{lemon}(951),\\ &\text{apple}(948),\\ &\text{orange}(950),\\ &\text{pomegranate}(957),\\ &\text{pineapple}(953)\}\ .\end{aligned}$$

These fruit classes are all well characterised by their colour.

Recent efforts, most notably from Geirhos et al. (2019), have shown that ImageNet-trained CNNs are biased towards texture information and tend to ignore shape. It is therefore relevant to ask whether networks rely on both colour and texture / pattern when classifying an image as a tiger for example, or on texture / pattern alone. To this end, we additionally define the animals trait group (with the corresponding ImageNet class index given in brackets)

$$\begin{aligned}\text{animals} = \{&\text{tiger}(292),\\ &\text{cheetah}(293),\\ &\text{brown bear}(294),\\ &\text{badger}(362),\\ &\text{zebra}(340),\\ &\text{ladybird}(301)\}\ .\end{aligned}$$

These classes are all jointly characterised (at least to human observers) by a combination of colour and texture / pattern.

3.3 Results

In this section we report our results producing super-stimuli for a range of ImageNet-trained convolutional networks for our two trait groups: fruits and animals. We compare super-stimuli for the following pretrained architectures: AlexNet (Krizhevsky et al., 2012), VGG-16 (Simonyan and Zisserman, 2014), Inception-V1 (Szegedy et al., 2015), MobileNet-V2 (Sandler et al., 2018), and ResNet-50 (He et al., 2016). To produce the super-stimuli, we use both the Fourier space technique from Olah et al. (2017) and the CPPN technique from Mordvintsev et al. (2018) that were explored in Section 3.1.

FIGURE 3.6: Super-stimuli for the 'fruits' trait group for a range of ImageNet-trained CNNs obtained by gradient ascent in Fourier space following Olah et al. (2017).

3.3.1 Fruits - Fourier Space

Figure 3.6 gives super-stimuli for the 'fruits' trait group produced by gradient ascent on image parameters in Fourier space following Olah et al. (2017). The results show that the networks do, on the whole, characterise the fruit classes by their colour. This is particularly the case for ResNet-50 where all of the super-stimuli include the correct colour for their corresponding class. The super-stimuli for AlexNet, VGG-16, and Inception-V1 do generally include the characteristic colour, but also multiple vibrant artefacts that would not typically be associated with the fruit classes such as blues and purples. These colours are also present in the super-stimuli for the other models but are much less dominant. This suggests that AlexNet, VGG-16, and Inception-V1 are less selective regarding colour when classifying images as one of our fruit classes compared to MobileNet-V2 and ResNet-50.

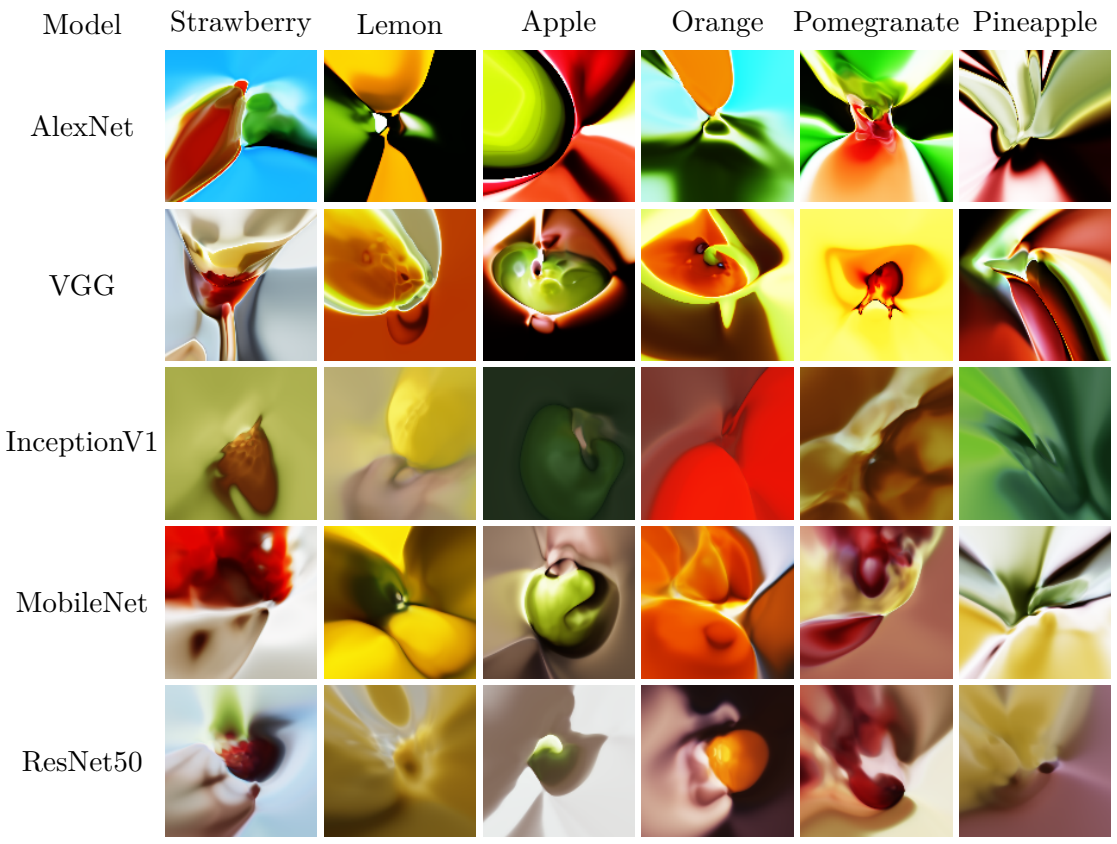


FIGURE 3.7: Super-stimuli for the 'fruits' trait group for a range of ImageNet-trained CNNs obtained by gradient ascent on the parameters of a CPPN following Mordvintsev et al. (2018).

3.3.2 Fruits - CPPN

Figure 3.7 gives super-stimuli of the ImageNet-trained models for the 'fruits' trait group obtained by learning the parameters of a CPPN following Mordvintsev et al. (2018). The results again demonstrate that CNNs do generally associate each fruit class with its characteristic colour. The observation that AlexNet, VGG-16, and Inception-V1 are not as selective regarding colour is reiterated here. For example, the pineapple super-stimuli do not include any of the characteristic yellow-orange of pineapple flesh and instead include intense reds that would be construed as pineapple coloured. Similarly, the orange super-stimuli for these networks mostly populated by greens, blues, and reds. Somewhat strikingly, the super-stimuli produced with this approach include characteristic shapes that are associated with the target class. For example, the Inception-V1 and ResNet-50 examples both depict a somewhat realistic strawberry shape. This is particularly true for the ResNet-50 which correctly places a green stalk at the top of the strawberry. For another example, consider the MobileNet-V2 super-stimulus for pomegranate, where seeds are enveloped by a layer of pith. This disagrees somewhat with the findings of Geirhos et al. (2019) which would suggest that we should predominantly recover texture and colour in these super-stimuli.

FIGURE 3.8: Super-stimuli for the 'animals' trait group for a range of ImageNet-trained CNNs obtained by gradient ascent in Fourier space following Olah et al. (2017).

3.3.3 Animals - Fourier Space

In Figure 3.8 we give super-stimuli for the 'animals' trait group produced by gradient ascent on image parameters in Fourier space following Olah et al. (2017). In contrast with the results for the 'fruits' group, the images here do not reflect the characteristic colours as clearly. Instead, the images presented here predominantly show the characteristic texture and occasionally shape cues of their target classes. For example, although there is a slight difference in colour between the tiger and cheetah images for AlexNet they are more clearly differentiated by the different patterns (stripes and spots) present. There are several exceptions to this, however. For example, all networks except Inception-V1 recover both the characteristic shape and deep red colour of a ladybird. There are additional cases where characteristic shapes are recovered. In particular, the bear images include multiple allusions to bear ears. Similarly, the zebra stimulus for MobileNet-V2 appears to resemble the hind legs of a zebra.

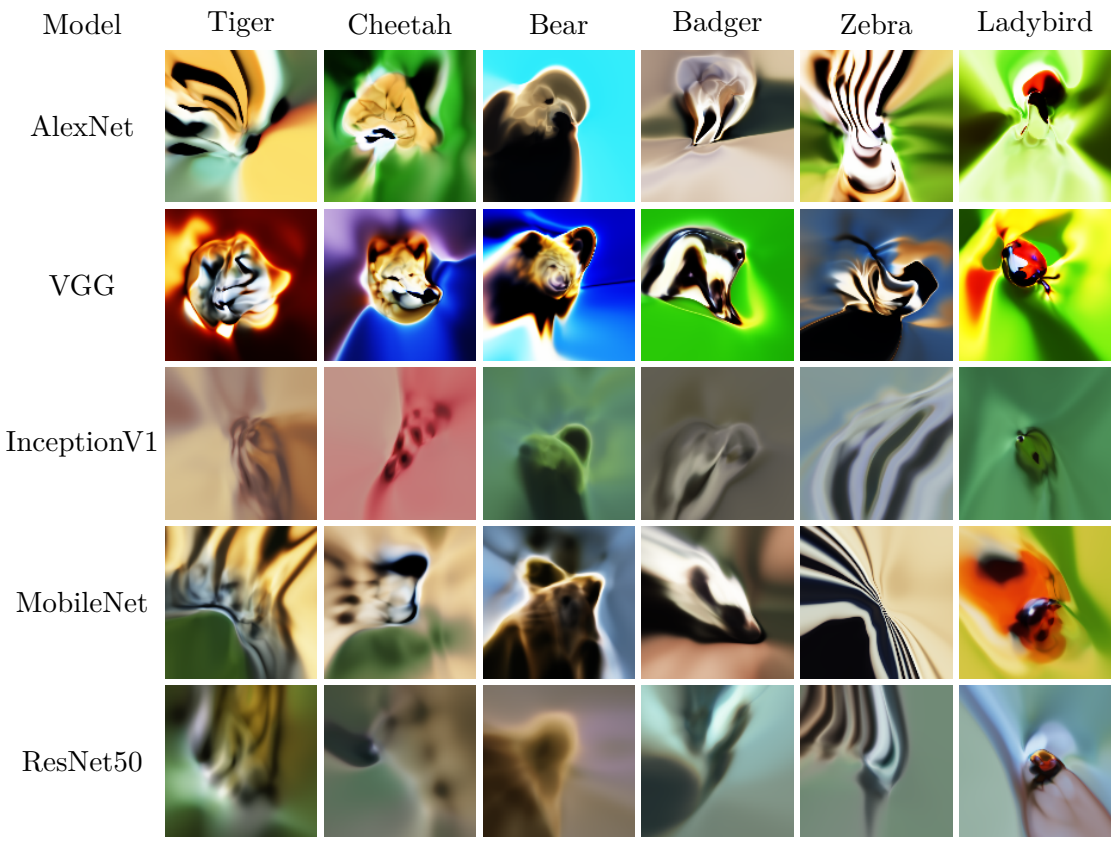


FIGURE 3.9: Super-stimuli for the 'animals' trait group for a range of ImageNet-trained CNNs obtained by gradient ascent on the parameters of a CPPN following Mordvintsev et al. (2018).

3.3.4 Animals - CPPN

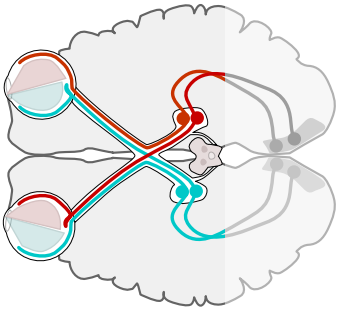
Figure 3.9 shows the super-stimuli obtained for the 'animals' trait group by optimising the parameters of a CPPN following Mordvintsev et al. (2018). Several interesting observations can be made regarding this figure. Firstly, the images do generally capture the characteristic colours of their corresponding classes. This is true with the notable exception of the Inception-V1 images. Here, the cheetah is given as red and the bear and ladybird as green. For the VGG-16, the tiger and cheetah images appear to be entirely noise artefacts without any real resemblance to their target classes. The second interesting observation that can be made is that the MobileNet-V2 and ResNet-50 images recover high-quality characteristic shape cues such as: a bears ear, a badgers snout, a tigers leg, or a ladybirds shell. This constitutes a clear demonstration that certain ImageNet-trained networks do employ an uncanny representation of shape when classifying images.

3.4 Summary

In this chapter we have considered whether CNNs use colour to their advantage when classifying images. To this end, we reviewed and re-implemented approaches to the production of super-stimuli with the aim of generating super-stimuli for classes that are well characterised by colour. Next, we proposed two groups of such classes which we refer to as trait groups: fruits, and animals. We generated stimuli for these groups for a range of CNN architectures using gradient ascent in Fourier space and gradient ascent on the parameters of a CPPN. Our results show that the characteristic colours of the classes are recovered in the super-stimuli. Furthermore, the super-stimuli for some models bore an uncanny likeness to their subjects not just in terms of colour but also in terms of shape and composition, particularly when generated with a CPPN. It could be suggested that there is an inductive bias in the CPPN approach that causes this shape likeness. However, although such a bias could explain the production of natural shapes, it cannot account for the recovery of characteristic shapes. Instead, the presence of characteristic shapes can only be explained by a selectivity of the model (if the model places a greater weight on colour and texture cues as observed by Geirhos et al. (2019)). Ultimately, our work here proposes a new application for super-stimuli when twinned with targets that are characterised by a particular trait of interest and presents new evidence for the representation of shape in deep convolutional networks.

Chapter 4

Architecture and Opponency



The tendency for learning machines to exhibit oriented-edge receptive fields, similar to those found in nature, has long been observed (Lehky and Sejnowski, 1988; Olshausen and Field, 1996; Bell and Sejnowski, 1997; Shan et al., 2007; Wang et al., 2015; Krizhevsky et al., 2012; Lindsey et al., 2019; Olah et al., 2020a). However, learning machines rarely exhibit the functional organisation found in nature. In convolutional neural networks, we typically find oriented-edge receptive fields in early layers, rather than a progression from centre-surround receptive fields to oriented-edge receptive fields as is common in biological vision (Hubel and Wiesel, 2004). In an important work, Lindsey et al. (2019) demonstrate that the addition of a bottleneck to a deep convolutional network can induce centre-surround receptive fields, suggesting a causal link between anatomical constraints and the nature of learned visual processing. In order to refine our understanding of this causal relationship, we pursue an electrophysiological interpretation of convolutional networks which incorporates opponency and colour tuning.

Cells with centre-surround and oriented edge receptive fields are spatially opponent. From the classic work of Kuffler (1953), Hubel and Wiesel (1962, 2004) and others (summarised in Troy and Shou (2002) and Martinez and Alonso (2003)), these neurons form the building blocks of feature extraction in the primary visual cortex. Formally, a neuron that is excited by a particular stimulus and inhibited by another in the same stimulus space is said to be opponent to that space. For example, if a neuron is excited by stimulus in some part of the visual field and inhibited in another, it is spatially opponent. Alternatively, if a neuron is excited by stimulus of a certain wavelength and inhibited by stimulus of another, it is spectrally opponent. Spectral opponency, first hinted at by the complementary colour system from Goethe (1840) and later detailed by Hering (1920), was only observed and characterised at a cellular level around 1960

(De Valois et al., 1958b; Wagner et al., 1960; Wiesel and Hubel, 1966; Naka and Rushton, 1966; Daw, 1967). Combined, the theories of spatially opponent feature extraction in the visual cortex (Kuffler, 1953; Hubel and Wiesel, 1962, 2004; Troy and Shou, 2002), trichromacy (Young, 1802; Helmholtz, 1852; Maxwell, 1860), and spectral opponency (De Valois et al., 1966) constitute a deep understanding of the early layers of visual processing in nature.

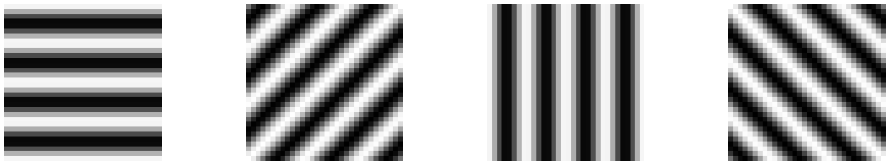
The notional elegance of the above theories has served to motivate much of the progress made in computer vision, most notably including the development of multi-layer (deep) Convolutional Neural Networks (CNNs) (Le Cun et al., 1990; Bottou et al., 1994; Le Cun et al., 1995) that are now so focal in our collective interests. Multi-layer CNNs are learning models designed to mimic the functional properties, namely spatial feature extraction and retinotopy, of the retina, Lateral Geniculate Nucleus (LGN), and primary visual cortex. By virtue of the ease with which one can train such models, multi-layer CNNs offer a unique opportunity to study the emergence of visual phenomena across the full gamut of constraints and conditions of interest. It is widely observed that trained convolutional neurons experience the same kinds of receptive fields as those found in nature, and that the learned features become successively more abstract with depth (Krizhevsky et al., 2012; Zeiler and Fergus, 2014; Olah et al., 2017, 2020a). However, we do not typically see structural organisation of these cell types. For example, edge and colour information is confounded in the first layer of ZFNet (Zeiler and Fergus, 2014), with some colour information also encoded in the second layer. Furthermore, as addressed by Lindsey et al. (2019), none of the convolutional neurons have centre-surround receptive fields of the kind observed in retinal ganglion cells. Rafegas and Vanrell (2018) analysed colour selectivity in a deep CNN, finding cells which are excited by two groups of stimuli that are roughly opposite in hue. To classify these cells as opponent would additionally require an understanding of the stimuli which inhibit each cell. There has been some exploration of the role of inhibition in deep CNNs (Olah et al., 2018), although we are not aware of any demonstration that learned convolutional cells are ever truly opponent in the sense that they are both inhibited below and excited above a baseline by some stimuli.

With the exception of recent developments in meta learning (e.g. Zoph and Le, 2016; Tan and Le, 2019), new convolutional architectures are typically designed with the aim of increasing either width (Zagoruyko and Komodakis, 2016) or depth (Szegedy et al., 2015; He et al., 2016) whilst preventing the vanishing gradient problem with: auxiliary losses (Szegedy et al., 2015), skip connections (He et al., 2016), dense connections (Huang et al., 2017), or stochastic depth (Huang et al., 2016) to name a few. However, the finding by Lindsey et al. (2019) that network architecture can impact the fundamental 'type' of function that is learned (rather than simply affecting capacity) suggests a new approach to both architecture design and interpretability. Specifically, if we can improve our understanding of the bias introduced by the network

architecture, we may be able to design new architectures with specific goals in mind or better interpret the performance of pre-existing ones.

Clearly, research in this space has the potential to impact our understanding of both deep learning and the neuroscience of vision. In order to realise this potential, large-scale studies are needed which properly establish the connections between the model architecture, the data space, and the kind of visual processing that is learned. Lindsey et al. (2019) mainly rely on qualitative assessment for the identification of centre-surround and oriented edge receptive fields, but do propose some quantitative analyses such as the variance in gradient with respect to different inputs as a measure of the linearity of the neuron. The highly detailed analyses of Olah et al. (2020b) give a comprehensive understanding of the function of particular neurons or circuits in deep networks, however, each functional unit or group is currently identified manually. The procedure proposed by Rafegas and Vanrell (2018) could be automated but involves the costly process of determining image patches which most excite each cell. The Brain-Score project from Schrimpf et al. (2018) is an attempt at providing an assessment of the similarity between a given network and various neural and behavioural recordings from primates. This is uninformative in the sense that it does not provide any information regarding precisely how the function of the network is similar to that of the primate visual system. The same could be said of the work of Gomez-Villa et al. (2019), who find evidence that CNNs are susceptible to the same visual illusions as those that fool human observers.

In this chapter, we develop a framework for automatically classifying convolutional cells in terms of their spatial and colour opponency, based on electrophysiological definitions from the neuroscience literature. In addition, we propose a method of obtaining a hue sensitivity curve for a given network, inspired by similar methods in psychophysics. Combined, these approaches provide a descriptor of the functions learned by CNNs that provides rich insight into how they encode information. We apply our framework on a colour variant of the model from Lindsey et al. (2019) and demonstrate that, following the introduction of a bottleneck, different cell types tend to be organised according to their depth in the network, with no such organisation found in networks without a bottleneck. We detail the relationship between data, architecture, and learned representation through a series of control experiments. In total, we have trained 2490 models over 9 different settings, all of which have been made publicly available, alongside code for all of our experiments, via PyTorch-Hub at `https://github.com/ecs-vlc/opponency`.



(A) $\theta = 0°$ (B) $\theta = 45°$ (C) $\theta = 90°$ (D) $\theta = 135°$

FIGURE 4.1: Examples of grating patterns used as stimuli for the spatial opponency experiments. These samples have been generated using PsychoPy (Peirce et al., 2019), with different angles (θ), frequency of 4, and phase of 0.

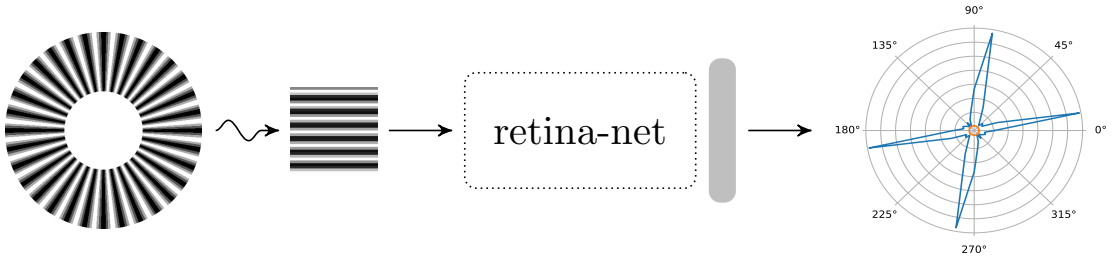


FIGURE 4.2: Our approach for obtaining a spatial tuning curve for cells in a deep network.

4.1 Methods

In this section we detail our methodology for classifying convolutional cells according to their spatial and colour processing. Generalising physiological definitions discussed in Section 2.1.2, to classify a cell as opponent we require: a set of stimuli, the ability to measure the response of the cell to each stimulus, and a measurement of the baseline response of the cell (in order to establish excitation and inhibition). The response of each neuron to the input is readily available in a deep network, and we define the baseline as the response of the cell to a black input (a matrix of zeros). If there exists a stimulus for which the cell is excited (responds above the baseline) and a stimulus for which the cell is inhibited (responds below the baseline), then the cell is opponent to the axis of variance of the stimuli set. We first describe the two stimuli sets that we will use for the classification of spatial and colour opponency. We go on to discuss automatic classification of a cell as double opponent and how we can infer the specific 'type' of an opponent cell. In addition, we introduce an approach for studying the hue sensitivity curve of a deep network, inspired by Bedford and Wyszecki (1958) The experiments laid out in this section will form our core results. We will later perform a control study to determine how well these results extend to different settings.

4.1.1 Spatial opponency

To classify spatial opponency, we require a set of stimuli that vary spatially. Following Johnson et al. (2001), we construct a set of high contrast greyscale gratings produced

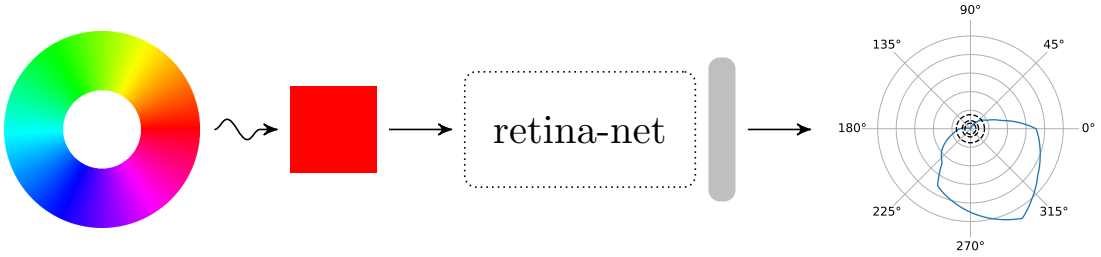


FIGURE 4.3: Our approach for obtaining a colour tuning curve for cells in a deep network.

from a sinusoidal function for a range of rotations, frequencies, and phases. Figure 4.1 gives some example stimuli generated using PsychoPy (Peirce et al., 2019) with various degrees of rotation. We compute the response of each cell to each stimulus and compare these to the baseline to obtain a tuning curve which can be used to perform an automatic classification. Figure 4.2 depicts our process for obtaining the spatial tuning curve of a cell.

In addition to spatially opponent, we have spatially non-opponent and spatially unresponsive. A non-opponent cell is one which may be excited or inhibited by the stimuli but does not cross the baseline. An unresponsive cell is one which is neither excited nor inhibited by any of the stimuli. We do not use any form of tolerance when making the above classifications. The reason for this is that each cell will activate in its own space, and so the relative effect of a fixed tolerance could vary greatly between cells. As a consequence of this design decision, it is likely that the output of any unresponsive cells lies in a clipped region of the activation function. For example, if using Rectified Linear Units (ReLUs), the cell response may remain at 0 for all of the stimuli. Such cells are either highly tuned to a particular, complex stimulus or merely unresponsive to all stimuli. That said, our interests here are primarily bound to the existence and distribution of opponent and non-opponent cells only; we are not aware of any demonstration that unresponsive cells are found in nature. Recall that although the described stimuli are oriented edges, they can still be used to infer centre-surround spatial opponency since a cell with a characteristic centre-surround receptive field would be highly tuned to a particular frequency and phase, but responsive to a broad range of angles.

4.1.2 Colour opponency

To classify spectral opponency, De Valois et al. (1966) vary the stimuli according to wavelength. For our experiments we propose using stimuli which vary in hue, rather than wavelength. The reason for this is that the trained networks will expect an RGB input and there is no exact mapping from wavelength to RGB. We could consider a more biologically valid colour representation such as the cone response space used by

Lehky and Sejnowski (1999) but opt for RGB as it is the standard practice in deep learning. Following the approach shown in Figure 4.3, we sample colours in the Hue, Saturation, Lightness (HSL) colour space for all integer hue values with saturation of 1.0 and lightness of 0.5. We then convert our stimuli to RGB before forwarding to the network and constructing the colour tuning curve. We can perform classification by following the same process of comparing to the baseline as in the spatial setting. We use the terms hue opponency, and colour opponency interchangeably to refer to the different cell types found through this process.

4.1.3 Double opponency

As discussed, we can automatically classify a cell as double opponent if it is both colour and spatially opponent. Our interests here lie in whether or not double opponent cells emerge in convolutional networks trained with a classification objective. Note that it has been observed that most spectrally opponent cells in macaque V1 are also orientation selective (Johnson et al., 2008), that is, they are double opponent. Unlike in the single opponent cases, we do not define a notion of double non-opponency or double unresponsiveness (although such classifications could be made if required).

4.1.4 Excitatory and inhibitory colours

Using the colour tuning curve, we can further determine the hue which most excites or inhibits each cell. Since cells are typically equipped with a non-linear activation function, there may be a wide range of stimuli for which they produce the lowest response. As such, we use the pre-activation output to infer the most inhibitory stimulus. This excitation and inhibition data will allow us to plot the distribution of colours to which cells in networks are tuned. Note that this distribution is insufficient to describe the type of opponency since it does not permit an understanding of whether there are distinct classes of opponent cell. For example, the distribution of excitation and inhibition does not distinguish between two groups of cells that are red / green opponent and blue / yellow opponent respectively, or many groups of cells that are red / green opponent, green / blue opponent, blue / red opponent etc. One option would be to applying a clustering technique to the most excitatory and inhibitory responses. However, this would introduce additional challenges through the need for appropriate algorithm and hyper-parameter choice. Instead, we can additionally study the conditional distribution of maximal excitation, given maximal inhibition by some colours in a chosen range. We suggest evaluation of these conditional distributions for the following hue ranges: red ($[315°, 45°)$), yellow ($[45°, 75°)$), green ($[75°, 165°)$), cyan ($[165°, 195°)$), blue ($[195°, 285°)$), and magenta ($[285°, 315°)$). By enabling direct assessment of the inhibition / excitation pairs, this will give a much deeper understanding of the kinds of opponency present in the networks being analysed.

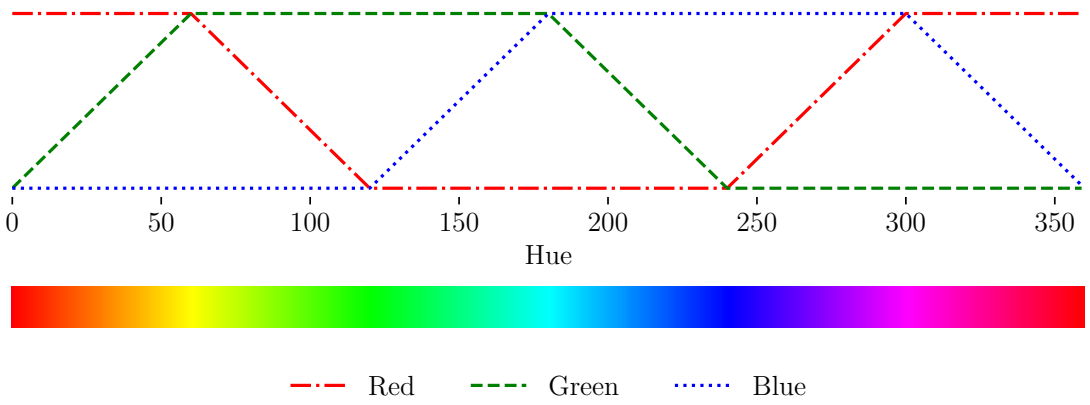


FIGURE 4.4: Plot of the Red, Green and Blue components of the Hue to RGB conversion. This function is piece-wise differentiable.

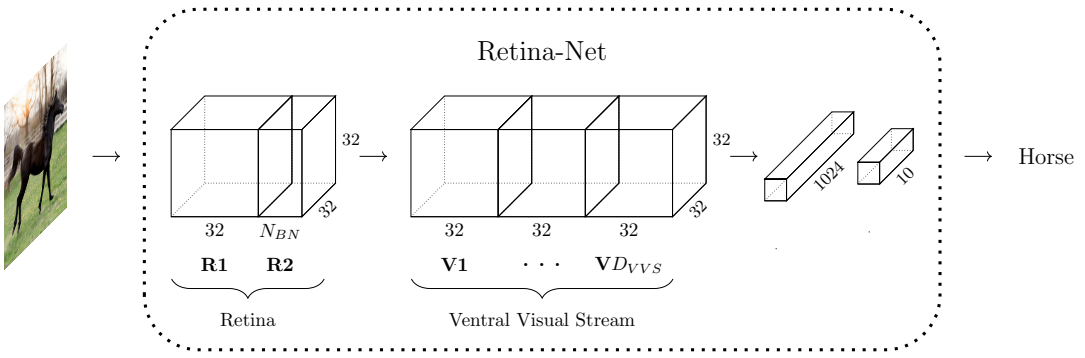
4.1.5 Hue Sensitivity

In addition to the hue tuning curve, we can consider the hue sensitivity of a network. Specifically, we look to replicate the experiments of Bedford and Wyszecki (1958), who showed that the change needed to elicit a just-noticeable difference in hue to a human observer is a complex function of wavelength. It is expected that such a sensitivity curve, though over hue rather than wavelength, will enable a more holistic view of colour tuning.

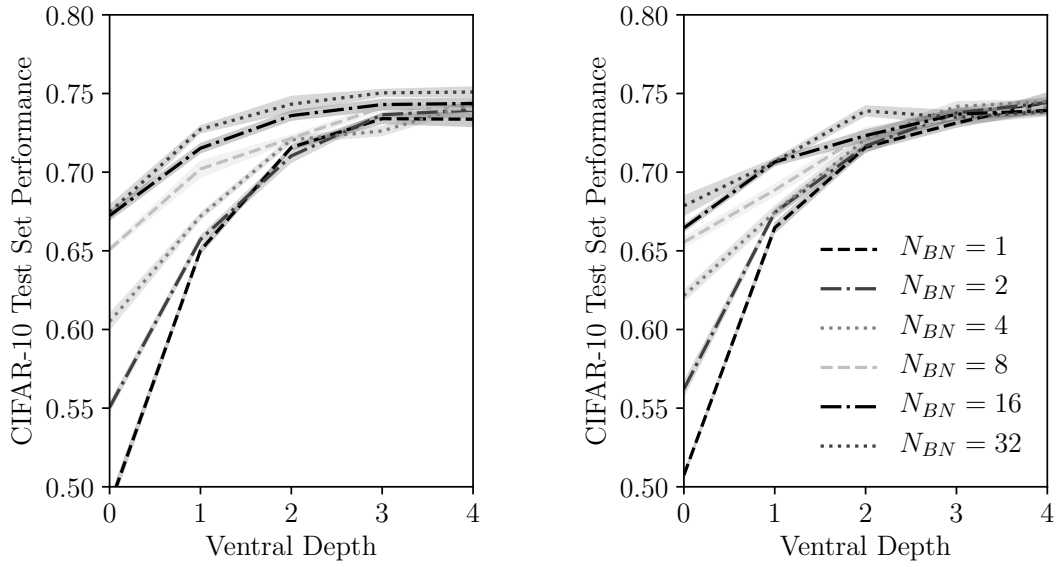
To perform a similar experiment to Bedford and Wyszecki (1958), note that the just-noticeable difference method is inversely related to the gradient of the perceived colour with respect to wavelength, which can be seen as a form of sensitivity. By virtue of automatic differentiation, it is trivial to obtain the gradient of the activation in a layer of our network with respect to the RGB input. Since the conversion from HSL to RGB (shown in Figure 4.4) is piece-wise differentiable, we can further obtain the approximate gradient of the activation with respect to hue. Note that we use the hidden layer activation of a network rather than a notion of 'perceived' colour, so it is unclear whether these results should reflect the biological data. Furthermore, in light of the above, one might expect that the predominant features of the sensitivity curve should derive from the relative responses of the RGB channels as a function of hue.

4.2 Results

We now present the results for our core experiments with Retina-Net models trained on colour CIFAR-10. We will later perform a control study and provide an in-depth discussion of the implications of these results, our aim in this section is merely to present the core findings of this work.



(A) Retina-Net schematic



(B) Greyscale (C) Colour

FIGURE 4.5: (a) Schematic of the Retina-Net model from (Lindsey et al., 2019). (b, c) CIFAR-10 test accuracy for the different combinations of retinal bottleneck and ventral depth explored in the experiments. Mean and standard error given over 10 trials.

4.2.1 Retina-Net

Since we are interested in understanding the casual link between architectural constraints and learned representation, we adopt the same deep convolutional model of the visual system as Lindsey et al. (2019), referred to as Retina-Net. This model, depicted in Figure 4.5a, consists of a model of the retina which feeds into a model of the visual cortex and Ventral Visual Stream (VVS). The retina model consists of a pair of convolutional layers with ReLU nonlinearities. The ventral network is a stack of convolutional layers (again with ReLUs) followed by a two layer MLP (with 1024 ReLU neurons in the hidden layer, and a 10-way softmax on the output layer). Note that Lindsey et al. (2019) additionally explore a model of the LGN, which can be considered as an extension of the retinal bottleneck (Ghodrati et al., 2017). We do not

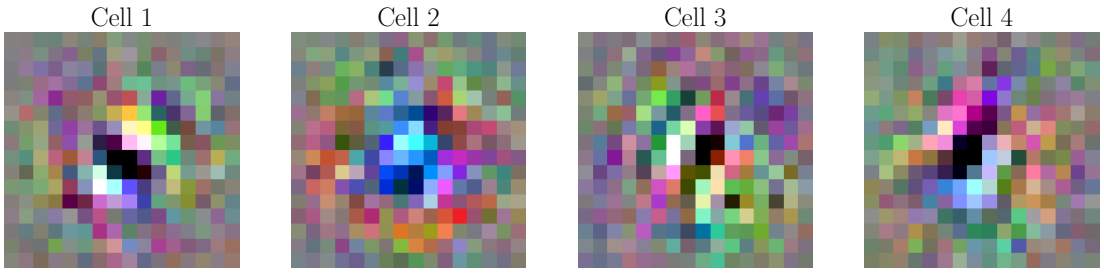
include such an exploration in this work as we are primarily focused on opponency and colour tuning in the bottleneck layer.

As with Lindsey et al.'s work, the networks are trained to perform classification on the CIFAR-10 data set (Krizhevsky, 2009), the only difference being that our model expects RGB inputs rather than greyscale. The choice of an object categorisation task is validated by previous studies which show there is a strong correlation between neural unit responses of CNNs trained on such a task and the neural activity observed in the primate visual stream (Yamins et al., 2014; Güçlü and van Gerven, 2015; Cadena et al., 2017). For further discussion of these results, please refer to Lindsey et al. (2019). Note that there may be many other learning tasks that are biologically valid in the sense that they yield similar functional properties. For example, self-supervised learning through deep information maximisation (Hjelm et al., 2018) and contrastive predictive coding (Hénaff et al., 2019) may present viable alternatives to the supervised object recognition used here.

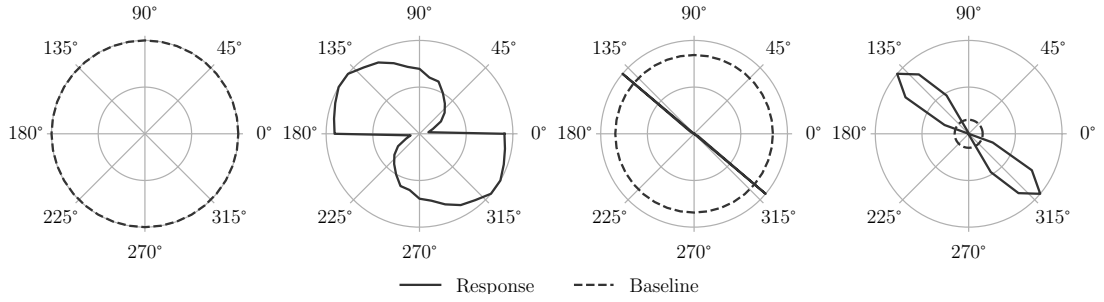
We train models across the same range of hyperparameters as Lindsey et al. (2019). Specifically these are: bottleneck width $N_{BN} \in \{1, 2, 4, 8, 16, 32\}$, and ventral depth $D_{VVS} \in \{0, 1, 2, 3, 4\}$. Again following Lindsey et al. (2019) we perform 10 repeats, with error bars denoting the standard deviation in result across all repeats. Networks were trained for 20 epochs with the RMSProp optimizer and a learning rate of $1e-4$ with initial weights sampled via the Xavier method (Glorot and Bengio, 2010). We note that in order to replicate the results from Lindsey et al. (2019), we required additional regularisation. Specifically, we use a weight decay of $1e-6$ and data augmentation (random translations of 10% of the image width/height, and random horizontal flipping). Figures 4.5b and 4.5c give the average terminal accuracy for models trained both on greyscale and colour images respectively. The greyscale accuracy curves match those given in Lindsey et al. (2019). The accuracy for networks trained on colour images is generally higher, particularly for networks with no ventral layers. We will later discuss additional training settings that are variants of the above.

4.2.2 Characterising Single Cells

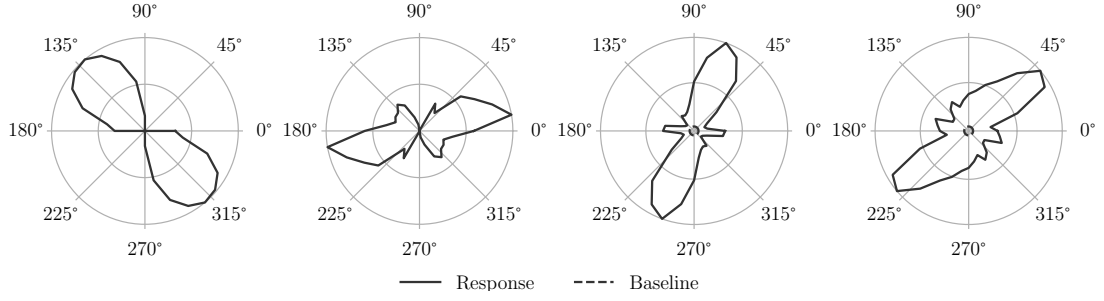
To begin, we illustrate our framework for characterising single cells. Figure 4.6 shows the first order receptive field approximations, orientation tuning curves, and colour tuning curves for four cells in the bottleneck layer of a network with $N_{BN} = 4$ and $D_{VVS} = 2$. Following Lindsey et al. (2019), the receptive field approximation is the gradient (obtained through back-propagation) of the output of a single convolutional filter in a single spatial position (that is, a single convolutional 'neuron') with respect to a blank input with a constant value of 0.01. This small positive amount is required to ensure that each of the cells is in the linear region of the ReLU activation function (that is, the gradient is non-zero). The gradient image is then normalised and scaled so



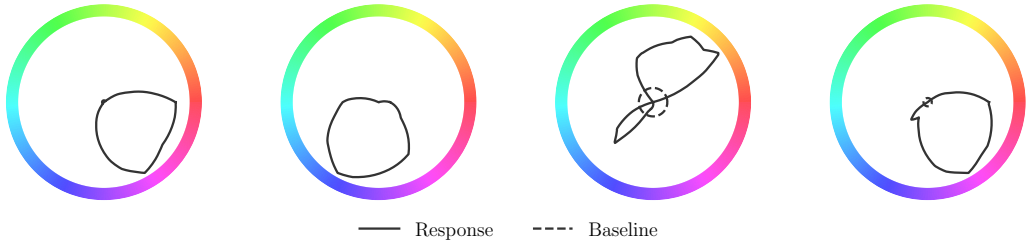
(A) First order receptive field approximation



(B) Orientation tuning curve for frequency and phase with minimal response



(C) Orientation tuning curve for frequency and phase with maximal response



(D) Colour tuning curve

FIGURE 4.6: Characterisation of the 4 cells in the second retinal layer of a network with $N_{BN} = 4$ and $D_{VVS} = 2$. (a) The receptive field approximation obtained from the gradient of the cell with respect to a blank image. (b) & (c) Orientation tuning curves for the frequency and phase combination that yielded the smallest and largest response respectively. (d) Colour tuning curve over the hue wheel. Cells 1, 3 and 4 are double opponent, cell 2 is non-opponent.

that it can be interpreted visually. Visually, cells 1, 3, and 4 appear to be greyscale edge filters, whereas cell 2 is red / blue or magenta / cyan centre-surround. However, the limitation of this analysis is the noise in the approximation. For example, one could argue that cell 1 is centre-surround with a dark centre and a magenta surround. Assessments given for any of the cells will be similarly contentious. Furthermore, this representation permits no understanding of inhibition. For example, cell 2 may be better described as tuned to blue hues in the interval $(180°, 270°)$, rather than centre-surround opponent.

To further characterise each cell, we employ our described approach. To characterise spatial opponency, in Figures 4.6b and 4.6c we provide orientation tuning curves for the frequency and phase which elicit the weakest and strongest responses respectively. If the cell responds above the baseline in one tuning curve and below in the other, or if either curve crosses the baseline, then the cell is spatially opponent. We can therefore say that, by our definition, cells 1, 3 and 4 are spatially opponent. In contrast, cell 2 is merely spatially non-opponent, always responding above the baseline for any choice of rotation, frequency, and phase. In addition to classifying opponency, we can identify the orientation tuning of each cell by further study of the curves in Figure 4.6c. Figure 4.6d gives the colour tuning curves for each cell. As hue is the only parameter to consider, classification here is simpler; the cell is hue opponent if the tuning curve crosses the baseline. Given this definition, we can say that cells 1, 3 and 4 are hue opponent, although the extent of inhibition is different in each case. Furthermore, for every cell we can identify the range of hues to which it is tuned.

Following interpretation of the tuning curves we can now state that cells 1, 3 and 4 are double opponent and that cell 2 is non-opponent both spatially and with regard to hue. Furthermore, for each cell we can state the orientation and hue to which it is tuned. For example, cell 2 is broadly excited by blue stimuli but with a distinct peak at a hue of around 240°. Cell 2 is spatially tuned to lines oriented in the interval $(0°, 45°)$. Although it is true that this approach gives us a deeper understanding of each cell, the real value is in the fact that each of the above steps can trivially be automated over the whole cell population. We therefore transition away from studying single cells, and instead consider the distributions of different cell types for the remainder of the chapter.

4.2.3 Characterising Cell Populations

For each result in this section, we automate cell classification following our described method and present the distribution of each cell type as a function of retinal bottleneck width and ventral depth. This allows us to understand the effect that these two architectural variables have on the kinds of cells that are learned and where they are found in the network. Note, however, that cells in deeper layers are expected to

have a highly non-linear response and thus may have receptive field properties that are quite different to the opponent cells observed in shallower layers. As such, observations regarding these deeper layers ('Ventral 2' in particular) should be considered only in the context of our approach and may not generally apply to the broader understanding of opponency.

Spatial opponency Figure 4.7 gives the distribution of spatially opponent, spatially non-opponent, and spatially unresponsive cells as a function of bottleneck width for a range of ventral depths. For a small bottleneck, the vast majority of cells in the second retinal layer are spatially opponent. Conversely, cells in the first ventral layer are predominantly spatially non-opponent. For deeper networks with less constrained bottlenecks the distributions are approximately equal in each of the layers. Almost all cells respond to some configuration of the grating stimulus, with only a small fraction of the population being spatially unresponsive. These findings are consistent with the observations that unresponsiveness has not been observed in the neuroscience literature and that the majority of cells in primate V1 are orientation tuned (Livingstone and Hubel, 1984). Regarding ventral depth, the results show a consistent reduction in spatial opponency in the last convolutional layer ('Retina 2' when depth is 0, 'Ventral 1' when depth is 1 etc.). There is a corresponding spike in spatial opponency in the penultimate convolutional layer ('Retina 2' when depth is 1, 'Ventral 1' when depth is 2 etc.). The average number of opponent cells in each layer does not differ greatly.

Colour opponency Curves showing how the distributions of the colour opponent classes change for the second retinal and first two ventral layers as the bottleneck is increased, for a range of ventral depths, are given in Figure 4.8. As the bottleneck decreases, the second retina layer exhibits a strong increase in hue opponency, nearing 100% for a bottleneck of one. Conversely, cells in the first ventral layer show a decrease in hue opponency over the same region. For all but the tightest bottlenecks, up to half of the cells are hue non-opponent. Hue non-opponent cells show almost the exact opposite pattern to hue opponent cells. The implication of this result is that networks with strong hue opponent representations in the bottleneck layer exhibit an increase in hue non-opponent cells in 'Ventral 1'. Since this spike in opponency returns in 'Ventral 2', we speculate that 'Ventral 1' merely preserves the opponent code from 'Retina 2' for downstream processing, and learns a set of filters that are tuned but non-opponent. This is inconsistent with the evidence that spatially tuned cells in primate V1 are also colour opponent (Lennie et al., 1990; Johnson et al., 2001). However, it should be stressed that our model of the primary visual cortex and ventral stream is highly simplified. In particular, we do not explicitly model the LGN or subsequent projections to different layers of V1 and greater similarity may well be observed in such a case. Similarly to the results for spatial opponency, there is a consistent reduction / spike in

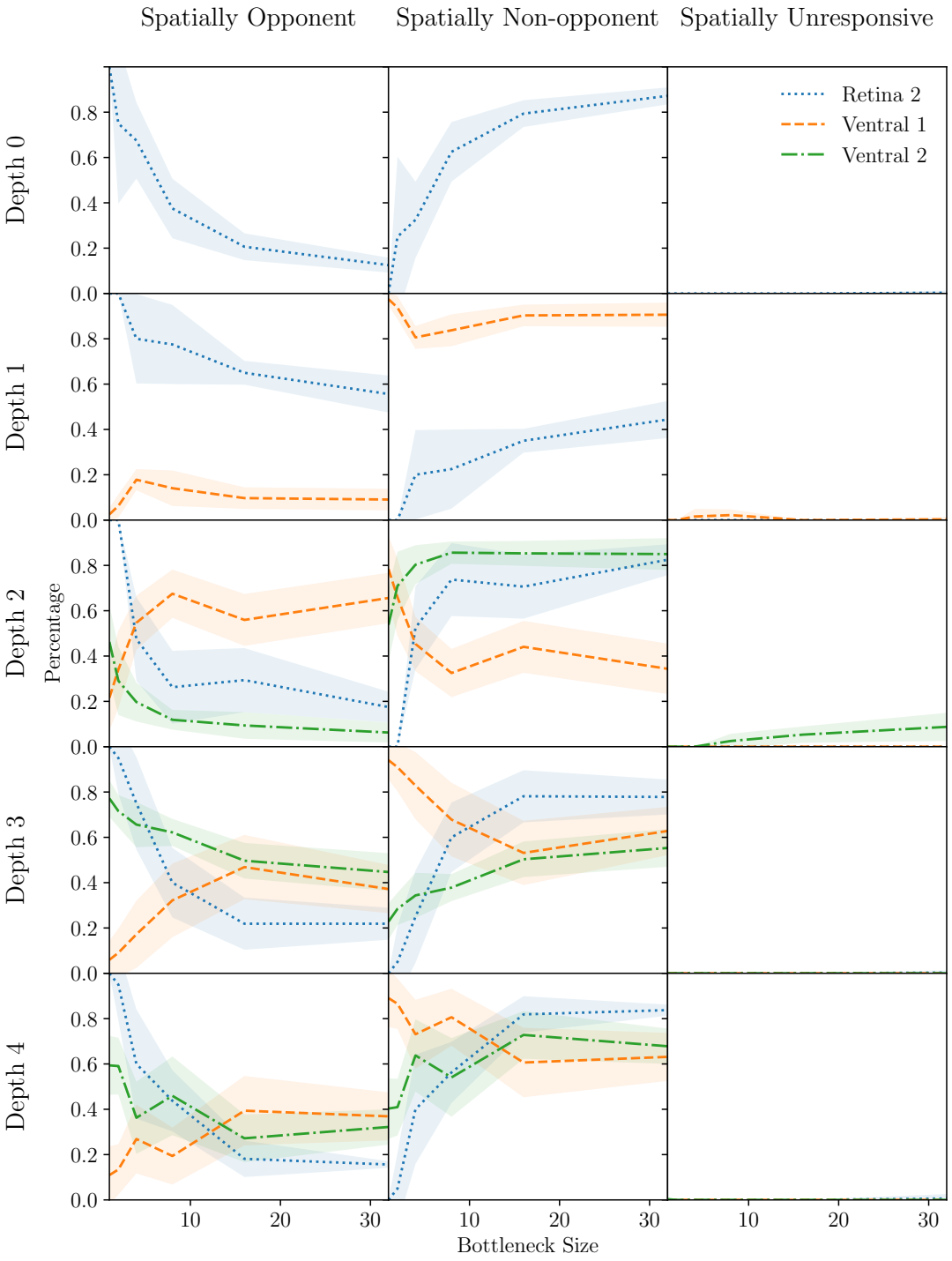


FIGURE 4.7: Distribution of spatially opponent, non-opponent, and unresponsive cells in different layers of our model as a function of bottleneck width, for a range of ventral depths. Functional organisation emerges for networks with tight bottlenecks. The last convolutional layer ('Retina 2' when depth is 0, 'Ventral 1' when depth is 1 etc.) exhibits a reduction in spatial opponency. The penultimate convolutional layer ('Retina 2' when depth is 1, 'Ventral 1' when depth is 2 etc.) exhibits an increase.

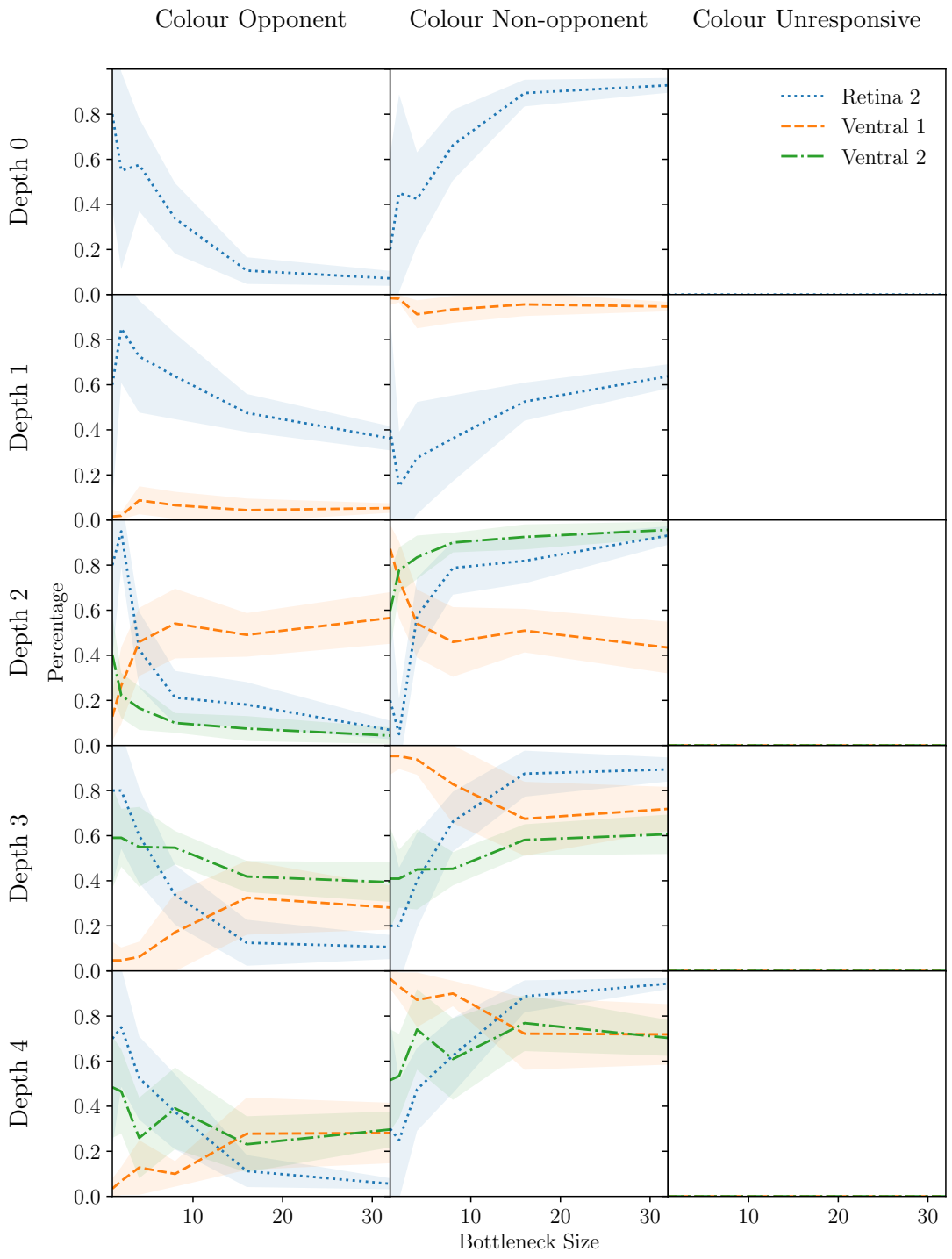


FIGURE 4.8: Distribution of colour opponent, non-opponent, and unresponsive cells in different layers of our model as a function of bottleneck width, for a range of ventral depths. Functional organisation again emerges for networks with tight bottlenecks. Furthermore, the last and penultimate convolutional layers exhibit a reduction and increase in colour opponency respectively. The echoes the spatial findings from Figure 4.7

hue opponency in the last and penultimate convolutional layers respectively. Averaged over bottleneck width, the number of hue opponent cells is generally lower than the number of spatially opponent cells.

Double opponency Figure 4.9 shows the distribution of double opponent cells as a function of bottleneck size and ventral depth, giving a similar picture to the spatial and hue opponency plots. The results suggest that the majority of hue opponent cells are also spatially opponent. This finding is in alignment with the observation that most hue opponent cells in the macaque V1 are also orientation selective (Johnson et al., 2008).

Types of opponency The plots in Figure 4.10 show the distribution over the hue wheel of the most excitatory and most inhibitory colours for cells in our models before and after training. The key observation here is that maximal excitation and inhibition before training is naturally aligned to the hues that correspond with RGB values of 255 or 0. This is a quirk of the convolutional architecture. Since at initialisation the function of the network is smooth, if the cell is excited by a particular channel, it will be most excited when that channel is maximised and vice-versa. The effect of training, regarding both excitation and inhibition, is to reduce the proportion of cells that are tuned to red, yellow, cyan, and blue, and increase the proportion of cells that are tuned to green and magenta. In addition some cells in the bottleneck layer ('Retina 2') become most excited by orange / red and cyan / blue. Unlike the random networks, this changes as a function of depth, tending to broaden the range of excitatory and inhibitory hues. Note that this corresponds to the network learning a complex, non-linear, colour system. Cells in deeper layers are not only excited by particular channels but by the specific hue of the input.

One could speculate that the distribution in Figure 4.10 indicates that the type of opponency that is learned corresponds well with the cone opponency observed in primates. However, as discussed, Figure 4.10 does not permit an understanding of the discrete types of opponency that are learned. Furthermore, the figure does not differentiate between the different model architectures. In Figure 4.11, we additionally plot the distribution of excitatory colours for all cells in the bottleneck layer, given that they are most inhibited by red, green, magenta, cyan, yellow, and blue respectively. These are plotted for 'Shallow' ($D_{VVS} \in \{0, 1\}$) and 'Deep' ($D_{VVS} \in \{3, 4\}$) networks with 'Narrow' ($N_{BN} \in \{1, 2, 4\}$) and 'Wide' ($N_{BN} \in \{8, 16, 32\}$) bottlenecks.

We can now observe that the primary opponent axis in our networks is green / magenta, with cells that are inhibited by red or magenta and excited by green being unique to the 'Wide' / 'Shallow' networks. In addition, we can say that the majority of hue opponent cells (that is, cells in the 'Narrow' networks) are channel opponent. In

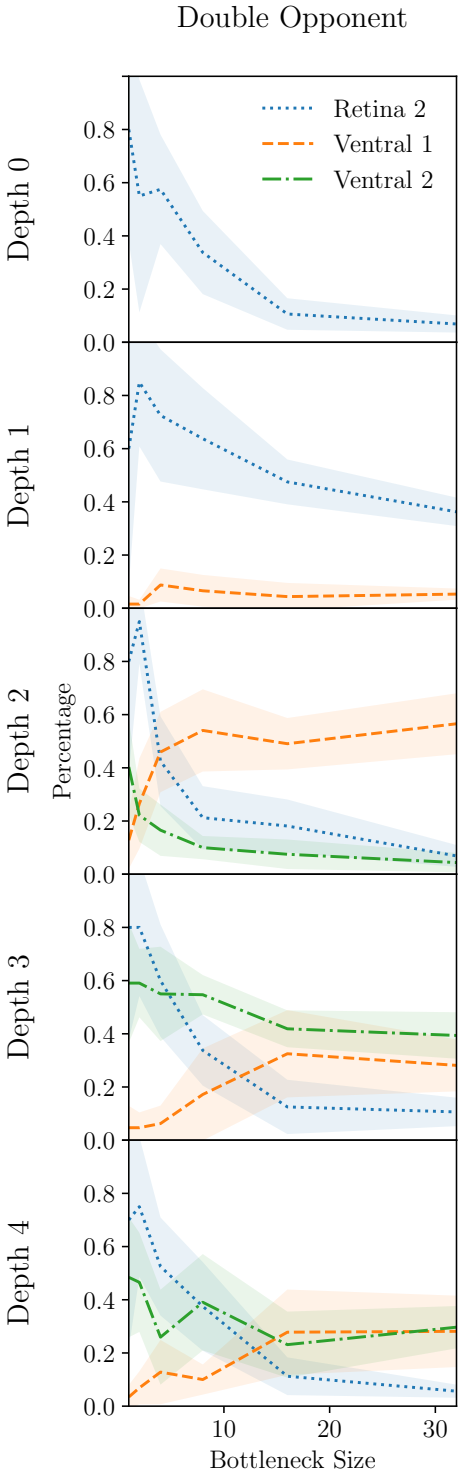


FIGURE 4.9: Distribution of double opponent cells in different layers of our model as a function of bottleneck width and ventral depth. Most spatially opponent cells are also colour opponent and so these distributions bare strong similarity to those in Figures 4.7 and 4.8

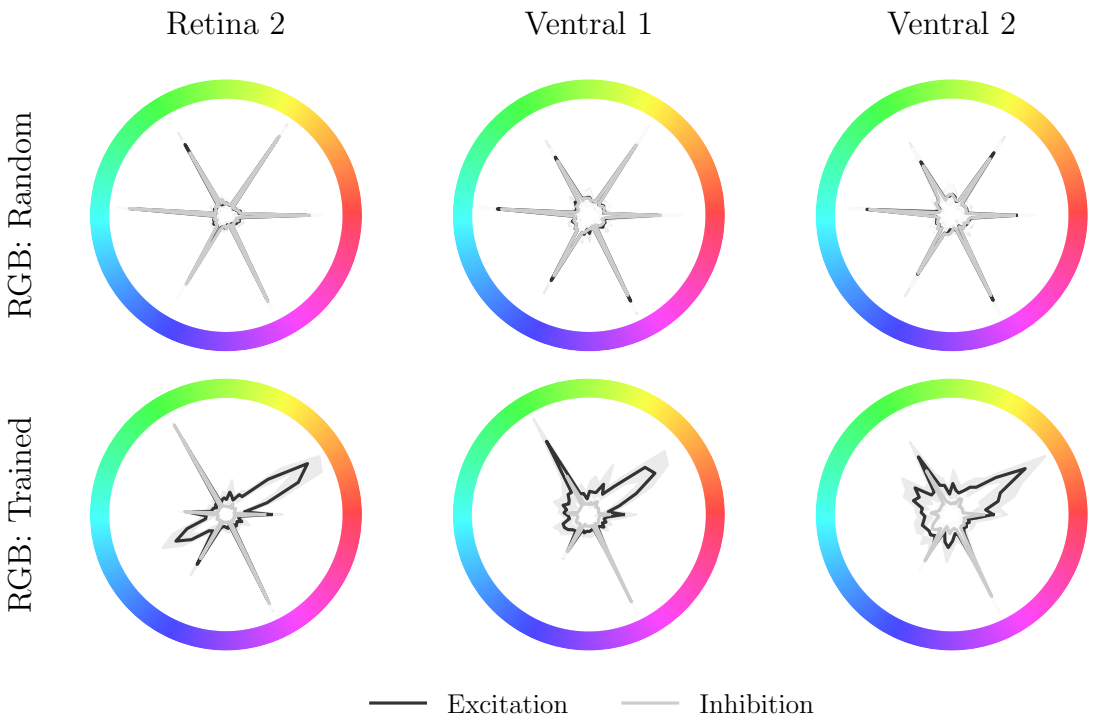


FIGURE 4.10: Distribution of excitatory and inhibitory hues for cells in different layers of networks with random weights, and networks trained on RGB images. Maximal excitation and inhibition before training is naturally aligned to the hues that correspond with RGB values of 255 or 0. Trained networks show a preference for green and magenta. Some cells are highly non-linear, maximally excited by orange / red and cyan / blue.

the 'Wide' networks, we find cells which are broadly excited by orange / red and cyan / blue. These cells persist in the first ventral layer, and are not typically present in 'Narrow' networks. This suggests that the 'Wide' networks are responsible for the peaks in Figure 4.10. We find the presence of cells which are excited by blue and inhibited by yellow, red, and green more prominently in the 'Deep' networks, with particular prevalence in the 'Narrow + Deep' networks. In general, the range of excitatory and inhibitory hues is greater in the 'Wide' networks, suggesting increased prevalence of complex, non-linear cells. This mirrors the finding from Lindsey et al. (2019) that cells in this setting tend to have a non-linear receptive field. Note that we have found that cells in the ventral layer (not included in the figure) are excited and inhibited by a much wider range of hues, particularly in the 'Narrow' networks. This suggests that the bottleneck induces an efficient colour code that enables cells in later layers to become attuned to highly specific hues. Recall that we observe an increase in the proportion of colour tuned but non-opponent cells in 'Ventral 1' in models with tight bottlenecks, corroborating this assertion.

Hue Sensitivity Figure 4.12 gives the results for the hue sensitivity experiment. We again provide plots for 'Shallow' and 'Deep' networks with 'Narrow' and 'Wide'

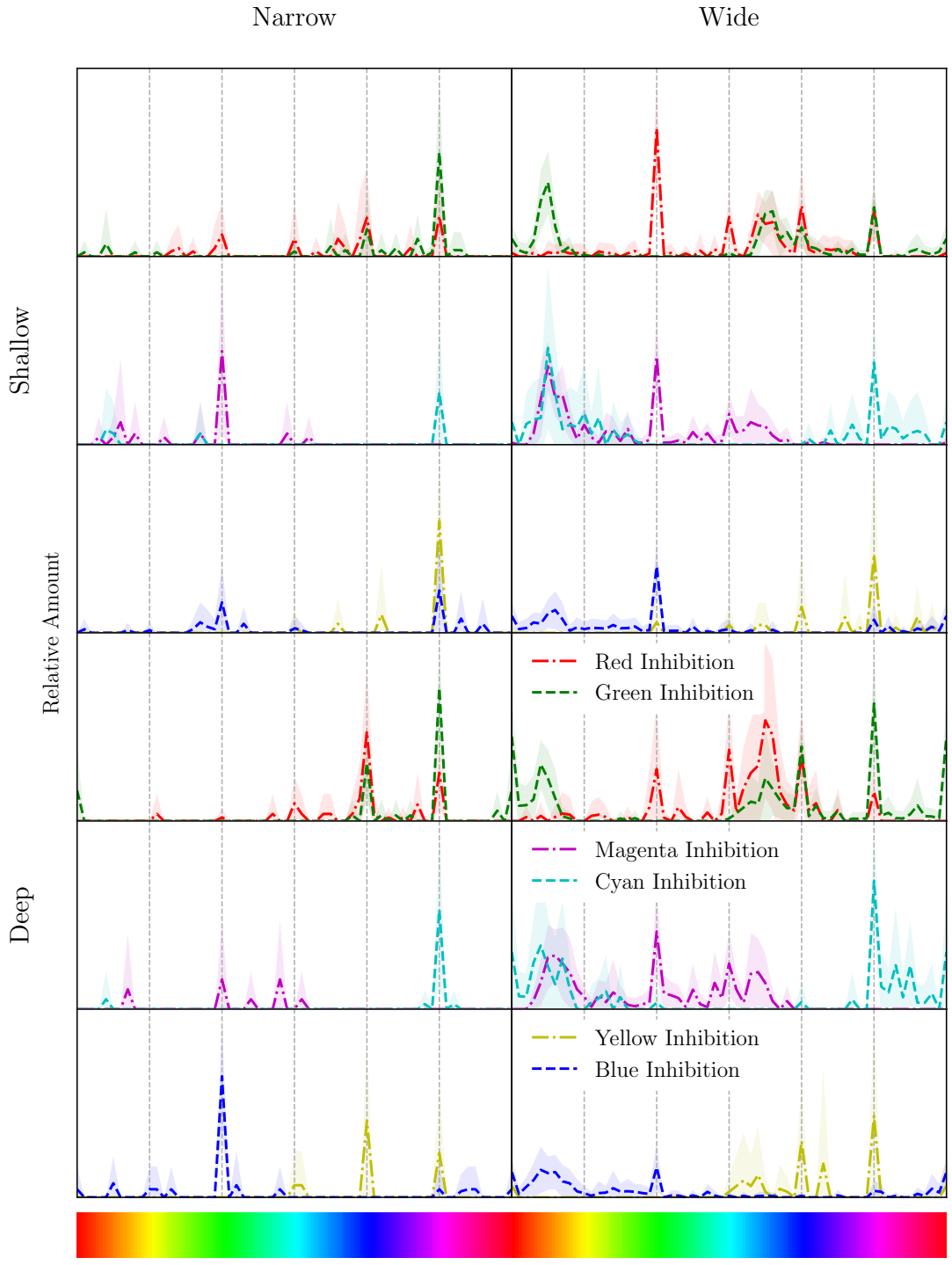


FIGURE 4.11: Conditional distribution of excitatory hues for cells that are most inhibited by red ($[315°, 45°)$), yellow ($[45°, 75°)$), green ($[75°, 165°)$), cyan ($[165°, 195°)$), blue ($[195°, 285°)$), and magenta ($[285°, 315°)$) for 'Shallow' ($D_{VVS} \in \{0, 1\}$) and 'Deep' ($D_{VVS} \in \{3, 4\}$) networks with 'Narrow' ($N_{BN} \in \{1, 2, 4\}$) and 'Wide' ($N_{BN} \in \{8, 16, 32\}$) bottlenecks. 'Narrow' networks learn a simple colour system, with cells that are maximally excited / inhibited by extreme RGB values (dashed vertical lines). 'Deep' networks show an increase in cells that are most excited by blue.

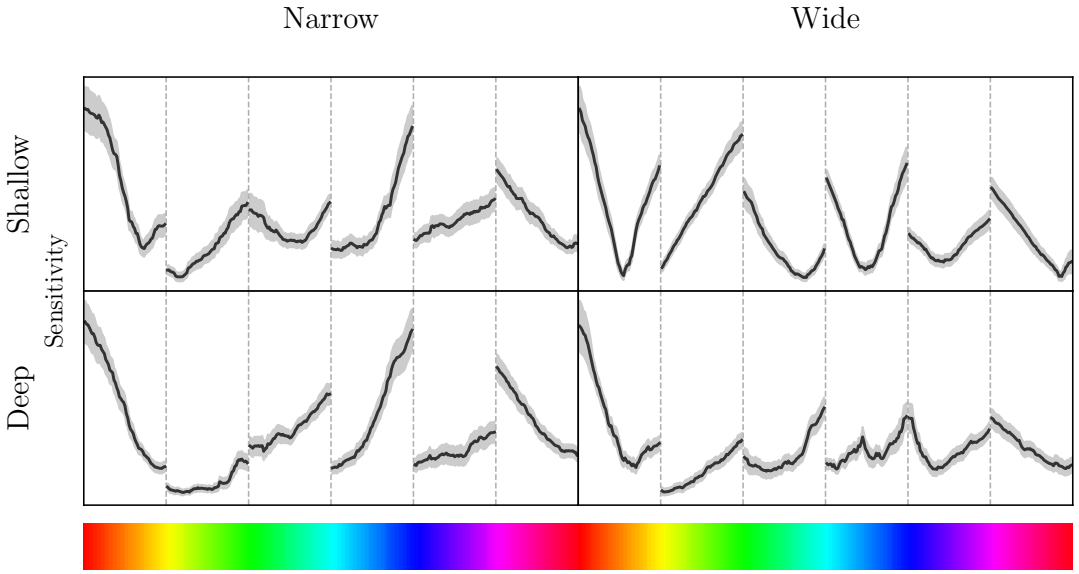


FIGURE 4.12: Mean gradient of the sum of the bottleneck layer response with respect to hue for 'Shallow' ($D_{VVS} \in \{0, 1\}$) and 'Deep' ($D_{VVS} \in \{3, 4\}$) networks with 'Narrow' ($N_{BN} \in \{1, 2, 4\}$) and 'Wide' ($N_{BN} \in \{8, 16, 32\}$) bottlenecks. The shaded region indicates the standard error across the trained models. Discontinuities derive from the conversion from HSL to RGB. Sensitivity is an approximately linear function of hue for 'Narrow' networks, and particularly in the 'Narrow + Deep' setting, again showing a simple colour code in the bottleneck layer. Conversely, 'Wide + Shallow' networks exhibit a highly non-linear sensitivity to hue.

bottlenecks so that these results can be understood in the context of the previous section. Since we are taking the gradient of the response, the sensitivity is undefined where there are discontinuities in the conversion from HSL to RGB (dotted vertical lines). The first point to note is that the straight lines in the sensitivity curves correspond to at most a quadratic response to hue. In contrast, non-linear sensitivity curves suggest a higher order hue response. In light of this, we can observe a general transition from highly non-linear hue response in the 'Wide + Shallow' networks to a more linear hue response in the 'Narrow + Deep' networks. This is in line with our findings in the previous section and again mirrors the findings from Lindsey et al. (2019). Regarding tuning to specific colours, Lehky and Sejnowski (1990) note that gradient of the tuning curve (such as the curves in Figure 4.6d) is maximal when the stimulus is to either side of the peak. As such, where there are peaks in the distribution of cells that are excited by a particular hue we should expect corresponding troughs in the sensitivity curve. As an example, note that the troughs in sensitivity to orange / red in the 'Wide + Shallow' (and to a lesser extent in the 'Narrow + Shallow' and 'Wide + Deep') networks matches the peaks in excitation observed in Figure 4.11. A similar observation can be made regarding the trough in sensitivity to cyan / blue in the 'Wide + Shallow' networks. The blue excitation that is a uniquely prominent feature in the 'Narrow' networks has also resulted in a corresponding dip in sensitivity. However, this has manifested as a sudden drop rather

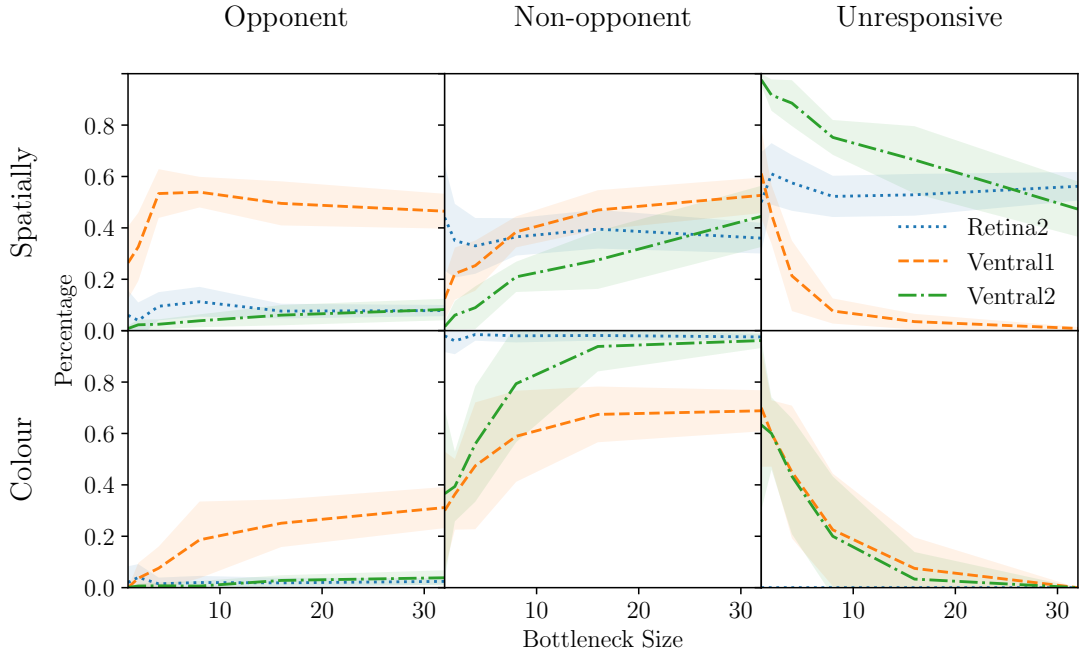


FIGURE 4.13: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of our models with Gaussian weights (mean and variance from filters of the same depth in a reference pre-trained model with $N_{BN} = 32$ and $D_{VVS} = 4$) as a function of bottleneck width. Some opponency is explained by simple statistics of the filters. Functional organisation emerges only as a result of training.

than a smooth transition since it lies at the discontinuous boundary between blue with a green component and blue with a red component. Ultimately these results demonstrate that low level analysis of the colour tuning distributions provides valid insights into the high level functional properties of the networks.

4.3 Control Experiments

In this section we perform a series of targeted experiments to assess how well our results extend to different settings. These experiments are intended to improve our understanding of the conditions under which the various forms of opponency emerge, supporting a comprehensive discussion.

4.3.1 Random weights

Although we have presented strong evidence that cells in trained networks exhibit spatial, colour, and double opponency, we have not yet demonstrated that this is learned. To determine if this opponency is learned, we require a demonstration that it is not present at initialisation (when the weights are random). We have therefore performed our experiments on randomly initialised models. We find that networks

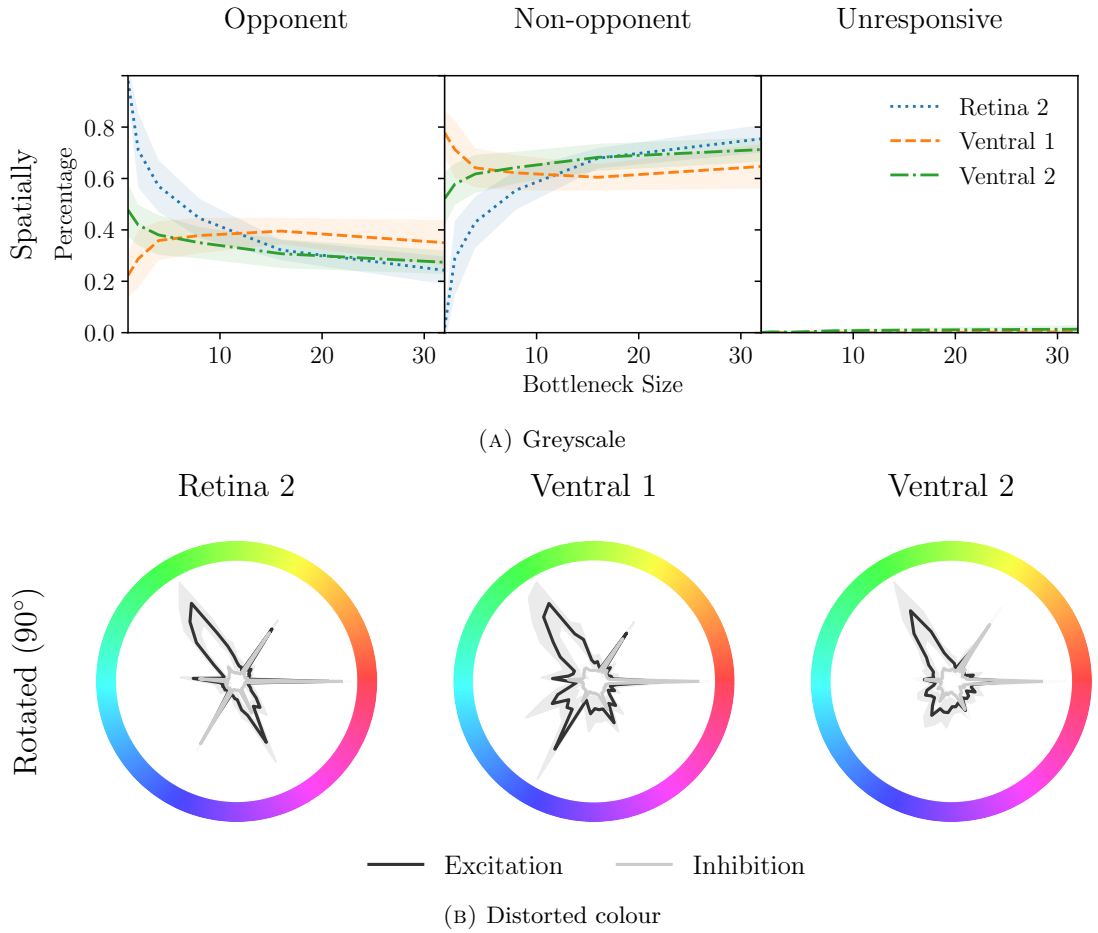
with random weights (that is, following the Xavier initialisation (Glorot and Bengio, 2010)) never exhibit spatial or hue opponency. Instead, most cells are non-opponent and their distribution over the layers does not change significantly with the bottleneck size. These results demonstrate that all of the opponency in our networks is learned. However, it could be the case that opponency derives from simple statistics of the convolutional filters. In order to understand this further we experimented with networks whose filter weights are Gaussian with the same mean and variance as the filters of the same depth in a reference pre-trained model. The results for this experiment are given in Figure 4.13. Although we do find some opponency in this case, we do not find the same structure. Of particular note are the unresponsive, not present in the trained networks. These findings reflect the fact that a degree of structure in the receptive fields is required in order for a cell to exhibit a consistent opponent or non-opponent response.

4.3.2 Greyscale

As previously mentioned, in addition to the colour models we trained a batch of models with greyscale inputs. The results in Figure 4.14a validate that spatially opponent cells still emerge and have a similar distribution throughout the layers as that of cells in models trained with RGB inputs. Furthermore, this validates our classification approach since, from Lindsey et al. (2019), it is known that the Retina-Net model learns centre-surround and oriented edge (both spatially opponent) filters with greyscale input.

4.3.3 Distorted colour

To further explore the idea that the opponency in our networks derives from the statistics of the data, we trained a batch of models on images with distorted colour. Specifically, we convert the images into HSV space and offset the hue channel by 90°, before converting back into RGB and forwarding to the network. Our interest here is not in whether opponency emerges, but in the effect this distortion has on it. Figure 4.14b shows the distribution of excitatory and inhibitory colours in networks trained with distorted inputs. Here, the most prevalent excitatory and inhibitory colours are aligned with the RGB extremes closest to a 90° rotation of the peaks in Figure 4.10. This is consistent with our observation that the vast majority of colour opponent neurons are channel opponent. In contrast, the additional excitation peak has been rotated by exactly 90° from orange / red to green. This demonstrates that the cells which are excited by specific hues emerge as a result of the statistics of the data, not of the input colour space.



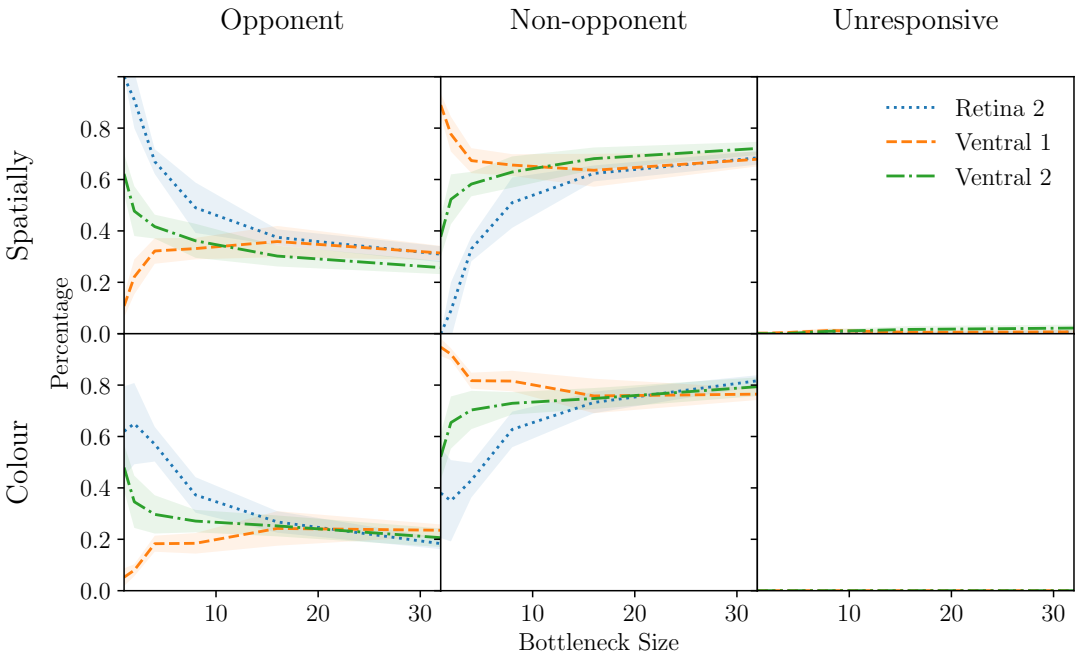
(A) Greyscale

(B) Distorted colour

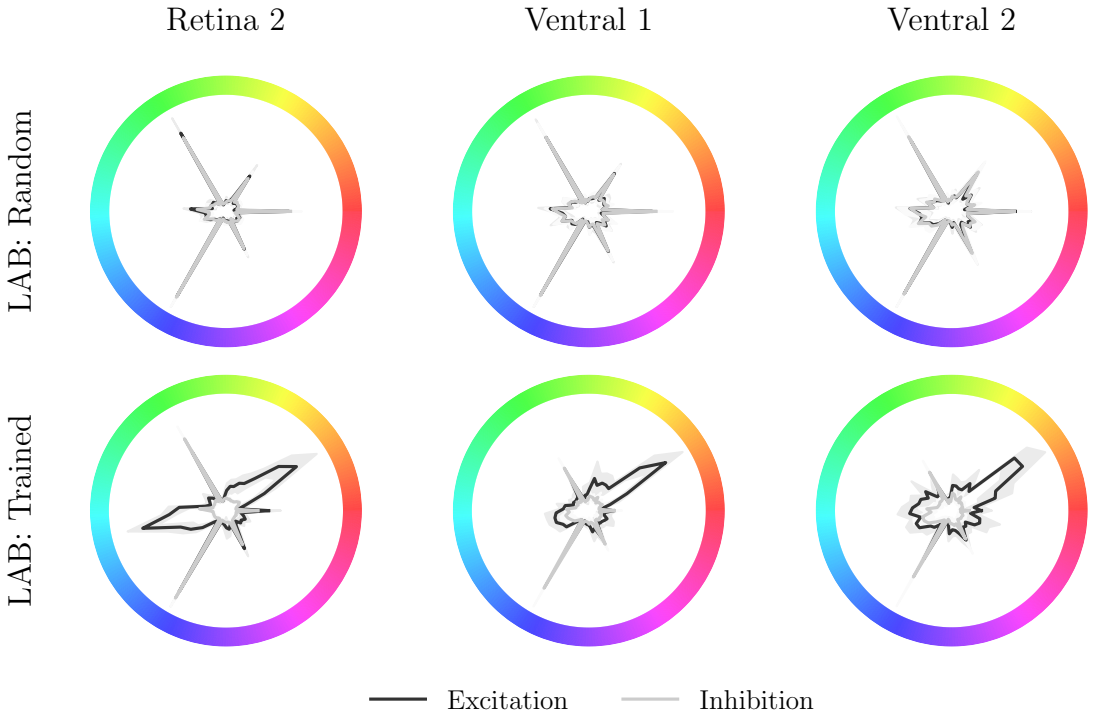
FIGURE 4.14: (a) Distribution of spatially opponent, non-opponent, and unresponsive cells in different layers of our model as a function of bottleneck width, for models trained with greyscale images showing that the known spatial opponency from Lindsey et al. (2019) is detected by our method. (b) Distribution of excitatory and inhibitory hues for cells in different layers of networks trained on images with distorted colour (hue rotation of 90°). The most prevalent excitatory and inhibitory colours are aligned with the RGB extremes closest to a 90° rotation of the peaks in Figure 4.10.

4.3.4 CIELAB space

In a similar vein to our experiments with distorted colour, we now perform experiments to validate whether opponency is still a feature in networks trained on images in the CIELAB colour space. The CIELAB colour space encodes colour in terms of lightness (L*), and two opponent axes: green / red (a*), and blue / yellow (b*). Each axis is non-linear such that uniform changes in CIELAB space correspond to uniform perceptual changes in colour. This will allow us to understand if functional organisation still emerges when receptive fields are naturally opponent, that is, structure would require the cells to learn to 'ignore' the inherent opponency of the a* and b* channels. Figure 4.15a shows the distribution of spatially and colour opponent cells in this setting. The distribution is nearly identical to that of networks trained on images in RGB space, with the same characteristic functional organisation. In Figure



(A) Opponency



(B) Excitatory and inhibitory colours

FIGURE 4.15: (a) Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on images in LAB space as a function of bottleneck width, showing that functional organisation is not unique to RGB. (b) Excitatory / inhibitory hues in LAB space for random and trained networks. Training increases prevalence of blue / green, and excitation by orange / red and cyan / blue.

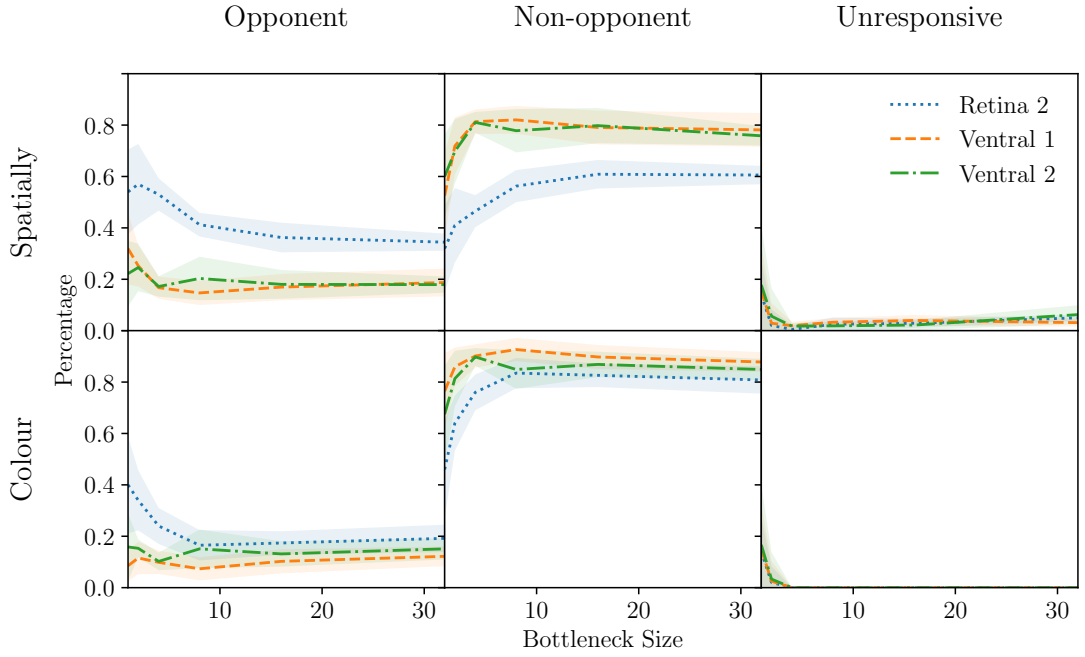


FIGURE 4.16: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on Street View House Numbers (SVHN) (Netzer et al., 2011) as a function of bottleneck width. Spatial opponency is present, with a similar distribution to the networks trained on CIFAR-10. Colour opponency is generally lower, increasing only slightly for networks with narrow bottlenecks.

4.15b, we plot the distribution of most excitatory and inhibitory colours in CIELAB space for random and trained networks. We again find that the distribution for random networks naturally aligns to hues which represent extreme values in the input colour space; CIELAB encodes colour on green / red and blue / yellow axes. Following training, we find that the majority of cells are most excited or inhibited by either green or blue. This bares some similarity to the RGB networks, which aligned to green and magenta following training. Again in accordance with the RGB networks, we find cells which are most excited by orange / red or cyan / blue, further showing that tuning to these particular colours is an artefact of the data rather than the colour space.

4.3.5 Street view house numbers

In addition to our experiments with CIFAR-10, we have trained a batch of networks on the Street View House Numbers (SVHN) data set (Netzer et al., 2011). This is a digit recognition (10 classes) problem with the same spatial resolution (32×32) as CIFAR-10. The distributions of the different cell types for these models are shown in Figure 4.16. The results show that spatial opponency is abundant in the second retina layer and increases for models with tight bottlenecks. In the first two ventral layers, the proportion of spatially opponent cells is generally lower ($\approx 20\%$) and does not change significantly with bottleneck size. Colour opponency is muted in comparison to

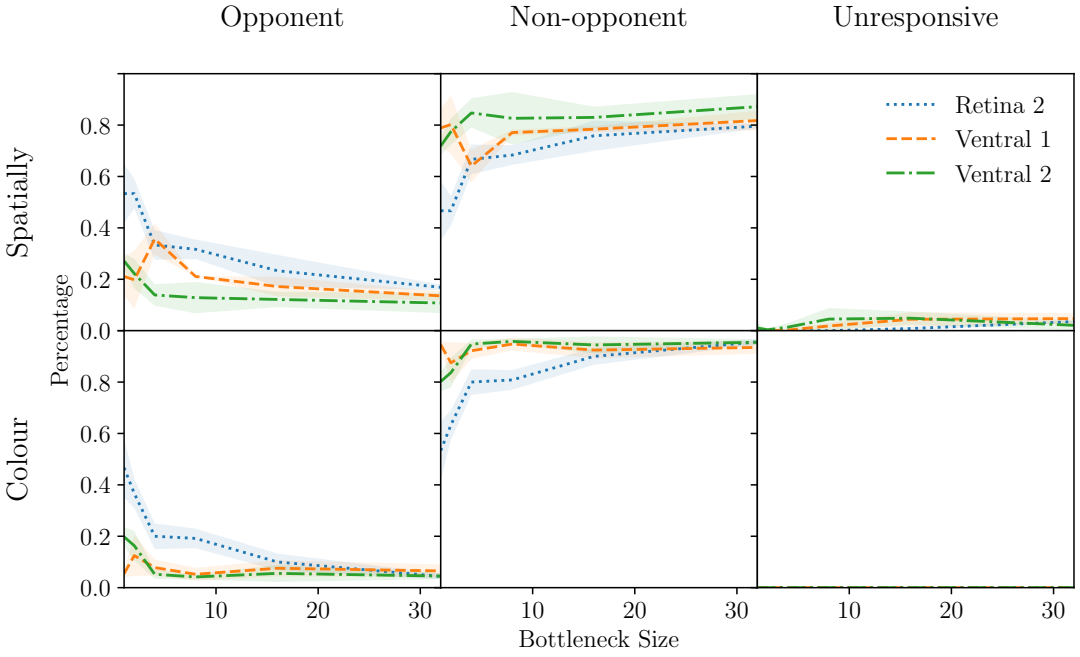


FIGURE 4.17: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on ImageNet (Russakovsky et al., 2015) as a function of bottleneck width, showing how our findings transfer to a higher resolution setting. There is an increase in opponency for narrow bottlenecks which decays rapidly. Emergent organisation is observed only partially in the networks with the tightest bottlenecks.

the CIFAR-10 experiments and increases in the second retina layer only slightly for tight bottlenecks. This is unsurprising since colour is not expected to be an important feature in the house number recognition problem. The largest networks in this setting achieved over 90% accuracy.

4.3.6 ImageNet

Our experiments thus far have focused on low resolution (32×32) images. It is now important to understand whether our findings generalise to a higher resolution setting. To that end, we have trained networks on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) data set (Russakovsky et al., 2015) at a resolution of 128×128. Due to hardware constraints, we perform only 3 repeats across the full range of ventral depths and bottleneck widths. We adapt the Retina-Net model slightly, adding average pooling with a window of size 4 before the first fully connected layer. Figure 4.17 gives the distributions of the different spatial and colour cell types respectively for models trained on ImageNet. As with CIFAR-10, the results show an increase in the proportion of spatially and colour opponent cells in the bottleneck layer of networks with a tight bottleneck. Unlike the CIFAR-10 results, this opponency decays rapidly and the emergent organisation is observed only partially in the networks

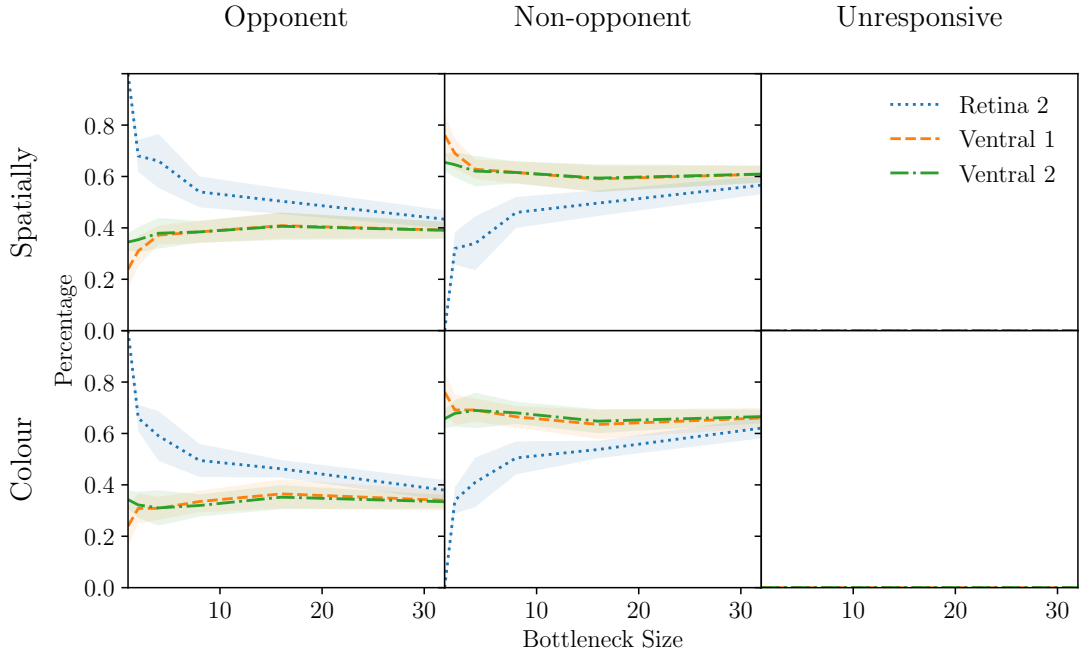


FIGURE 4.18: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on the Intel scene classification challenge data set (Intel, 2018) as a function of bottleneck width. With fewer classes (6 in this case), the number of opponent cells is much higher. The distribution of opponent cells in 'Retina 2' bares strong similarity with the results from CIFAR-10. This does not extend to the ventral layers, which have near-identical cell distributions.

with the tightest bottlenecks. The percentage of opponent cells was generally lower than in networks trained on CIFAR-10. This could be related to the fact that the Retina-Net model does not effectively 'fit' to ImageNet; the trained models achieved an accuracy between 5% and 20%. Although the number of double opponent cells in this setting (not shown in the figure) is much lower, the vast majority of spatially opponent cells are also colour opponent. We do not find a spike in opponency in the penultimate convolutional layer in general in the ImageNet trained models.

4.3.7 Intel scene classification

In addition to our experiments with ImageNet, we have trained a batch of models on the Intel scene classification challenge (Intel, 2018) data set. This is a natural scene classification problem with 6 classes and the same spatial resolution as ImageNet. As such, these models allow us to explore opponent cell types in a high resolution setting where the model obtains stronger performance (up to 80% for the largest models). The results in Figure 4.18 show that models trained in this setting exhibit a much higher percentage of spatially and colour opponent cells than in models trained on ImageNet, particularly in the second retina layer. We again find that the percentage of opponent cells in the first two ventral layers is equal and broadly independent of bottleneck size,

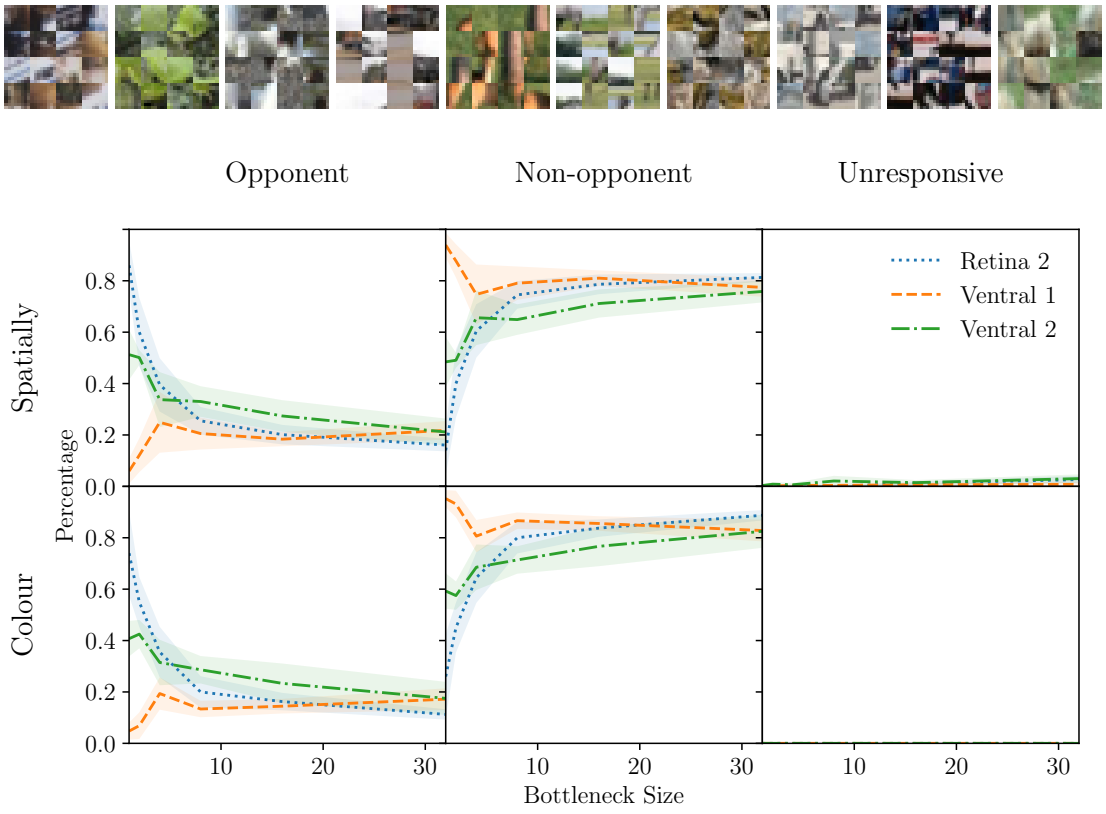


FIGURE 4.19: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on mosaic images as a function of bottleneck width with example mosaic images. These results show that when the spatial structure of the input is removed some spatial opponency, particularly in 'Retina 2', is removed also. Colour opponency is similarly affected, suggesting a complex dependence between spatial and colour processing.

not showing the emergent structure observed in CIFAR-10 except in the extreme case of $N_{BN} = 1$.

4.3.8 Classifying mosaics

We also performed experiments to determine whether there are conditions under which certain types of opponency can be removed. To attempt to ablate spatial oppponency, we trained a batch of models to classify mosaic images. These are images that have been separated into smaller squares which have then been shuffled (see examples in Figure 4.19 for reference). Figure 4.19 gives the distribution of spatially and colour opponent cells in these networks. The figure shows that some of the spatial opponency is removed in this setting. Notably, 'Retina 2' exhibits a moderately lower proportion of spatial opponency, such that it is now in line with 'Ventral 2'. This could be due to the fact that the impact of the mosaic images depends heavily on the size of the receptive field. We further note that colour opponency is affected to the same extent as

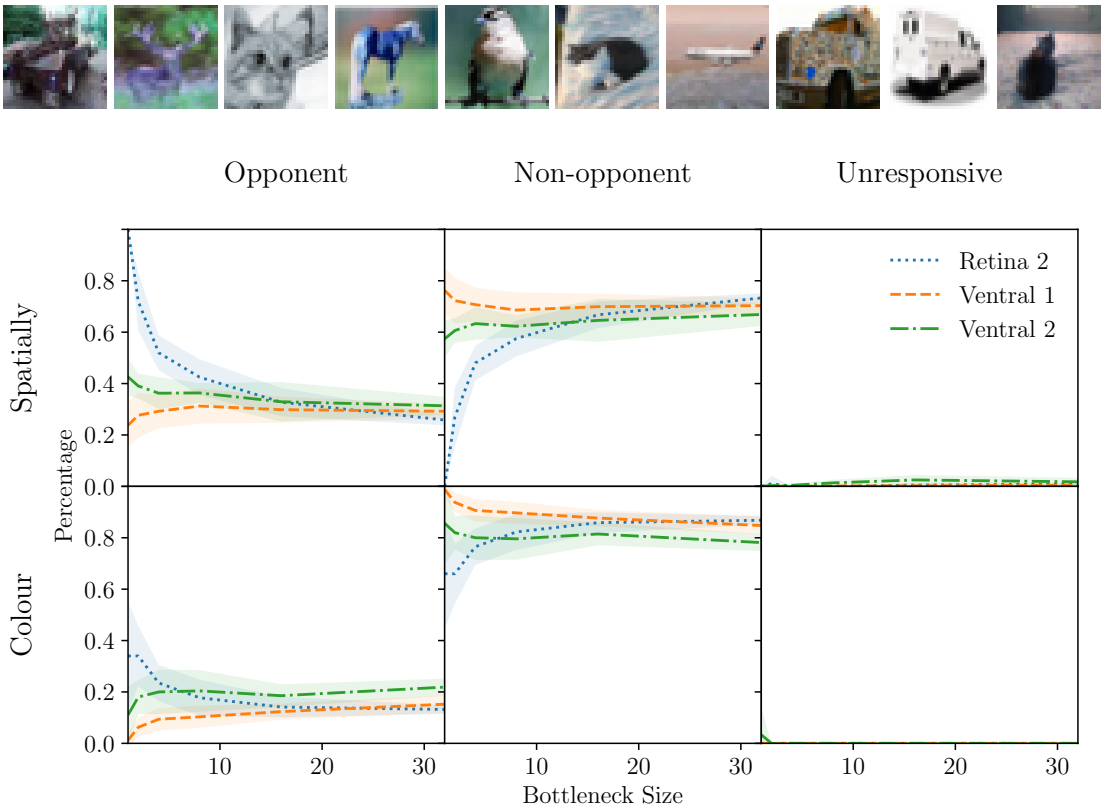


FIGURE 4.20: Distribution of spatially and colour opponent, non-opponent, and unresponsive cells in different layers of models trained on images with shuffled colour channels as a function of bottleneck width with example shuffled images. When consistent colour information is removed, most colour opponency is also removed. Spatial opponency remains.

spatial. This suggests that the efficacy of colour opponent cells is in some way reduced in the mosaic setting.

4.3.9 Shuffled colour channels

To attempt to ablate colour opponency, we remove colour information by randomly shuffling the channels of inputs to the network (see examples in Figure 4.20 for reference). The resultant distribution plots in Figure 4.20 show that this alteration removes the vast majority of colour opponent cells, whilst spatial opponency remains. Since the information present in shuffled images is the same, this experiment demonstrates that colour opponency arises out of a need to consistently infer the colours in the inputs. We speculate that this aids in classification since each class will be associated with a set of features that vary both spatially and in hue. By shuffling the channels we remove the ability to repeatedly associate a particular input tuple with a particular class. This view is supported by the fact that the models in this setting generally reached a lower accuracy than the models trained on standard colour images and sometimes failed to match the models trained on greyscale images.

4.4 Summary

Equipped with the results of our experiments, we now summarise the conclusions which can be drawn regarding spatial and colour processing in convolutional neural networks.

Our primary finding is that the addition of a bottleneck in the Retina-Net model induces functional organisation when trained on colour (RGB) CIFAR-10. We have further shown that this finding generalises to networks trained on images in the CIELAB colour space. There is some evidence that this result differs when the networks are trained on other data sets; although the key finding, that structure emerges only with the tightest bottlenecks, remains. In the case of ImageNet, more experimentation with a model capable of fitting to the data would be required to understand this fully. Regarding network depth, our experiments have uncovered an increase in the number of opponent cells in the penultimate convolutional layer of the network and a corresponding decrease in the last convolutional layer. Our experiments with random networks demonstrate that all of the discussed opponency is learned and that most opponency is not a result of simple statistics of the weights.

In addition to these high level observations, we have shown that an analysis based on approaches from neuroscience can yield a rich understanding of the function performed by a trained network. For example, we have shown that the deep Retina-Net model with a tight bottleneck learns a set of double opponent filters in the bottleneck layer, followed by a set of spatially and colour tuned but non-opponent filters in the first ventral layer, with opponency returning in the second ventral layer. Cells which are maximally excited by blue are a unique feature of these networks not present when the bottleneck is relaxed. Furthermore, these networks tend to learn linear, channel opponent, neurons rather than neurons which are opponent to specific hues. We speculate that this is due to the increased need to learn an efficient colour code in the tight bottleneck case.

The key implication of our core findings is that the model architecture can be the source of an inductive bias towards the number of opponent cells. While this finding alone may be of interest, whether it is of any practical significance depends on whether opponency is desirable. By virtue of the fact that opponent cells represent a more efficient encoding of the input, one might speculate that an increase in opponency could lead to increased generalisation performance. This view is mildly supported by the plot in Figure 4.5c, where the networks with $N_{BN} = 2$ and $D_{VVS} = 4$ obtained the highest accuracy.

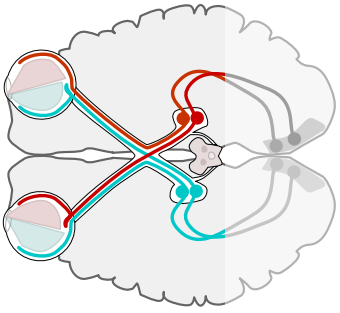
We have also demonstrated a number of similarities between the learned representations of our networks and representations observed in nature. The large amount of double opponent cells we find in the retina layer of networks with tight bottlenecks is consistent with what is known about cells in the retina and LGN (Hubel

and Wiesel, 2004). There are some consistencies and some inconsistencies between the ventral layers of the model and what is known about spatial and colour processing in the visual cortex. However, as discussed, it is not clear that the ventral convolutional architecture is a good analogue of the structure of the visual cortex and so such comparisons should be treated with scepticism. Our finding that the type of opponency learned is aligned with extreme values in the input colour space accords with the physiological finding that opponency in early stages of the visual pathway is aligned with cone responses (Shevell and Martin, 2017).

The consequence of these demonstrations is not to suggest that convolutional neurons and biological neurons are similar. Instead, we have shown that similarity in the data space, architecture, and problem setting can give rise to similarity in the emergent functional properties. In addition to the above, we have demonstrated some settings in which opponency is either hindered or removed entirely. This kind of controlled experiment may enable the exploration of hypotheses relating to the neuroscience of vision. Specifically, through construction of a data set which mimics an environment, or an architecture which mimics an anatomy, one might seek a better explanation of the differences in visual processing between species. This potential is hinted at by our experiments with SVHN, which show that networks trained on the digit recognition task have fewer colour opponent cells.

Chapter 5

Opponency and Robustness



Convolutional Neural Networks (CNNs) with a bottleneck designed to mimic the anatomical constraint imposed by the optic nerve have recently been shown to learn representations which more closely align with those found in primate visual systems. Specifically, Lindsey et al. (2019) showed that a reduction in the number of channels or convolutional neurons in the second layer of a CNN gives rise to centre-surround receptive fields in the first two layers followed by orientation selectivity in subsequent layers. Lindsey et al.'s experiments were based on a relatively simple model of the visual system (aka 'Retina-Net') constructed with a model of the retina (the first two layers up to the bottleneck), a model of the ventral stream (stacked convolutional layers) and a pair of fully connected layers to produce a classification. These models were trained on a greyscale version of the CIFAR-10 data set (Krizhevsky, 2009). In Chapter 4, we showed that the same network architecture (albeit with colour input) learns a simple opponent colour code in the bottleneck layer followed by a reduction in colour opponency in later layers. We further demonstrated that the general trend that the bottleneck induces opponency was true across a range of data sets from CIFAR-10 to ImageNet (Russakovsky et al., 2015).

Modern neural networks for solving large-scale classification problems like ImageNet look very different to the aforementioned 'Retina-Net' and its colour variant, and do not incorporate any form of bottleneck near the input layer. At the same time, other recent research has demonstrated that there may be benefits to better modelling of the visual system in state-of-the-art models. For example, Geirhos et al. (2019) show that CNNs trained on ImageNet are biased towards texture information in contrast to Human observers who make classifications predominantly according to shape information. They subsequently show that promoting a shape bias in trained models can improve both accuracy and robustness. Dapello et al. (2020) show that simulating

the function performed by the primary visual cortex over the input increases robustness to adversarial attacks in which images which have been imperceptibly (to a human) adjusted such that the classification given by the trained network is no longer correct. Their model for simulating V1 (V1Block) is based on the classical observations (for example by Hubel and Wiesel (2004)) that the receptive fields of cells in V1 are orientation selective (modelled as Gabor filters) and that these cells can be described as either simple or complex (modelled with different acitvation functions). Parameters for the Gabor filters in the V1Block are sampled according to known distributions from the neuroscience literature rather than being learned. The model also assumes that the input to the V1Block is just the normalised pixel values (subtract 0.5 and divide by 0.5 for each RGB colour channel) and eschews any notion of processing performed by the retina itself, or by the Lateral Geniculate Nucleus (LGN).

Since convolutional neurons already reflect many of the functional properties of cells found in the early visual system, it is prudent to ascertain whether more simple modifications, such as the bottleneck explored in Chapter 4, can be used to the same effect. This chapter starts to ask what happens if we place a bottleneck at the beginning of a performant classification network trained on the ImageNet data set. We first show that the bottleneck induces similar emergence of relevant cell types, but with some important differences. In particular, we find that very tight bottlenecks give rise to cells that are perhaps better described as luminance opponent rather than colour opponent. In addition, we find that the stark differences between receptive fields of cells in networks with different bottlenecks, observed by Lindsey et al. (2019), are more nuanced with centre-surround receptive fields seeming to emerge for all bottlenecks. We further show that the introduction of the bottleneck introduces a small but consistent shape bias. Finally, we consider the adversarial robustness of our models, showing that the bottleneck slightly improves robustness to natural adversarial examples but actually reduces robustness measured as a worst-case accuracy following multiple attacks.

5.1 Methods

This section outlines our approach to constructing an anatomically constrained ResNet model. In order to add a bottleneck to an existing model, we propose pre-pending the 'Retina' portion of the model from Lindsey et al. (2019). This consists of two convolutional layers with 32 channels and N_{BN} channels respectively, where N_{BN} is the bottleneck width. Unlike the model of Lindsey et al. (2019) which uses a kernel size of 9, we choose the kernel size to ensure a large enough receptive field whilst preserving consistency with the first layer of the target model where possible (e.g. 7 for a ResNet). Padding is chosen such that the resolution of the output is the same as the

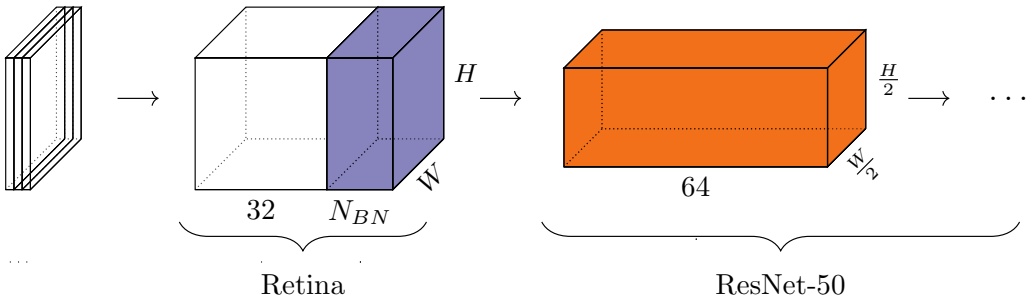


FIGURE 5.1: Anatomically constrained ResNet-50 model diagram. The only modification to the ResNet-50 architecture is that it expects an input with N_{BN} channels. We trained such models on the ImageNet data set for $N_{BN} \in \{1, 2, 4, 8, 16, 32\}$.

resolution of the input. In this way the bottleneck acts as a kind of parameterised pre-processing of the input.

We trained constrained variants of a ResNet-50 (depicted in Figure 5.1) for the range of bottleneck widths used by Lindsey et al. (2019) ($N_{BN} \in \{1, 2, 4, 8, 16, 32\}$) on the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovsky et al., 2015) data set. We trained three repeats of each model variant, giving 18 models in total. Models were trained on the ImageNet 2012 training set on nodes equipped with four NVidia Quadro RTX 8000 GPUs. The following setup was used:

Batch size:	1024	
Optimiser:	SGD with the following:	
	initial lr:	0.1
	momentum:	0.9
	weight decay:	1e-4
	schedule:	lr drops by factor of 10 at 30 and 60 epochs
	epochs:	90
Data Augmentation:	Random resized cropping to 224x224	
	random horizontal flipping	
	normalisation using standard ImageNet mean and s.d.	

5.2 Early Visual Representations in Anatomically Constrained ResNets

In this section, we study early visual representations in our constrained ResNet model using the methodologies introduced in Chapters 4 and 3 and other methods from the literature.

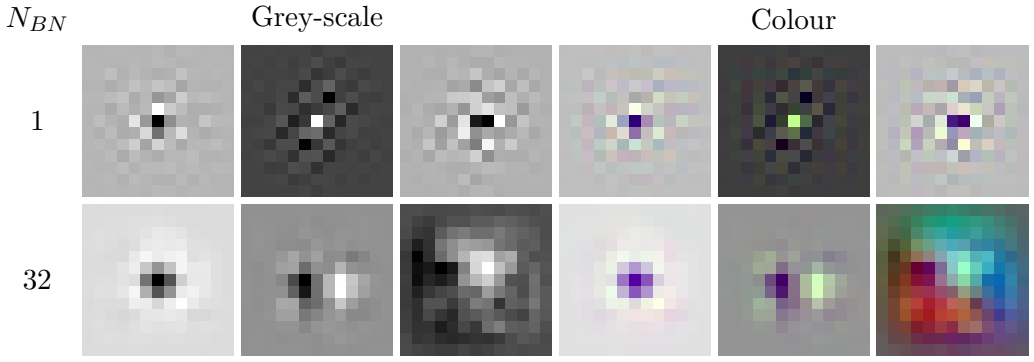


FIGURE 5.2: Grey-scale and colour receptive field visualisations for networks with $N_{BN} = 1$ and $N_{BN} = 32$.

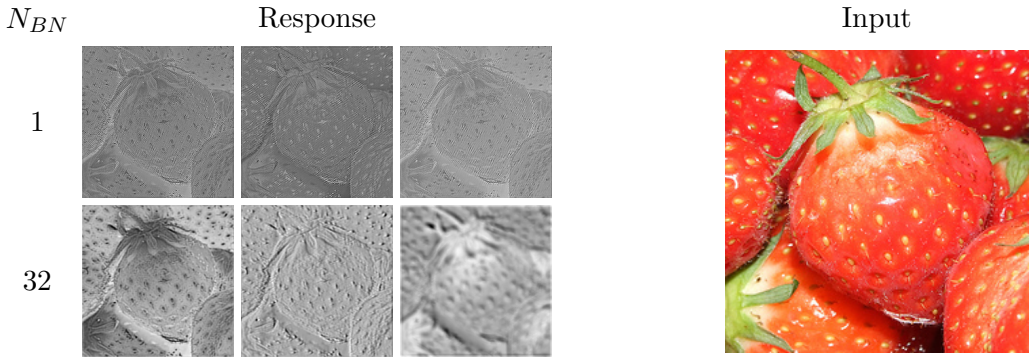


FIGURE 5.3: Cell responses to an example input for the same cells used in Figure 5.2 from networks with $N_{BN} = 1$ and $N_{BN} = 32$.

5.2.1 Receptive Field Visualisations

We begin our analysis with a subjective assessment of the receptive fields of cells in our trained networks. Figure 5.2 shows the receptive field visualisations in grey-scale and colour for cells in the Retina-2 layer of networks with $N_{BN} = 1$ and $N_{BN} = 32$. These are computed by taking the gradient of the response of the centre neuron to a blank stimulus (constant value of zero) and then applying min / max normalisation to obtain an image. The results show that the cells in networks with $N_{BN} = 1$ are somewhat centre-surround, but with quite a small extent. In contrast, the cells in networks with $N_{BN} = 32$ have much smoother receptive fields that are sensitive to changes over a much larger region of the input. Although it is true that we find exclusively centre-surround cells in the networks with $N_{BN} = 1$, we also find some centre-surround cells in the networks with $N_{BN} = 32$.

To try to better understand the functions performed by these cells, we plot the response of each cell to an example input in Figure 5.3. The results show that when $N_{BN} = 1$, the Retina-2 cells preserve most of the edge information in the image (with perhaps a mild edge enhancement effect) and respond in accordance with the brightness of the input (albeit inverted for two of the cells). In contrast, the Retina-2

cells in networks with $N_{BN} = 32$ distort the input heavily, performing specific colour opponent edge detection functions.

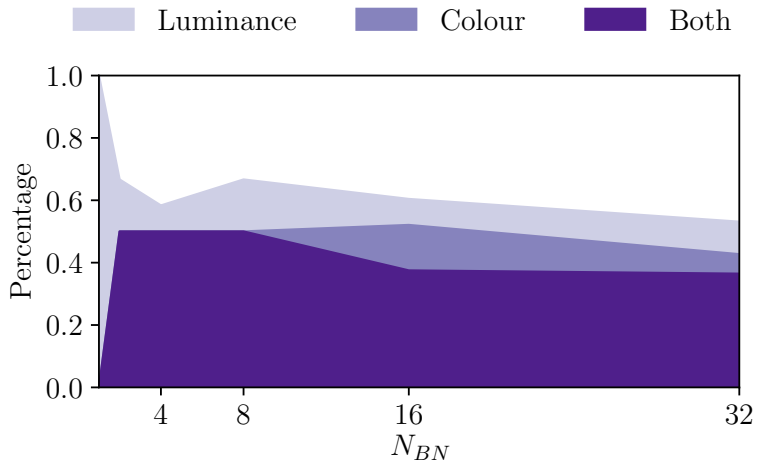
5.2.2 Opponency

Next, we compute the distribution of opponent cell types, following the approach from Chapter 4, in the second retina layer (Retina-2) and first ResNet layer (ResNet-1). This approach consists of presenting a set of stimuli which vary in hue to the cell in order to obtain a hue response curve. The cell is colour opponent if the response curve crosses the baseline response of the cell to a zero stimulus. Interestingly, our results show that networks with $N_{BN} = 1$ have no colour opponent cells in either of the layers we studied. This contrasts with the finding from Chapter 4 that almost all cells in the retina layers of more constrained networks were colour opponent. We presume that the range of colour stimuli in ImageNet is sufficiently broad that when restricted to one channel, better performance can be obtained with a luminance opponent encoding than a colour opponent one. To test this, we additionally present a set of stimuli which vary in luminance (i.e. full field grey-scale stimuli with values ranging from zero to one) and perform the same process of comparing to the baseline (response of the cell to constant stimulus of 0.5 for this experiment) to infer luminance opponency. Note that any cell whose response is linear in the sense that extremes of input value correspond to extremes of response will be classified as luminance opponent. It is therefore appropriate to say that any cells which are not luminance opponent are to some degree non-linear.

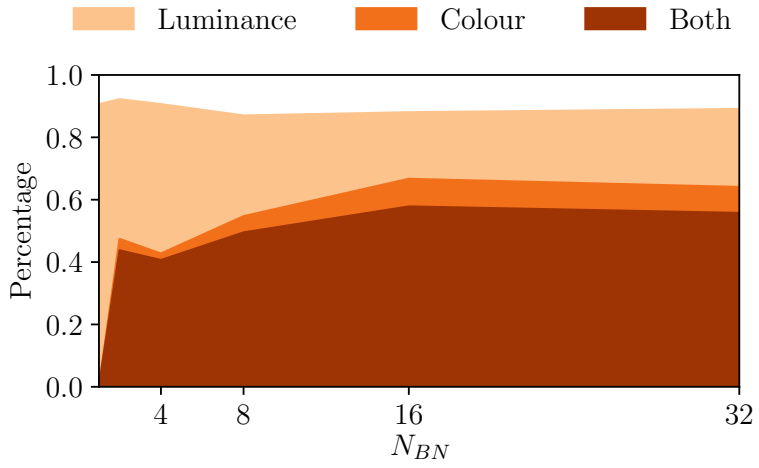
Figures 5.4a and 5.4b give the results for this experiment in the Retina-2 and ResNet-1 layers respectively. The results show that the vast majority of colour opponent cells are also luminance opponent, particularly in the more constrained networks. In the Retina-2 layer, there is an increase in cells which are luminance opponent, but not colour opponent, for networks with $N_{BN} < 4$. This in turn suggests an increase in the linearity of cell response for these networks, which accords with the findings of Lindsey et al. (2019). Almost all cells in the first ResNet layer are luminance opponent. In both layers, approximately half of the cells are also colour opponent in networks with all but the tightest bottlenecks, where the percentage of colour opponent cells drops to zero. Ultimately, the pattern in the colour opponent cell distribution shown in Chapter 4 is replaced with a similar pattern in luminance opponent cell distribution in these networks. This is accords with our findings in Section 5.2.1.

5.2.3 Super-stimuli

In the neuroscience literature, the distinction between luminance opponency and colour opponency is often unclear. For example, it has been argued that many of the



(A) Retina-2 opponent cell distribution



(B) ResNet-1 opponent cell distribution

FIGURE 5.4: Colour and luminance opponency distributions in the Retina-2 and ResNet-1 layers of anatomically constrained ResNet-50 models.

cells which preferred achromatic stimuli in Monkey V1, described by Hubel and Wiesel (1968), may be better described as colour opponent but with imbalanced cone response such that they are also responsive to luminance (Johnson et al., 2001; Lennie et al., 1990; Schluppeck and Engel, 2002; Shapley and Hawken, 2011). We therefore must ascertain whether the luminance opponent cells are still able to convey valuable information regarding the colour of the input.

To do this, we employ the approach from Chapter 3 to obtain super stimuli which the networks classify with high confidence as a range of classes which are characterised by a particular colour (a trait group). Figure 5.5 shows these stimuli for the 'fruits' trait group for models with $N_{BN} = 1$ and $N_{BN} = 32$. From the figure we can see that the networks with a single ($N_{BN} = 1$) luminance opponent channel in the bottleneck layer still make use of colour to inform their decisions, learning the characteristic red of a strawberry, yellow of a lemon, green of an apple, and so on. This is possible since these

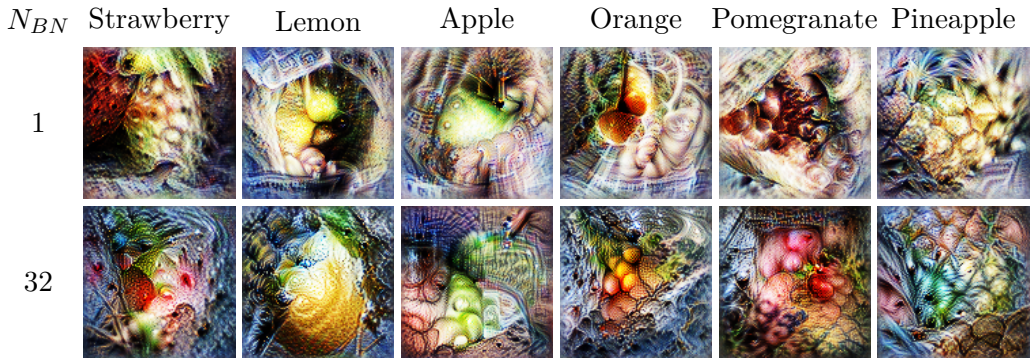


FIGURE 5.5: Super-stimuli for various classes of fruit for networks with $N_{BN} = 1$ and $N_{BN} = 32$ obtained by gradient ascent in Fourier space following Olah et al. (2017). Even when the networks are restricted to a single channel ($N_{BN} = 1$), they still learn a strong representation of colour.

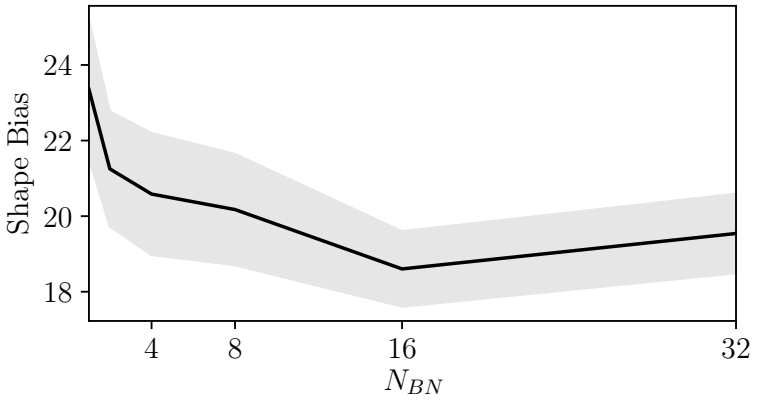


FIGURE 5.6: The shape bias (following Geirhos et al. (2019)) across bottleneck widths.

luminance opponent cells may still have a broad non-opponent hue response that uniquely encodes colour over at least part of the hue wheel.

5.2.4 Shape Bias

It is clear from Figure 5.5 that neither of the models is particularly biased towards shape in that the super stimuli have appropriate textures and colours but not shapes. To quantify this, we finally measure the shape bias of the models. Following Geirhos et al. (2019), we evaluate the models on the 'StylizedImageNet' data set [1], which contains images with a texture-shape cue conflict, requiring networks to recognise objects based on shape rather than texture. Figure 5.6 shows that networks with tighter bottlenecks exhibit a small increase in shape bias. The model with a bottleneck width of 1 trained on ImageNet achieves the highest shape bias, outrunning a standard (i.e. without our modification) ResNet-50 model. Note, however, that the shape bias of

[1] Available from https://github.com/rgeirhos/Stylized-ImageNet

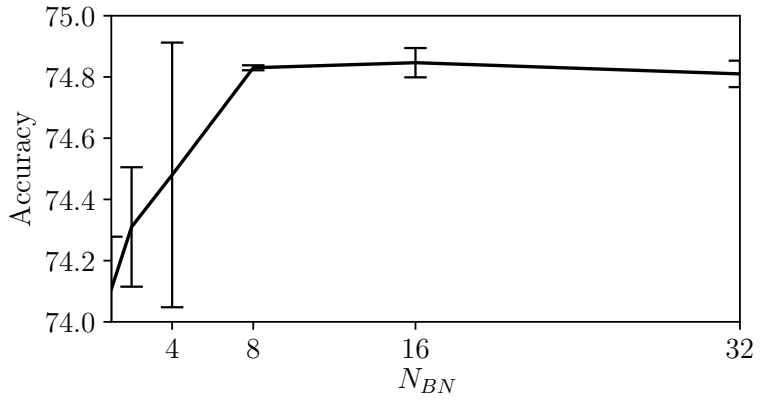


FIGURE 5.7: Top-1 accuracy on the ILSVRC validation set across different bottleneck widths.

all of the models is far lower than the models trained with the approach from Geirhos et al. (2019).

5.3 What Effect Does a Bottleneck Have on Performance?

In this section we explore whether the changes in function induced by the bottleneck correspond to any tangible improvement in performance. The most obvious first step is to consider the effect of the bottleneck on the accuracy of the model. Figure 5.7 shows that networks with bottleneck width of 4 or greater all achieve validation set performance that is within the margin of error (one standard deviation) of the others and that networks with tighter bottlenecks show a small but clear drop in performance. Importantly, all of the networks perform within 2% of the pre-trained ResNet-50 without our modification provided in torchvision (76.15%), which has minor differences in the training procedure that might explain this. This contrasts with the networks considered by Dapello et al. (2020) which show a larger reduction of around 4.5% reaching an accuracy of 71.7% (although again note that there are differences in the training procedure). Note that in addition to small differences in training procedure, our bottleneck layers do not make use of modern conventions such as BatchNorm which are known to provide important performance benefits, particularly in deeper networks.

5.3.1 Robustness to natural adversarial examples

We now consider robustness to the ImageNet-A data set (Hendrycks et al., 2019), which contains 7500 naturally occurring images which are consistently mis-classified by pre-trained models. The images correspond to 200 ImageNet classes in which normal models (ResNet50, VGG16, etc) have a top-1 accuracy of more than 90% using the

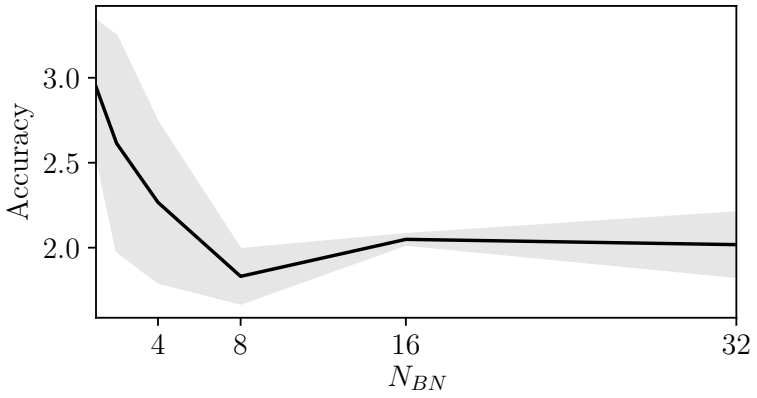


FIGURE 5.8: Robustness to natural adversarial examples from the ImageNet-A data set (Hendrycks et al., 2019) as a function of bottleneck width.

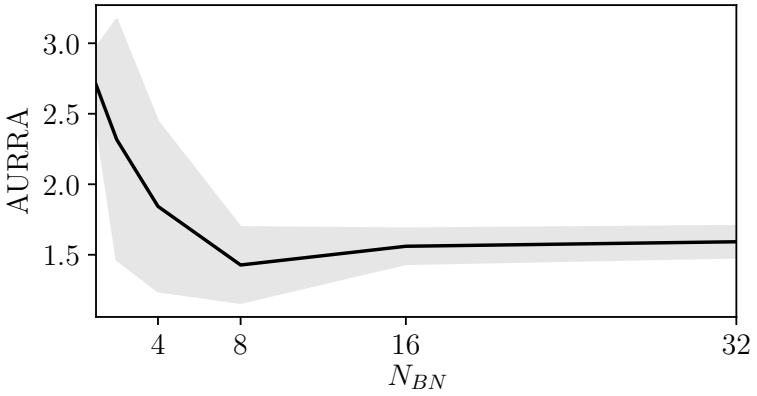


FIGURE 5.9: Model calibration, Area Under the Response Rate Accuracy curve (AURRA), on ImageNet-A (Hendrycks et al., 2019) as a function of bottleneck width. Networks with tighter bottlenecks exhibit improved calibration.

ImageNet validation data set. The ImageNet-A data set typically results in a top-1 accuracy of around 2-3% on the same models. Figure 5.8 shows the top-1 accuracy of our models on the ImageNet-A data set. Figure 5.9 shows the Area Under the Response Rate Accuracy curve (AURRA) which measures how calibrated the models are (how a model's confidence in its prediction relates to its accuracy). Both of these figures illustrate that the tighter bottlenecks do induce slightly better performance on this task.

5.3.2 Robustness to L_∞ constrained artificial attacks

To study a more general notion of adversarial robustness, we follow the guidelines given by Carlini et al. (2019). We use FoolBox (Rauber et al., 2017, 2020) to compute the worst case performance of the models under a range of attacks. We explore a range of perturbations using different L_∞ thresholds between 0 and 1.0. The attacks include Projected Gradient Descent, which was utilised by Dapello et al. (2020) to

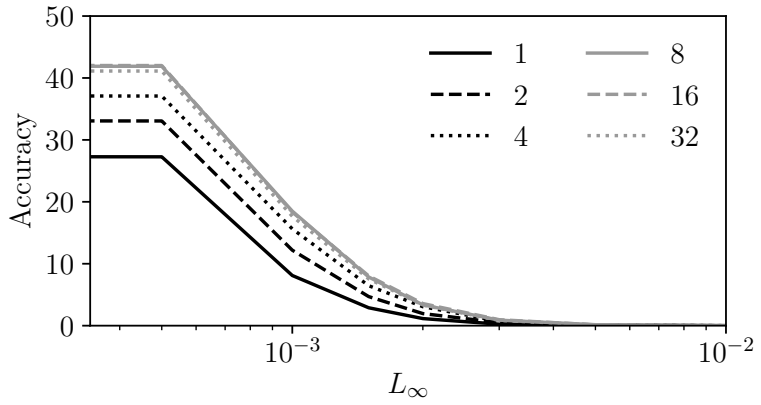


FIGURE 5.10: Robust accuracy of different bottlenecks (line styles) against L_∞ adversarial attack budgets.

demonstrate improved robustness of their model. We use the following adversarial attacks with L_∞ perturbation budgets between 0.0 and 1.1. Each attack uses the FoolBox (Rauber et al., 2020) implementation with default parameters.

- Fast Gradient Sign Method (Goodfellow et al., 2015)
- Projected Gradient Descent (Madry et al., 2018)
- Basic Iterative Method (Kurakin et al., 2017)
- Additive Uniform Noise
- DeepFool (Moosavi-Dezfooli et al., 2016)

Figure 5.10 shows the robust accuracy, the worst-case performance following all attacks, of the models across different L_∞ perturbation budgets. Results were computed by sampling 10000 images from the ImageNet validation set uniformly for each class. The results show that the increased opponency and small increase in shape bias actually correspond to a decrease in robustness to targeted adversarial examples.

It is now pertinent to ask whether this result is unique to the ImageNet-trained ResNet models or also applies to the models trained on CIFAR-10 from Chapter 4. We therefore perform the same suite of attacks as in our ImageNet experiments but over the whole CIFAR-10 test set. Figure 5.11 shows the worst case accuracy for these models following all attacks. The results again show that no significant difference is obtained through the introduction of a bottleneck.

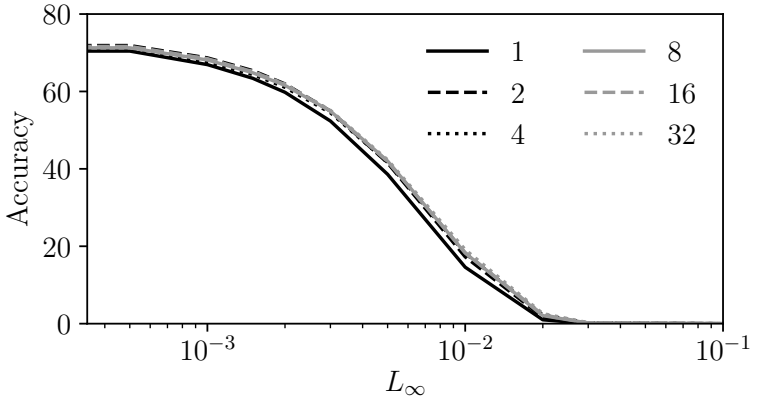


FIGURE 5.11: Robust accuracy of the CIFAR-10 trained models from Chapter 4 with different bottleneck widths (line styles) against L_∞ adversarial attack budgets.

5.4 Summary

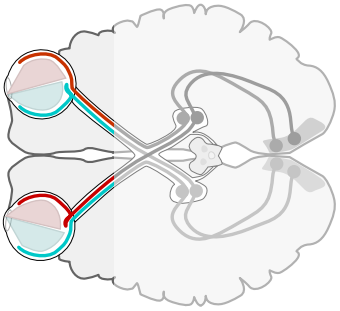
In this chapter we have explored the addition of a bottleneck to ImageNet trained ResNet-50 models. We have shown that cells in the bottleneck layer of models with tight bottlenecks learn opponent receptive fields in accordance with the CIFAR-10 results from Chapter 4. In contrast to those results, the opponent cells we find here are best described as luminance opponent rather than colour opponent. That said, we have shown that these luminant opponent cells still exhibit some form of spectral response and that, even in the extreme case of $N_{BN} = 1$, the model is able to learn about colour. Whether the cells in networks with tight bottlenecks have centre-surround receptive fields of the kind observed by Lindsey et al. (2019) is unclear, although there is some evidence that they do. It is clear, however, that the Retina-2 cells respond more linearly in networks with tight bottlenecks, echoing the finding of Lindsey et al. (2019). We have further shown that these differences in function correspond to a small but consistent increase in the shape bias of the networks.

Our investigation into the performance of the different bottlenecked models is surprising in the sense that it perhaps asks more questions than it answers. At the outset our hypothesis was that the bottleneck might encourage both shape bias as well as adversarial robustness; the typical nature of centre-surround opponent cells observed in prior studies would for example suggest that the early network would perform isotropic band-pass filtering (like a Laplacian of Gaussian) which one presumes would naturally bias to towards shape representations. In terms of accuracy, the different bottlenecks only have a mild effect with the tightest bottlenecks (below 8) having slightly worse performance. The results of testing the models on natural adversarial examples indicates that the tightest bottlenecks do perform better on the ImageNet-A data set (both in terms of accuracy and calibration). However, it is unclear whether this increase corresponds to an increase in robustness or merely indicates that the type of function learned differs for tighter bottlenecks. In any case,

the worst case accuracy, computed after a set of targeted adversarial attacks, shows that adversarial robustness actually decreases slightly as the bottleneck is tightened.

Chapter 6

Foveation and Function



In nature, visual attention is the movement of the eyes (Yarbus, 1967), and often body (Held and Hein, 1963), which enables us to process intractable quantities of information by restating the problem as one of time rather than bandwidth or functional capacity. In the previous chapters, we have shown that each convolutional filter in a convolutional neural network can be seen to have functional expressivity similar to that of a retinal ganglion cell. The tiling of this single filter over the image is accounted for by the spatial redundancy observable throughout the human visual system as a space preserving retinal mapping (retinotopy) (Purves et al., 1997). Indeed, modern convolutional neural networks (particularly models of attention) tick a lot of boxes in comparison to the simplified biological model. There is, however, one glaring omission: foveation.

In the context of deep learning, foveation is not well explored. This is likely due to the fact that many practitioners believe foveation to be superfluous to the objectives of artificial vision. However, even if foveation is truly without value to the goal of greater generalisation performance, an effective model of foveation may yet prove useful. Specifically, if we are able to characterise the conditions required for the emergence of the phenomena surrounding foveation, we may help to understand why our vision is foveated if not for performance reasons.

One such phenomenon, and as a consequence of the interconnectivity found throughout the visual pathway, is that neurons may integrate information over multiple scale spaces. In the Inferior Temporal (IT) cortex (part of the 'what' pathway), evidence has been found for neurons which modulate their receptive field according to the input. In Rolls et al. (2003), the authors studied neurons in macaque IT as they performed an object search task. Their results show neurons that have a large receptive field when the objects are placed in a blank scene. In contrast, when

the scene is complex, the receptive fields are markedly reduced in size. It is suggested that this translation variance facilitates an unambiguous representation of object features in complex scenes (Rolls and Deco, 2002). The potential for integration of information over multiple scale spaces and modulation of receptive field size is exhibited in foveation. Specifically, downstream neurons can trade off between relying on information from just one scale space or integrating features over multiple scale spaces. This enables a much more dramatic change in receptive field sizes than would be possible without foveation. It is therefore possible that foveation may provide value in a setting where a network is required to perform both a localisation task and a recognition task in parallel. Furthermore, we may look to such a network for the emergence of these multi-modal cells.

Regarding colour processing, it has been widely observed that colour sensitivity in humans and trichromatic primates decays with eccentricity (Mullen, 1991; Stromeyer III et al., 1992). In Chapter 4 we extensively studied the notion of opponency in neural networks. We further showed that the cells in the Retina-Net models are in fact channel (RGB) opponent. This is an analogue of the cone opponency observed in nature where the red-green (RG) and blue-yellow (BY) opponent systems pit L cones against M cones and S cones against a combination of L and M cones respectively. Several works have additionally shown differences between the cone contrast sensitivity of the RG and BY opponent systems with eccentricity (Mullen and Kingdom, 2002; Mullen et al., 2005; Murray et al., 2006; Hansen et al., 2009). RG sensitivity is much greater than BY sensitivity at the fovea but decays more rapidly, matching BY sensitivity within 25° of eccentricity (Mullen and Kingdom, 2002). The RG and BY opponent systems also exhibit distinct morphology. For example, there is evidence of BY opponency in the koniocellular layers of primate LGN with RG opponency mediated only by cells in the parvocellular layers (Hendry and Reid, 2000). These distinct pathways consequently innervate distinct areas of the primary visual cortex (Chatterjee and Callaway, 2003). Trichromatic vision, supporting RG opponency, is a recent evolutionary development compared to the older dichromatic vision, supporting BY opponency, of many primate species (Lucas et al., 2003). These unique evolutionary origins offer one explanation for the differences between the presentation of and discrepancy between the two cone opponent systems. However, it may also be the case that there are inductive biases in the object detection task that give rise to changes in RG and BY sensitivity with eccentricity.

In order to consider the above phenomena we require a model of foveation that incorporates both a scale-space and a spatially variant function. In this chapter, we start by introducing the foveated convolution. This is a stack of convolutional layers, each sampling at a different scale, that receive as input successively smaller centre crops of the image.

In a first experiment, we show that the Spatial Transformer Network (STN) (Jaderberg et al., 2015), a popular deep model of visual attention (Ablavatski et al., 2017), fails to perform in a visual search problem (scattered CIFAR-10: a CIFAR-10 (Krizhevsky, 2009) image is placed randomly on a large blank canvas and the network is required to classify it) that is trivial to a human observer. Although many components (for example, a visual memory (Baddeley and Hitch, 1974)) contribute to the success of human visual attention we show that the addition of a foveated convolution can dramatically improve the localisation performance of deep networks augmented with STNs. These attention policies are learned without any enhancement to or supervision of the localisation network (as in Ablavatski et al. (2017)) or increase in the total number of parameters.

In our second experiment, we study a variant of the STN model where the localisation and classification networks have shared weights. This enables us to study a setting where the convolutional network is tasked with both localisation and object recognition in parallel. We show that the addition of a foveated convolution improves performance on this task. Furthermore, by analysing the receptive fields of neurons in these networks we provide evidence for a multi-modal representation, enabled by the foveated convolution, that supports the dual objective.

In our final experiment, we extend the foveated convolution model to allow for multiple layers of spatially variant function. Studying opponency in these models, we find a reduction in opponent cells in the periphery, echoing the findings from nature. We show that this pattern does not emerge in a baseline model with a locally connected layer (which supports spatially variant function without a scale space). We subsequently propose an approach for inferring RG or BY stimulus preference, extending our work from Chapter 4. Applying this analysis to our multi-layer foveated models, we show a reduction in RG preference moving from the centre to the periphery. We again report results for a baseline locally connected network, in this instance showing that it exhibits the same reduction in RG preference in the periphery. On the basis of these findings, we suggest that the reduction in opponency in the periphery emerges as a result of the scale space representation offered by foveation. In contrast, the reduction in RG preference does not require a scale space and is instead in some way induced by the object detection task.

6.1 Foveated Convolutions

In this section we draw inspiration from the distribution of ganglion cells on the retina in order to construct a convolutional model of foveation following the observations from Section 2.1.3. To model foveation we require oversampling at the centre and undersampling at the periphery. We model oversampling with two transpose

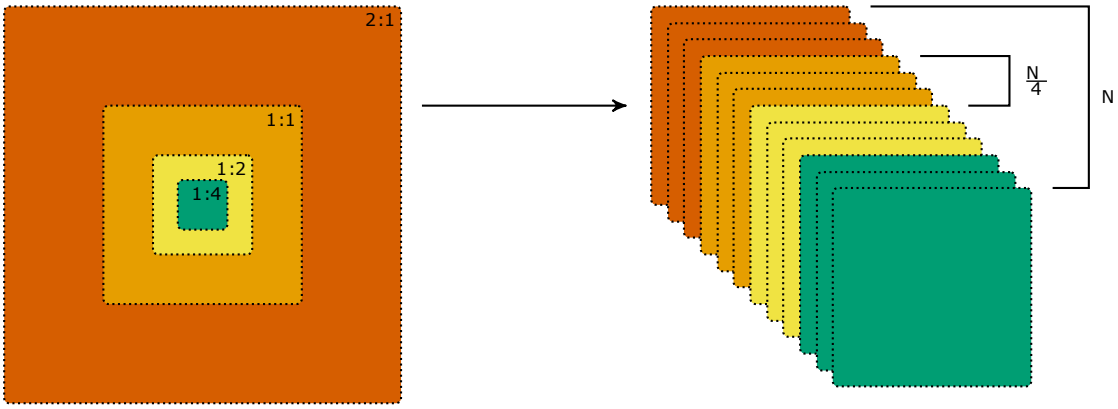


FIGURE 6.1: The foveated convolution operation. Left shows the crop area and 'input : output' pixel ratio for each convolution, right shows the output tensor for a foveated convolution with N channels. To add this component to the model from Figure 6.2 we replace the first layer of the classifier with a foveated convolution.

convolutions which operate only on the central regions of the input. We subsequently have a traditional convolutional layer with a stride of one which operates on a slightly larger region. Finally, we use layers with increased dilation and stride to model the undersampling that occurs towards the periphery. We take dilation and stride to be equal since the relative increase in dendritic spread is approximately inversely proportional to the change in ganglion-cone cell ratio. The input crop is chosen so that the output from each convolutional layer is equal and maximal (that is, the layer with the highest stride operates on the full image). Combining each of these design choices, we obtain a layer of the form depicted in Figure 6.1. This layer bares a strong resemblance to the sampling process in image foveation but with a fixed number of discrete scale spaces rather than a continuous transition.

The output of a foveated convolution with 32 channels (the number used in our experiments) will contain 8 channels from each of four convolutional layers. The first acts on the whole image with a stride and dilation of 2. The second, acting on a centre crop of the image with half the width and height, has a stride and dilation of 1. The third layer, acting on a centre crop of the image with a quarter of the width and height, is a transpose layer with a stride of 2. The final layer, acting on a centre crop of the image with an eighth of the width and height, is a transpose layer with a stride of 4. By virtue of their stride, the dilated layers address only a fraction of the pixels in the image. Although this is appropriate since there are fewer cone cells towards the periphery, we can further model the increase in size of peripheral cone cells by swapping the dilated layer with an average pooling followed by a convolution with a stride and dilation of 1. All convolutional layers have a kernel size of 3 in our foveated layer.

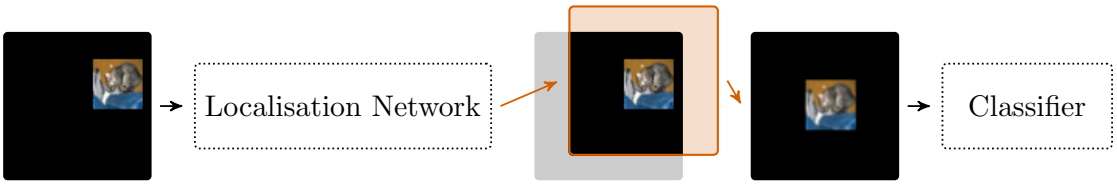


FIGURE 6.2: The basic network architecture for a translation-only STN. At each step, the output of the localisation network is used to construct a sampling grid. This is then interpolated over the image to obtain the classifier input.

6.2 Experiment One: The Impact of Foveation on Localisation Performance

In this section we present our first experiment studying the impact of foveated convolutions on localisation performance of a Spatial Transformer Network (STN) (Jaderberg et al., 2015). We construct the scattered CIFAR-10 data set which illustrates a limitation of standard STNs and show that the addition of foveated convolutions can alleviate this problem.

6.2.1 Scattered CIFAR-10

Translation invariance is a property long thought desirable in image processing models (Nixon and Aguado, 2012). A model can be considered translation invariant if the quality of its inference is not altered by translations of its inputs. However, translation invariance is not always desirable; if the feature extraction following a spatial attention mechanism is translation invariant, there will be no motivation for the model to learn an interesting policy. By virtue of their action over the input, convolutional networks exhibit a degree of translation invariance. This makes attention mechanisms such as STNs followed by standard convolutional layers sometimes struggle to localise the input.

Consider a visual search problem we call scattered CIFAR-10 where a CIFAR-10 (Krizhevsky, 2009) image is placed randomly on a large blank canvas and the network is required to classify it. It is trivial for a human observer to determine the location of the image in the canvas and subsequently attempt to classify it. An illustration of the scattered CIFAR-10 problem and the architecture of a translational STN, where the attention mechanism is limited to only translating the input rather than a full affine transform, is given in Figure 6.2. To perform well on scattered CIFAR-10, the attention mechanism will need to learn to align the images. If that alignment is made to a repeatable point, such as the centre of the image, the challenge becomes effectively solved. Specifically, one can envisage a two stage process where the attention policy is first learned using a simple classification network and then fixed to act as a pre-processor for a more complicated network. In this setting, if the

FIGURE 6.3: The transformed results, after training on scattered CIFAR-10, for: a Spatial Transformer Network (STN), an STN with multi-scale extraction (STN-MS), a FoveaNet (FN), and STN with a full affine glimpse (STN-FA), and a FoveaNet with a full affine glimpse (FN-FA).

TABLE 6.1: Classification error of a three layer CNN equipped with different attention mechanisms on scattered CIFAR-10.

Model	Error
CNN	$58.77 \pm 0.29\%$
CNN with global pooling	$52.71 \pm 0.47\%$
STN	$60.02 \pm 2.18\%$
STN multi-scale	$57.55 \pm 2.72\%$
FoveaNet	$\mathbf{50.41 \pm 1.39}\%$
Pooled FoveaNet	$\mathbf{49.53 \pm 0.82}\%$
STN *full affine*	$47.70 \pm 1.70\%$
Pooled FoveaNet *full affine*	$\mathbf{45.83 \pm 0.61}\%$

localisation network (which decides the glimpse parameters) performs well, it should be possible to get the same accuracy as can be obtained on vanilla CIFAR-10 with any architecture.

6.2.2 Localisation Performance

The second row of Figure 6.3 shows the transformed images emitted by the attention mechanism of a simple CNN with a translational STN over the input (top row). The classification error for this model and a non-attentional baseline with the same classification architecture are given in Table 6.1. This model uses a localisation network of four convolutional layers (kernel sizes of 7, 5, 3 and 3) followed by a linear layer with 32 neurons and a final linear layer which regresses the 2 translation parameters. Following the translational affine transformation, a three layer convolutional network (kernel size of 3 and stride of 2 in each layer), 128 neuron linear layer and final 10 neuron softmax layer perform the classification step. We use ReLU nonlinearity throughout and train for 50 epochs with the Adam optimiser (initial learning rate of 0.0001). From the figure we can see that the model fails to appropriately localise the input, instead collapsing to zero translation in both directions.

We additionally conduct experiments using the multi-scale approach employed by Ba et al. (2014); Mnih et al. (2014). This approach is costly when using STNs since the image needs to be differentiably scaled multiple times on the forward pass. The third row of Figure 6.3 shows the outcome of this experiment, extracting glimpses at 3 scales. The figure shows an improvement over the baseline when using multi-scale glimpse extraction.

The fourth row of Figure 6.3 shows the result of our model augmented with a foveated convolution layer (substituted for the first layer of the classifier). This has the same number of weights as the previous model but is now able to almost perfectly solve the localisation problem. Consequently, the terminal error of the model (given in Table 6.1) has significantly improved. The foveated layer with pooling instead of strided downsampling (Pooled FoveaNet) gives a small improvement in localisation performance and accuracy as shown in Table 6.1.

It could be argued that the STN performs poorly as a result of the constraint that it can only translate the input. The fifth row of Figure 6.3 shows the transformed images in a model with a full affine STN. The network still fails to localise in this setting, whilst adding unnecessary rotation and scale. The sixth row show the transformed images for a FoveaNet with a full affine mechanism. In this setting we still find unnecessary rotation and scale but the scattered images have been correctly pose normalised. These observations are also echoed by the classification error in Table 6.1.

An alternative approach to translation invariance is to equip a CNN with global pooling. This takes a spatially global average of each feature map such that the terminal activations are translation invariant. As Table 6.1 shows, the performance is still worse than that of a foveated convolution, and of course without any of the localisation benefits.

6.2.3 Foveated Pre-processing

The results in Figure 6.3 show that when the localisation problem is effectively solved the scattered image is reliably transformed to the centre of the canvas. This property makes it possible to use the pre-trained STN mechanism followed by a centre crop as a preprocessing technique. Since the STN is lightweight for inference and does not need to be trained any further, this can be achieved with a negligible change in computation cost.

We now perform experiments using this pre-processing followed by a PreAct ResNet-18 for classification. In this setting, the full advantage of foveated convolutions becomes clear. The results, given in Table 6.2, show that foveated convolutions can be used to completely solve scattered CIFAR-10, obtaining a superior result to the ResNet trained on vanilla CIFAR-10. We suspect that the reason for the increase is that the attention

TABLE 6.2: Classification error of a PreAct ResNet-18 following preprocessing by applying a centre crop on the transformed output of a pre-trained attention mechanism on scattered CIFAR-10.

Model	Error
No crop	$\mathbf{6.17 \pm 0.08}\%$
STN + crop	$53.77 \pm 0.63\%$
STN multi-scale + crop	$38.66 \pm 9.08\%$
Pooled FoveaNet + crop	$\mathbf{6.13 \pm 0.08}\%$
Vanilla CIFAR-10	$9.28 \pm 0.09\%$

No crop: 10 hours, 12GB GPU memory
Others: 2 hours, 2GB GPU memory

step acts as a small augmentation, similar to a padded random crop. As a further baseline, we measure the performance of the ResNet-18 trained on full scattered CIFAR-10 images. Although this performs very well, the increased resolution resulted in an approximately five fold increase in memory usage and training time. Strangely, the ResNet trained on scattered CIFAR-10 outperforms the same network on vanilla CIFAR-10; we again suspect that this is because the scattering acts as an extremely costly augmentation.

6.3 Experiment Two: Multi-modal Neurons in Foveated Networks

In this section we propose a variant of the model from Section 6.2 where the weights are shared between the localisation network and the classifier and the convolutional architecture is based on the Retina-Net model from Lindsey et al. (2019) and Chapter 4. We train this model on a more challenging variant of the scattered CIFAR-10 task where additional distractor fragments are added that we refer to as cluttered CIFAR-10. Finally, we study the receptive fields of cells in these networks and find evidence for a change in receptive properties depending on the stimulus that we suggest facilitates the dual objective of localisation and classification.

6.3.1 Cluttered CIFAR-10

Inspired by Cluttered MNIST used in (Ba et al., 2014), we now construct Cluttered CIFAR-10 by embedding CIFAR-10 images in a canvas along with randomly sampled patches from the rest of the dataset. This is a more challenging variant of the visual search task explored in Section 6.2. Figure 6.4 shows the cluttered CIFAR-10 input and transformed outputs for two models. The first model is the result of replacing the

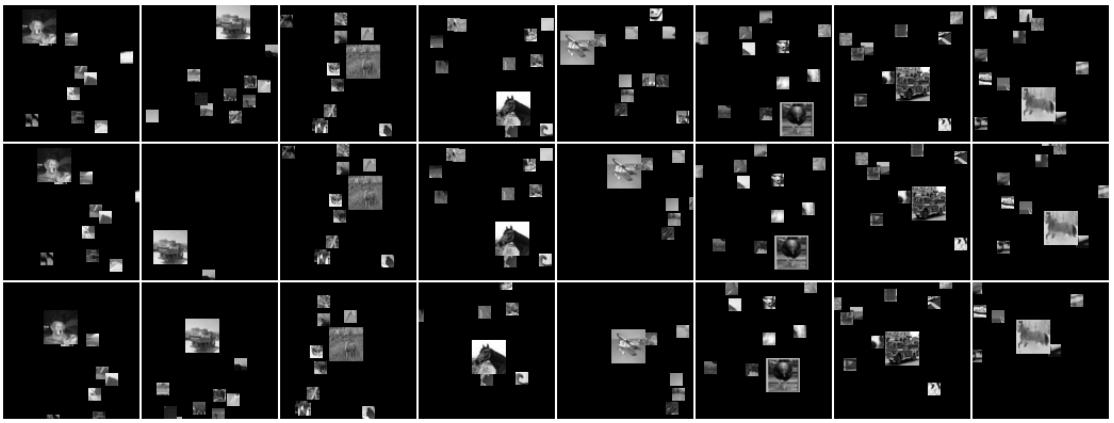


FIGURE 6.4: Top: The model input for cluttered CIFAR-10. Middle: The transformed outputs for the non-foveated model. Bottom: The transformed outputs for the foveated model

localisation and classification networks with the Retina-Net architecture with a bottleneck of 8 and a ventral depth of 4 (refer to Section 4.2.1 for details on the meaning of these hyper-parameters). The second model is a variant of the first where the first layer (Retina 1) has been substituted with a foveated convolution. For both models the weights of the localisation and classification network are shared. The figure reflects the results from 6.2 in that the standard model (middle row) generally regresses to the identity transform, whereas the model augmented with foveated convolutions manages to correctly localise the image in at least some of the examples despite the increased complexity of the problem. This improved localisation is also reflected in the terminal classification errors; the non-foveated model achieved an error of 64.4%, whereas the foveated model achieved an improvement at 57.8%.

6.3.2 Receptive Field Analysis

In order to visualise the receptive fields of neurons in our models, we can employ the approach of Lindsey et al. (2019) where the gradient of the cell is taken with respect to a blank stimulus. For our purposes, we need a way to differentiate between receptive fields with respect to simple stimulus (representing the background) and complex stimulus (representing the object). Since the cluttered CIFAR-10 images do not resemble natural images, we note that any stimulus where the input is greater than zero corresponds to stimulus from an object. We therefore study receptive fields with respect to two inputs: one containing zeros representing the background and another containing ones representing the foreground.

Figure 6.5 gives the receptive fields for neurons in the 'V1' layer of the non-foveated model in response to bright stimulus (top) and dark stimulus (bottom). The figure shows that the receptive fields with respect to the two stimuli are largely consistent,

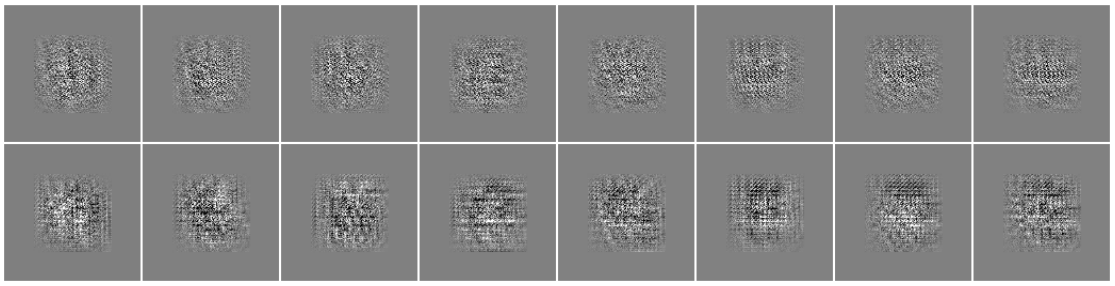


FIGURE 6.5: Top: Receptive fields for cells in layer 'V1' of the non-foveated model following a bright stimulus. Bottom: Receptive fields for the corresponding cells in response to a dark stimulus.

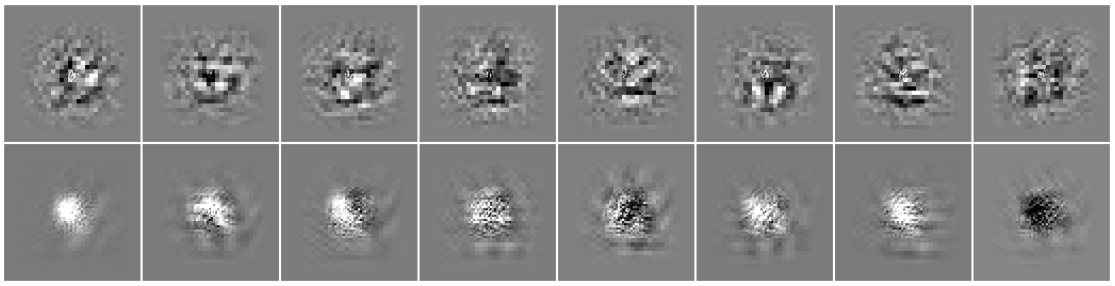


FIGURE 6.6: Top: Receptive fields for cells in layer 'V1' of the foveated model following a bright stimulus. Bottom: Receptive fields for the corresponding cells in response to a dark stimulus.

both in terms of the size of the receptive field and the sensitivity (brightness of the image here corresponds to the magnitude of the gradient).

Figure 6.6 gives the receptive fields for neurons in the 'V1' layer of the non-foveated model. Here we find a stark contrast between the receptive fields with respect to the two stimuli. With respect to the bright foreground stimulus, receptive fields are large and complex. With respect to the dark stimulus receptive fields are small, typically with a single region of consistent high sensitivity at the centre.

There are a few observations we can make regarding these results. The first is that the cells in the foveated network are highly non-linear in their activation. This derives from the fact that a linear cell would have the same gradient with respect to all stimuli. The second observation is that these receptive fields correspond to two modes of operation for the neurons: classification, and localisation. In the classification mode (bright stimulus) the cells exhibit a wide and detailed receptive field required to convey information about the features in the object (the CIFAR-10 image). In the localisation mode (dark stimulus) the cells exhibit a small, specific receptive field needed to convey information about the precise location of the stimulus. The presence of these multi-modal neurons is enabled by the scale space representation of the foveated convolution layer.

Having characterised the receptive fields on neurons in the 'V1' layer, it is now pertinent to ask whether these receptive field properties are upheld in later layers. To

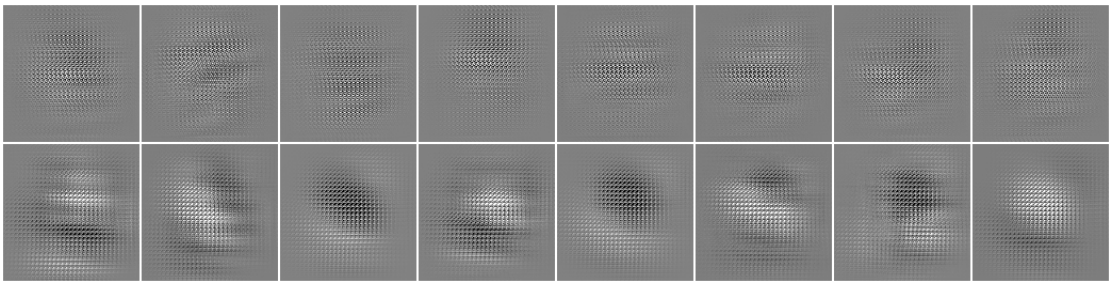


FIGURE 6.7: Top: Receptive fields for cells in layer 'V4' of the non-foveated model following a bright stimulus. Bottom: Receptive fields for the corresponding cells in response to a dark stimulus.

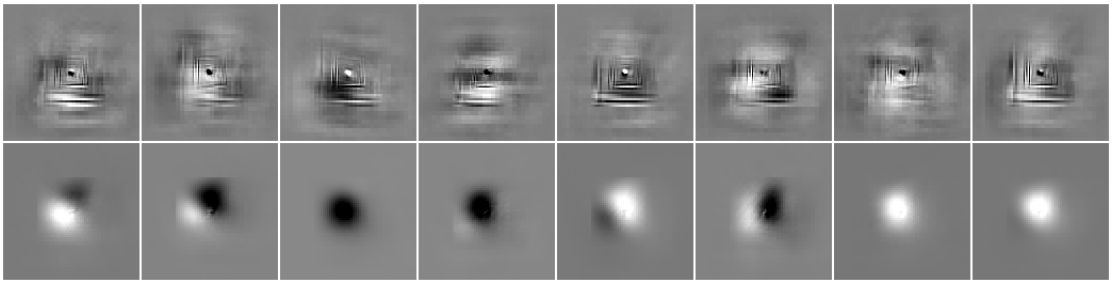


FIGURE 6.8: Top: Receptive fields for cells in layer 'V4' of the foveated model following a bright stimulus. Bottom: Receptive fields for the corresponding cells in response to a dark stimulus.

this end, in Figures 6.7 and 6.8 we plot the receptive fields for cells in the 'V4' layer of the non-foveated and foveated models respectively. Our findings here echo our findings in the 'V1' layer. At this higher level of abstraction, the 'V4' cells in the foveated model in localisation mode (dark stimulus) now appear to act as rudimentary edge and blob detectors. These cells convey features that are of clear utility when trying to find the CIFAR-10 image in a cluttered canvas.

6.4 Experiment Three: Opponency and Eccentricity

In this section we study the distributions of opponent cells against eccentricity according to the definitions from Chapter 4 in networks equipped with foveated convolutions. We analyse these distributions in order to determine whether the characteristic differentials observed in nature emerge. We additionally propose a new method for determining the stimulus preference of cells in our networks in order to further study the distribution of RG and BY preference.

6.4.1 Multiple Layers of Foveation

For our experiments here we require two layers of foveated function since we have observed in Chapter 4 that receptive fields akin to those of retinal ganglion cells are

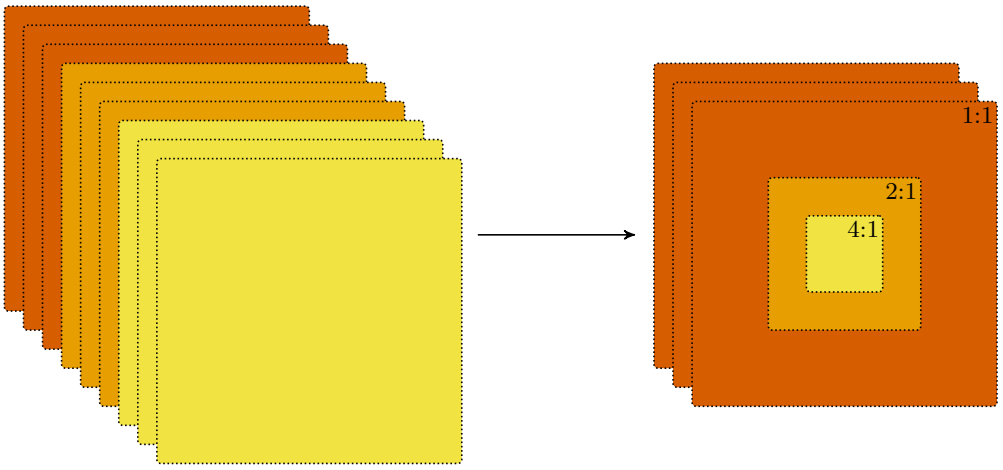


FIGURE 6.9: The inverse foveation pooling operation. Groups of feature maps from different scale spaces are resampled and inserted in to each other to produce a single feature map.

predominantly found in second layer units. In general, strided and transpose convolutions can be substituted for interpolation followed by a standard convolution. Using this principle, we can separate the foveated convolution model into a stack of interpolation layers, which each receive different centre crops of the input, followed by a stack of convolutional layers which all have the same hyper-parameters. In essence this is as a spatial pyramid followed by parallel independent convolutional layers.

In addition to multiple layers of foveation, we require the ability to increase the number of scales so that we have a smooth transition in scale space across the image (in the retina this transition is continuous). The number of scales in the pyramid, N_S, is a hyper-parameter of the model. The size of the centre crop is chosen such that the scaled output has the same spatial resolution as the original image. We then sample the scale factors for the interpolation uniformly sampled between 1 and 4. Note that we avoid downsampling such that when $N_S = 1$ this is equivalent to a standard non-foveated model. We follow the spatial pyramid with two layers of grouped convolutions in N_S groups with kernel size of 3, stride of 1, and no input padding. Grouped convolutions are equivalent to parallel layers with the same hyper-parameters. There are $32 \times N_S$ filters in the first layer and $64 \times N_S$ in the second layer.

Following the foveated portion of the network we need to re-scale the different groups to all be in the same scale space. In addition, we need to remove the centre portions of the peripheral groups since they should only communicate information about the periphery. This can be achieved with an inverse of the spatial pyramid operation, where successive scales are downsampled and inserted into the centre of the previous scale. The output from this layer is a single block of feature maps where the distance of a given pixel from the centre determines which group of convolutions it derives from. A diagram of this inverse pooling operation is given in Figure 6.9. For the remainder of the model we use four convolutional layers with 128, 256, 512, and 512 units

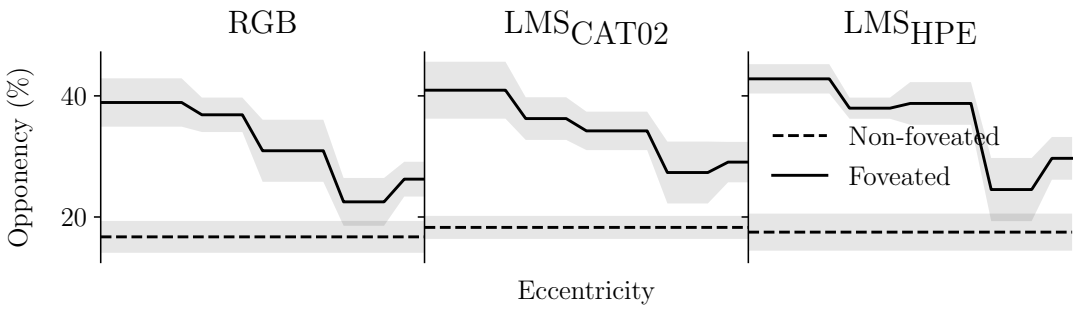


FIGURE 6.10: Distribution of colour opponent cells against eccentricity in foveated ($N_S = 7$) and non-foveated ($N_S = 1$) variants of our model trained on CIFAR-10. Colour opponency increases in the centre of the visual field to over twice that of the periphery.

respectively followed by a fully connected layer. We use kernel size of 3, stride of 2, and input padding of 1 for the convolutional layers.

We additionally report results using locally connected networks. These are akin to a convolution where the weights of each filter at each spatial position are learnable independently of each other. These locally connected networks exhibit spatially variant function but not spatially variant scale.

To study the impact of channel responses on opponency, we train networks on images in multiple colour spaces that have different cone response characteristics. Specifically, we use: RGB, $\text{LMS}_{\text{CAT02}}$, and LMS_{HPE}. A visual depiction of the RGB / LMS channel responses to wavelength and a description of the approach for converting between the colour spaces is given in Section 2.1.1.

In order to follow deep learning best practices we normalise inputs to have zero mean and unit deviation according to statistics computed over the whole data set after transformation into the target colour space. All models are trained for 200 epochs using Stochastic Gradient Descent (SGD) with momentum of 0.9, an initial learning rate of 0.1, and L2 weight decay of 10^{-4}. We divide the learning rate by 10 at epochs 100 and 150. For data augmentation, we use random horizontal flipping only. We use batchnorm layers followed by ReLU activation functions for all hidden layers in all models. For the experiments in this section, the error region gives the 95% confidence interval following 10 repeats with independent random initialisation of the weights.

6.4.2 Distribution of Opponent Cells

Figure 6.10 gives the distribution of opponent cells against eccentricity in models trained on the CIFAR-10 data set in the RGB, $\text{LMS}_{\text{CAT02}}$, and LMS_{HPE} colour spaces respectively. The figure shows a peak in opponency in the centre of the visual field and a reduction in the periphery that is consistent in all three colour spaces. The

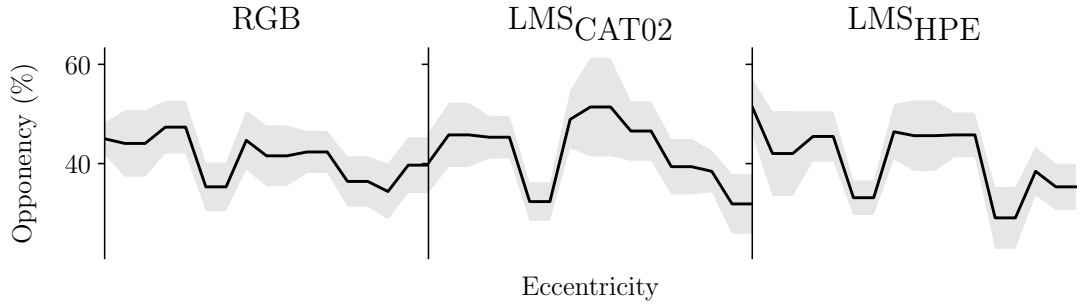


FIGURE 6.11: Distribution of colour opponent cells against eccentricity in our locally connected model trained on CIFAR-10. Locally connected layers, where receptive field size is constant across the visual field, do not exhibit a change in opponency with eccentricity.

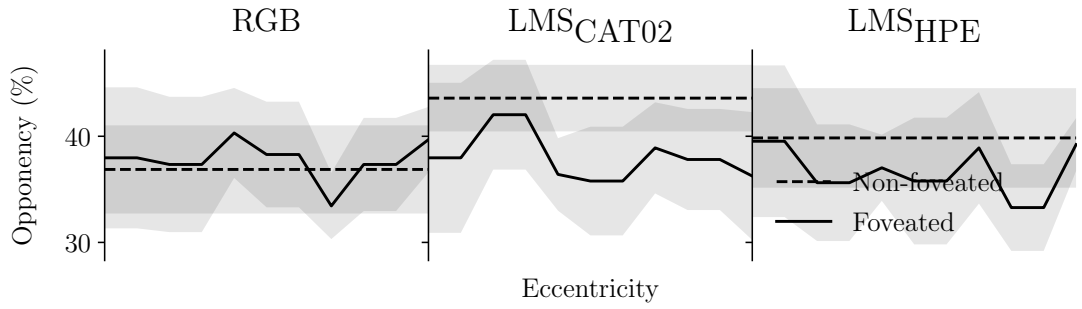


FIGURE 6.12: Distribution of colour opponent cells against eccentricity in foveated ($N_S = 7$) and non-foveated ($N_S = 1$) variants of our model with random weights. Random networks do not exhibit a change in opponency with eccentricity.

non-foveated networks exhibit opponency at the same level as the perihperal cells in the foveated networks. This is to be expected since these neurons have the same receptive field size and scale space.

Figure 6.11 gives the opponency distribution with eccentricity for networks with locally connected layers. The results show that colour opponency stays broadly constant across the visual field in these models. This suggests that the scale space representation is required in order for the differential distribution of opponency to emerge.

We additionally report results for the same analysis in networks with random weights in Figure 6.12. The figure shows that the opponency distribution does not change significantly with eccentricty. Furthermore, the results for foveated and non-foveated networks are within the margin of error of each other. This suggests that the differential distribution of opponency is a learned result rather than an innate property of the networks.

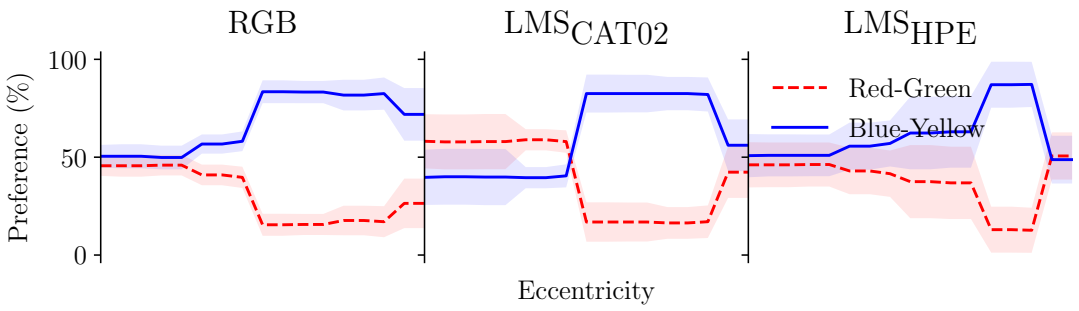


FIGURE 6.13: Distribution of channel opponent stimulus preference against eccentricity in a foveated ($N_S = 7$) model trained on CIFAR-10. RG preference increases in the centre of the visual field and is traded for BY preference the periphery.

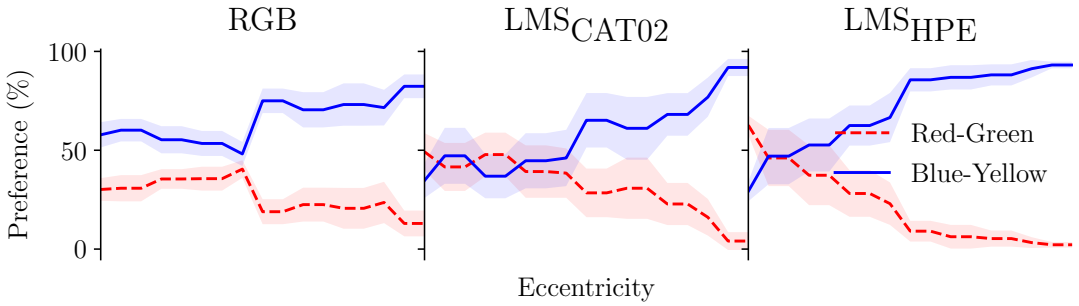


FIGURE 6.14: Distribution of channel opponent stimulus preference against eccentricity in our locally connected model trained on CIFAR-10. RG preference increases in the centre of the visual field and is traded for BY preference the periphery.

6.4.3 Distribution of RG / BY Preference

We now wish to determine the preference for RG / BY channel opponency in our networks. To this end we construct two sets, RG and BY, of sinusoidal grating stimuli using PsychoPy (Peirce et al., 2019). Each grating set is configured to isolate RG or BY inputs. We first determine the maximum response of each unit to each grating set over all angles, frequencies, and phases. We then label each unit as either preferring RG or BY stimuli depending on which grating set elicits the strongest response. Note that a consequence of this approach is that all cells are labelled as either RG or BY preferring. This contrasts with the approach from biology where both RG and BY sensitivity can decay with eccentricity.

In Figure 6.13 we plot the distributions of RG and BY preference with eccentricity for networks equipped with foveated convolutions. The figure shows that RG preference is at it's highest in the centre of the visual field and decays towards the periphery. There is a drop in RG preference at the very edge of the visual field that we suspect is an artefact of the foveated convolution.

Figure 6.14 gives the RG and BY preference curves for our locally connected model. The figure shows that our locally connected variant exhibits the same characteristic

change in RG and BY preference that we found in networks with foveated convolutions. We do not find the drop in RG preference at the edge of the visual field in this setting. This accords with our assertion that these drops are an artefact of the foveated convolution.

6.5 Summary

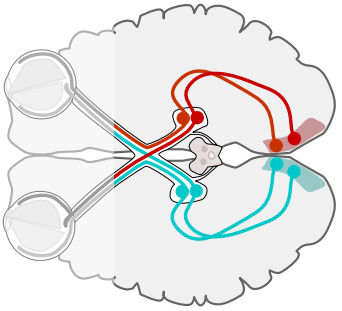
In this chapter we have taken inspiration from biology to design the foveated convolution, a layer of stacked convolutions with a foveated receptive field. We have shown that the addition of a foveated convolution dramatically improves localisation performance in a visual search task and that this gives a corresponding reduction in classification error. This unsupervised localisation is sufficiently reliable that a PreAct ResNet-18 can be trained on top to completely solve scattered CIFAR-10 with an approximately five fold reduction in complexity over training on the full images. We have further shown that foveation can facilitate a multi-modal representation where cells switch between a localisation mode and a classification mode depending on the nature of their input.

One interpretation of the scale space representation of the foveated convolution is that it changes the data distribution with eccentricity. Specifically, patches from each scale are sampled with a different scale factor and thus contain distinct visual features as a result. In Chapter 4 we showed that the data distribution plays a critical role in the emergence of opponency. For example, cells in networks with random weights do not exhibit the same opponent characteristics. Furthermore, experiments with different data sets yielded changes in the opponency distribution. On this basis, we contend that foveation gives rise to the differential distribution of opponency with eccentricity by changing the data distribution. This is consistent with the results presented here where a foveated convolution, which exhibits both a spatially variant scale space and function, gave rise to a differential distribution of opponency with eccentricity. In contrast, a locally connected network, which exhibits spatially variant function but consistent scale, did not give rise to a differential distribution of opponency with eccentricity.

Regarding the differential distribution between RG and BY, we find the same characteristic curve for both foveated convolutions and locally connected networks. This finding suggests that the RG / BY differential is not caused by the scale space representation of foveation but instead emerges from the data distribution and the object classification task. However, more experimentation would be needed to understand this relationship fully.

Chapter 7

Attention and Memory



In the previous chapters we have considered techniques for analysing specific aspects of deep networks at a cellular level using insight from electrophysical and psychophysical studies. However, much of the effort applied to understanding visual intelligence has been devoted to a higher-level, psychological, model of vision. In this chapter we look to build a deep model that accords with a high level view of the visual system and thereby understand the challenge and potential of such an approach.

David Marr posited that vision is composed of stages which lead from a two dimensional input to a three dimensional contextual model with an established notion of object (Marr, 1982). This higher order model is built up in the visual working memory as a visual sketchpad which integrates notions of pattern and texture with a notion of pose (Baddeley and Hitch, 1974). Visual attention plays a key role in the construction of these percepts by enabling objects to be appreciated from a range of perspectives and their spatial footprint to be inferred.

In deep learning, there exists an extensive collection of approaches to visual attention (Ablavatski et al., 2017; Ba et al., 2014; Gregor et al., 2015; Jaderberg et al., 2015; Mnih et al., 2014; Sønderby et al., 2015). Directly inspired by visual attention in nature, deep visual attention corresponds to adaptive filtering of the model input typically through the use of a glimpsing mechanism which allows the model to select a portion of the image to be processed at each step. With the objective of integrating pose and texture, visual attention models face two key challenges:

- to derive notions of pose and object from visual features,
- and to effectively model long range dependencies over a sequence of observations.

Various models have been proposed and studied which hope to enable deep networks to construct a notion of pose. The representation of pose can either be implicit or explicit in the nature of the objective. For example, transformational attention models learn an *implicit* representation of object pose by applying one or more transforms to an image with the objective of object classification (Jaderberg et al., 2015; Ablavatski et al., 2017). The transform policy in these cases has the objective of improving downstream performance, that is, making it easier for the resultant image to be classified. It is presumed therefore that the transform is pose-normalising in the sense that it removes information about the objects pose and retains information about its class. Other models such as Transformational Autoencoders and Capsule Networks harness an explicit understanding of positional relationships between objects and object-parts (Hinton et al., 2011; Sabour et al., 2017).

A long range dependency is a connection between two observations that are far apart (at a long range) in a sequence. In visual attention this concept encompasses several important properties. For example, the realisation that when you look away from and then back to an object it is still the same object (object permanence) requires understanding of a long range dependency. In deep learning, sequence processing tasks are typically handled using Recurrent Neural Networks (RNNs).

Short term memories have previously been studied as a way of improving the ability of RNNs to learn long range dependencies. The ubiquitous Long Short-Term Memory (LSTM) network is perhaps the most commonly used example of such a model (Hochreiter and Schmidhuber, 1997). More recently, the fast weights model, proposed by Ba et al. (2016) provides a way of imbuing recurrent networks with an ability to attend to the recent past and thus establish connections between current and past observations. From these approaches, it is evident that memory is a central requirement for any method which attempts to augment deep networks with the ability to attend to visual scenes.

The core concept which underpins memory in neuroscience is synaptic plasticity, the notion that synaptic efficacy, the strength of a connection, changes as a result of experience (Purves et al., 1997). These changes occur at multiple time scales and, consequently, much of high level cognition can be explained in terms of the interplay between immediate, short and long term memories. An example of this can be found in vision, where each movement of our eyes requires an immediate contextual awareness and triggers a short term change. We then aggregate these changes to make meaningful observations over a long series of glimpses. Fast weights (Ba et al., 2016) draw inspiration from the Hebbian theory of learning (Hebb, 1949) which gives a framework for how this plasticity may occur. In Miconi et al. (2018), the authors show that weights updated by a Hebbian rule can be combined with weights learned through backpropagation to act as a content-addressable memory.

In this chapter, we propose expanding transformational visual attention models to incorporate visual memory at multiple time-scales. These are: weights learned through back-propagation (long term), a parts-based representation formed from a series of sketches (working / medium-term), plastic weights (short-term), and the hidden state of an RNN (immediate). In the process of this endeavour, we wish to understand the challenges faced when attempting to use deep learning to realise a high-level psychological model. At its core, our model consists of LSTM layers that produce a series of glimpses (different affine transforms of the input). Our model then includes a Hebbian mechanism based on the Perceptron from Rosenblatt (1962) which integrates features from these glimpses. We show that this model can be trained to classify images from the MNIST and CIFAR-10 datasets. We then demonstrate that it is possible to learn this memory representation in an unsupervised manner by painting images, similar to the Deep Recurrent Attentive Writer (DRAW) network (Gregor et al., 2015). Using this representation, we demonstrate competitive classification performance on MNIST and performance on CIFAR-10 that exceeds a baseline VAE with self supervised features. Furthermore, we demonstrate that the model can learn a disentangled space over the images in CelebA (Liu et al., 2015) separating the faces from their backgrounds, shedding light on some of the higher order functions that are enabled by visual attention and associative memory. Finally, we show that the model can perform both classification and sketching tasks in parallel to produce a visually interpretable classifier.

7.1 An Associative Visual Working Memory

In this section, we detail our model. At a high level, we build on the model from Ablavatski et al. (2017) to include a Hebbian network that integrates glimpse features and a mechanism for image reconstruction (auto-encoding) that we refer to as a visual sketchpad. The Hebbian mechanism, our 'Memory Network', is akin to a short-term memory in the sense that it acts as a store for information from the whole sequence. We will later show that a key advantage of the Hebbian approach lies in the ability to project multiple query vectors through the memory and obtain different, goal-oriented, latent representations for one set of weights. Derived from the memory network are sketch features, sketches, and finally the full canvas. The canvas is the result of transforming and combining a sequence of sketches according to positional information from the glimpses. Combined, these components are an analogue of a working memory, where the same information is represented in parallel at multiple levels of abstraction.

In Section 7.1.1 we detail the plastic 'Memory Network' and update policy. Then in Section 7.1.2 we describe our model of visual attention which is used to build up the memory representation for each image. Finally, in Section 7.1.3 we will discuss the

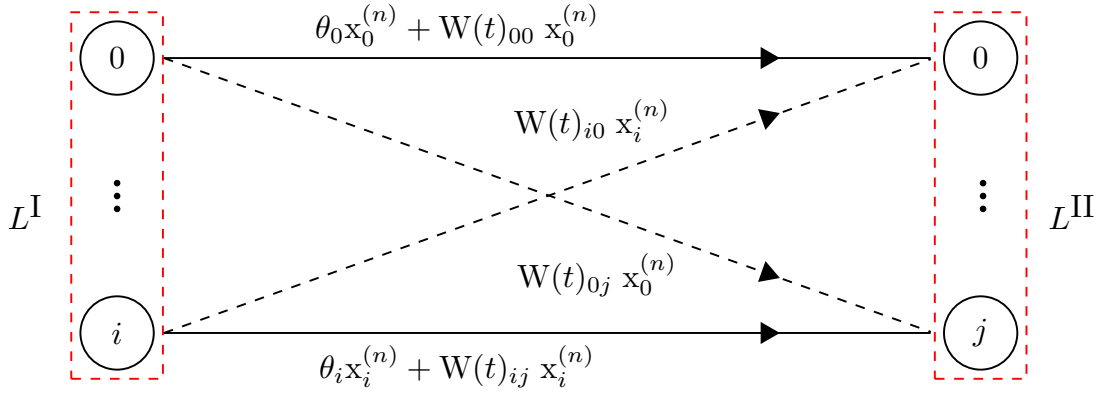


FIGURE 7.1: The two layer Hebb-Rosenblatt memory architecture. This is a novel, differentiable module inspired by Rosenblatts perceptron that can be used to 'learn' a Hebbian style memory representation over a sequence and can subsequently be projected through to obtain a latent space.

different model 'heads' which use these spaces to perform different tasks, including our sketchpad mechanism.

7.1.1 Hebb-Rosenblatt Redux

We require our Hebbian mechanism to concurrently perform two tasks, learning from the conditioning stimuli (the glimpses) and being queried to perform downstream objectives (classification and sketching). We draw inspiration from various models of plasticity to derive a differentiable learning rule that can be used to iteratively update the weights of a memory space, during the forward pass of a deep network. This memory layer can subsequently be queried to obtain context dependant latent spaces. Early neural network models used learning rules derived from the Hebbian notion of synaptic plasticity (Hebb, 1949; Block, 1962; Block et al., 1962; Rosenblatt, 1962). The phrase 'neurons that fire together wire together' captures this well. If two neurons activate for the same input, we increase the synaptic weight between them; otherwise we decrease. The information in biological networks is typically seen as being encoded by the timing of activations and not their number. However, a windowed calculation of the firing rate provides a reasonable, continuous valued approximation (Dayan and Abbott, 2001; Rolls and Deco, 2002) that we can use to create a differentiable short-term memory unit that can be integrated with deep networks.

Consider the two layer network inspired by the early multi-layer perceptron models of Rosenblatt (1962) shown in Figure 7.1 with an input layer L^{I} and output layer L^{II}, with weights at time t given by $\mathbf{W}(t) \in \mathbb{R}^{S_m \times S_m}$, where S_m denotes the size of the memory. Consider now that we present a sequence of conditioning stimuli

$$X = (\mathbf{x}^{(n)})_{n=0}^{N_s},\ \mathbf{x}^{(n)} \in \mathbb{R}^{S_m} \tag{7.1}$$

to the model. For the stimulus $\mathbf{x}^{(n)}$, L^{I} will have an activation $\phi(\mathbf{x}^{(n)})$, where

$$\phi : \mathbb{R}^{S_m} \to [0, \Phi]^{S_m} \subset \mathbb{R}^{S_m} \tag{7.2}$$

is a neuron-wise bounded rectified function. A corollary of this in nature is that the firing rate in a given window is non-negative and has a maximum value governed by the refractory period (Dayan and Abbott, 2001). For simplicity, we will consider the memory input to have been pre-processed such that $\mathbf{x}^{(n)} = \phi(\mathbf{x}^{(n)})$.

Cells in L^{II} receive stimulus via two sets of afferent connections and consequently activate according to two terms, the bias and the projection. Let

$$\boldsymbol{\beta}^{(n)} = \mathrm{diag}(\boldsymbol{\theta})\mathbf{x}^{(n)} \tag{7.3}$$

represent the bias term (where $\boldsymbol{\theta} \in \mathbb{R}^{S_m}$ are tunable, neuron-wise bias weights) and

$$\boldsymbol{\gamma}^{(n)}(t) = \mathbf{W}(t)\mathbf{x}^{(n)} \tag{7.4}$$

represent the projection term. The response of L^{II} to stimulus $\mathbf{x}^{(n)}$ at time t can be written

$$\mathbf{y}^{(n)}(t) = \phi(\boldsymbol{\beta}^{(n)} + \boldsymbol{\gamma}^{(n)}(t)) \ . \tag{7.5}$$

The Hebbian change in some synaptic weight, following the presentation of the nth stimulus for some period Δt is proportional to the product of the activations in the input and output neurons. For some weight $W(t + \Delta t)_{ij}$, we increment by some amount of the product of activations $x_i^{(n)} y_j^{(n)}$ and decrement by some amount of its previous value $W(t)_{ij}$. We can therefore use the outer product to write

$$\mathbf{W}(t + \Delta t) = \mathbf{W}(t) + \Delta t \mathrm{diag}(\boldsymbol{\eta})(\mathbf{y}^{(n)}(t)\mathbf{x}^{(n)\top}) - \Delta t \mathrm{diag}(\boldsymbol{\delta})\mathbf{W}(t) \ , \tag{7.6}$$

where at each step, we reduce the weights matrix by some neuron-wise decay rate $\boldsymbol{\delta} \in \mathbb{R}^{S_m}$ and apply the increment with some neuron-wise learning rate $\boldsymbol{\eta} \in \mathbb{R}^{S_m}$. This rule allows for the memory to make associative observations over the sequence of input stimuli, capturing salient information. By virtue of the bias term (not present in traditional Hebbian formulations) we can initialise the weights to zero ($\mathbf{W}(0) = \mathbf{0}_{S_m, S_m}$) and still learn a meaningful representation. This should enable more effective learning over short sequences of stimuli which are each presented only once since we do not introduce a 'false memory'. Note that if we remove the projection term and the activation function ϕ and set $\boldsymbol{\theta} = \mathbf{1}_{S_m}$ we obtain the term used in the fast weights model of Ba et al. (2016). Furthermore, it can be seen that our model is a special case of the differentiable plasticity architecture (Miconi et al., 2018), where the matrix of error-corrected weights is constrained to be diagonal and the activation functions are bounded rectifiers.

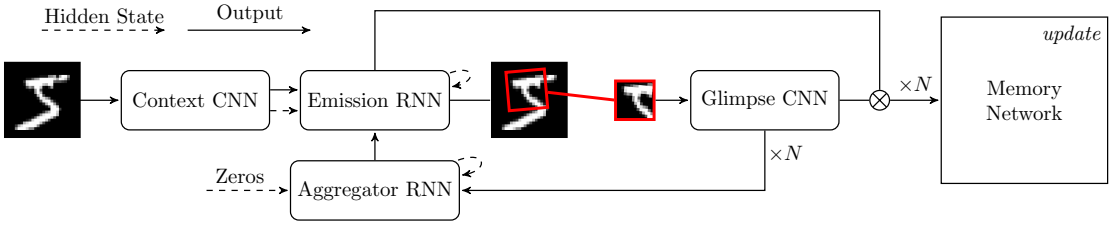


FIGURE 7.2: The attentional working memory model which produces an *update* to the memory. See Section 7.1.2 for a description of the structure and function of each module.

One central innovation of our approach lies in our specific treatment of the plastic network as a memory mechanism. For this setting, we consider the response of the second layer to some query stimulus $\mathbf{x}^{(q)}$, not present in the conditioning sequence. Query stimuli do not trigger a plastic update and are not affected by the bias term, instead they give rise to latent representations conditioned on the plastic weights. Specifically, the representation derived from a query stimulus is given by $\boldsymbol{\gamma}^{(q)}(t) = \mathbf{W}(t)\mathbf{x}^{(q)}$, where we omit the bias term so that the projection is not dependant on the specific query used, only on the content of the memory.

Defining $m_{nq} = \mathbf{x}^{(n)\top}\mathbf{x}^{(q)}$, a fuzzification of the number of cells in L^{I} which are active for both stimuli n and q, we can show that $\boldsymbol{\gamma}^{(q)}(t)$ depends on the m_{nq} of the conditioning stimuli. From the definition of $\boldsymbol{\gamma}^{(q)}$ we have

$$\boldsymbol{\gamma}^{(q)}(t_0 + \Delta t) - \boldsymbol{\gamma}^{(q)}(t_0) = (\mathbf{W}(t_0 + \Delta t) - \mathbf{W}(t_0))\mathbf{x}^{(q)} \,, \tag{7.7}$$

substituting in Equation 7.6 we get

$$\begin{aligned}
\boldsymbol{\gamma}^{(q)}(t_0 + \Delta t) - \boldsymbol{\gamma}^{(q)}(t_0) &= \Delta t \left[\mathrm{diag}(\boldsymbol{\eta})(\mathbf{y}^{(n)}(t_0)\mathbf{x}^{(n)\top})\mathbf{x}^{(q)} - \mathrm{diag}(\boldsymbol{\delta})\mathbf{W}(t_0)\mathbf{x}^{(q)}\right] \\
&= \Delta t \left[\mathrm{diag}(\boldsymbol{\eta})\mathbf{y}^{(n)}(t_0)(\mathbf{x}^{(n)\top}\mathbf{x}^{(q)}) - \mathrm{diag}(\boldsymbol{\delta})\mathbf{W}(t_0)\mathbf{x}^{(q)}\right] \\
&= \Delta t \left[\mathrm{diag}(\boldsymbol{\eta})\mathbf{y}^{(n)}(t_0)m_{nq} - \mathrm{diag}(\boldsymbol{\delta})\boldsymbol{\gamma}^{(q)}(t_0)\right] \,.
\end{aligned} \tag{7.8}$$

This is the key property of associative learning, that the response of the network is dependent on the association between the current input (in our case a query) and the previous inputs (the glimpse features). For our model, this suggests that different query vectors can be used to obtain draw out context dependent features from the series of glimpses.

7.1.2 The Short Term Attentive Working Memory Model (STAWM)

Here we describe in detail the STAWM model shown in Figure 7.2. This is an attention model that allows for a sequence of sub-images to be extracted from the input so that we can iteratively learn a memory representation from a single image. STAWM is

based on the Deep Recurrent Attention Model (DRAM) and uses components from Spatial Transformer Networks (STNs) and the Enriched DRAM (EDRAM) (Ba et al., 2014; Jaderberg et al., 2015; Ablavatski et al., 2017). The design is intended to extend the previous models of visual attention to incorporate our suggested working memory based on a visual sketchpad, which starts with a Hebbian network.

At the core of the model, a two layer RNN defines an attention policy over the input image. As with EDRAM, each glimpse is parameterised by an affine matrix, $\mathrm{A} \in \mathbb{R}^{3\times 2}$, which is sampled from the output of the RNN. At each step, A is used to construct a flow field that is interpolated over the image to obtain a fixed size glimpse in a process denoted as the glimpse transform

$$t_{\mathrm{A}} : \mathbb{R}^{H_i \times W_i} \rightarrow \mathbb{R}^{H_g \times W_g} \ , \tag{7.9}$$

where $H_g \times W_g$ and $H_i \times W_i$ are the sizes of the glimpse and image respectively. Typically the glimpse is a square of size S_g such that $H_g = W_g = S_g$ and we can write the glimpse sequence as

$$G = (\mathbf{g}^{(n)})_{n=0}^{N_g}, \mathbf{g}^{(n)} \in \mathbb{R}^{S_g \times S_g} \ . \tag{7.10}$$

Features obtained from the glimpse are then combined with the location features and used to update the memory with Equation 7.6. Each glimpse is used to update the memory only once, such that $N_g = N_s$. This differs from the approach in Miconi et al. (2018) where each stimulus is used to update the weights multiple times.

Context and Glimpse CNNs Our model includes two Convolutional Neural Networks (CNNs) for obtaining visual features from the image and from the glimpses (which are essentially smaller images). We refer to these networks as the context and glimpse CNNs. The context CNN is given the full input image and expected to establish the top-down contextual information required when making decisions about where to glimpse. Output from the context CNN is used as the initial input and initial hidden state for the emission RNN to support this behaviour. From each glimpse we extract features with the glimpse CNN that are then collected in the 'Memory Network' and fed back into the attention network to inform the next glimpse.

Aggregator and Emission RNNs The aggregator and emission RNNs, shown in Figure 7.2, formulate the glimpse policy over the input image. The aggregator RNN receives the features from the series of glimpses in its hidden state (initialised to zero when no glimpses have yet been made). The emission RNN takes this aggregate of knowledge about the image and an initial hidden state from the context network to define the glimpse policy over the image.

By initialising the hidden states in this way, we expect the model to learn an attention policy which is motivated by the difference between what has been seen so far and the total available information. We use LSTM units for both networks because of their stable learning dynamics (Hochreiter and Schmidhuber, 1997). We give these layers the same size so that they can be conveniently implemented as a two layer LSTM. We use two fully connected layers to transpose the output down to the six dimensions of A for each glimpse. The last of these layers has the weights initialised to zero and the biases initialised to the affine identity matrix as with spatial transformer networks (Jaderberg et al., 2015).

Memory Network The memory network takes the output from a multiplicative 'what, where' pathway and passes it through our Hebbian layer with weights $\mathbf{W} \in \mathbb{R}^{S_m \times S_m}$, where S_m is the memory size. The 'what' and 'where' pathways are fully connected layers which project the glimpse features ('what') and the RNN output ('where') to the memory size. We then take the elementwise product of these two features to obtain the input to the memory, $\mathbf{x} \in \mathbb{R}^{S_m}$. We can think of the memory network as being in one of three states at any point in time, these are: *update*, *intermediate* and *terminal*. The *update* state of the memory is a dynamic state where any signal which propagates through it will trigger an update to the weights using the rule in Equation 7.6. In STAWM, this update will happen N_g times per image, once for each glimpse stimulus.

For the *intermediate* and *terminal* states, no update is made to the plastic weights for signals that are projected through. In the *intermediate* state we observe the memory at some point during the course of the attention policy. Conversely, in the *terminal* state we observe the fixed, final value for the memory after all N_g glimpses have been made. We can use the *intermediate* or *terminal* states to observe the latent space of our model conditioned on some query vector. That is, at different points during the glimpse sequence, different query vectors can be projected through the memory to obtain different latent representations. For a self-supervised setting we can fix the weights of our model so that the sequence of conditioning stimuli, X, cannot be changed by the optimizer.

7.1.3 Using the Memory

We now have an architecture that can be used to build up or 'learn' a Hebbian memory over a series of glimpses of an image. We use the output from the context CNN as a base that is projected into different queries for the different aims of the network. We do this using linear layers, the biases of which can learn a static representation which is then modulated by the image context. We 'detach' the image context meaning that no gradient can propagate back from the query to the context

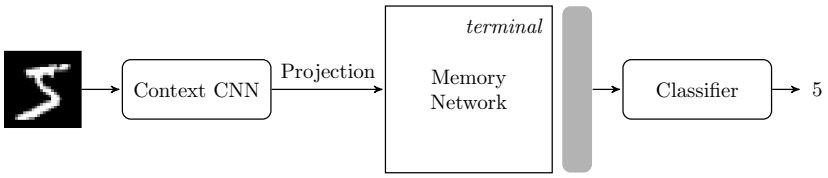


FIGURE 7.3: The classification model which uses the *terminal* state of the memory.

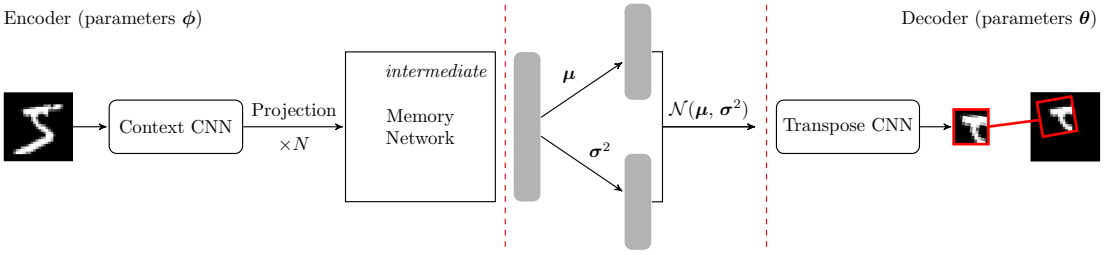


FIGURE 7.4: The drawing model which uses *intermediate* states of the memory.

CNN. This ensures that the context network is not polluted by divergent updates and prevents a trivial case where the glimpses are ignored and the query is used to convey information about the image rather than to address the memory. In this section we will characterise the two network 'heads' which make use of latent vectors (the queries) derived from the memory to perform the tasks of classification and drawing. We will also go on to discuss ways of constraining the sketches in a variational setting to force the model to learn a 'space' of sketches that can be sampled enabling direct assessment of what the model has learned.

Classification For the task of classification, the query vector is projected through the memory in exactly the same fashion as a linear layer to derive a latent vector. This latent representation is then passed through a single classification layer which projects from the memory space down to the number of classes as shown in Figure 7.3. A softmax is applied to the network output and the entire model (including the attenion mechanism) is trained by backpropagation to minimise the categorical cross entropy between the network predictions and the ground truth targets. We therefore expect STAWM to learn a context dependant attention policy that is motivated by the need to classify the images from the series of glimpses.

Learning to Draw To construct a visual sketchpad we propose a novel approach with the auxiliary model depicted in Figure 7.4. The drawing model uses each intermediate state of the memory to query a latent space and compute an update, $\mathbf{U} \in \mathbb{R}^{H_g \times W_g}$, to the canvas, $\mathbf{C} \in \mathbb{R}^{H_i \times W_i}$, that is made after each glimpse. Computing the update or sketch is straightforward, we simply use a transpose convolutional network ('Transpose CNN') with the same structure as the glimpse CNN in reverse. However, as the features were observed under the glimpse transform, t_{A}, we allow the

emission network to further output the parameters, $\mathrm{A}^{-1} \in \mathbb{R}^{3\times 2}$, of an inverse transform

$$t_{\mathrm{A}^{-1}} : \mathbb{R}^{H_g \times W_g} \rightarrow \mathbb{R}^{H_i \times W_i} \tag{7.11}$$

at the same time. The sketch will be warped according to $t_{\mathrm{A}^{-1}}$ before it is added to the canvas. To add the sketch to the canvas there are a few options. We will consider two possibilities here: addition and masking.

The addition method is to simply add each update to the canvas matrix and finally apply a sigmoid after the glimpse sequence has completed to obtain pixel values. This gives an expression for the final canvas

$$\mathbf{C}_{N_g} = \sigma\left(\mathbf{C}_0 + \sum_{n=0}^{N_g} t_{\mathrm{A}^{-1}}(\mathbf{U}_n)\right) \ , \tag{7.12}$$

where $\mathbf{C}_0 \in \{c\}^{H_i \times W_i}$, σ is the sigmoid function and N_g is the total number of glimpses. We can modulate c to alter the base colour of the canvas. Choosing $c = 0$ gives a base colour of grey, which is appropriate for colour settings such as CIFAR-10, whereas $c = -6$ gives a black canvas which is more appropriate for MNIST. The virtue of this method is its simplicity. However, for complex colour reconstructions, overlapping sketches will be required to counteract each other such that they may not be viewable independently of each other.

An alternative approach, the masking method, could help to prevent these issues. Ideally, we would like the additions to the canvas to be as close as possible to painting in real life. In such a case, each brush stroke replaces what previously existed underneath it, which is dramatically different to the effect of the addition method. In order to achieve the desired replacement effect we can allow the model to mask out areas of the canvas before the addition is made. We therefore add an extra channel to the output of the transpose CNN, the alpha channel. This mask, $\mathrm{P} \in \mathbb{R}^{H_g \times W_g}$, is warped by $t_{\mathrm{A}^{-1}}$ as with the rest of the sketch. A sigmoid is applied to the output from the transpose CNN so that the mask contains values close to one where replacement should occur and close to zero elsewhere.

Ideally, the mask values would be precisely zero or one. To achieve this, we could take P as the probabilities of a Bernoulli distribution and then draw

$$\mathrm{B} \sim \mathit{Bern}(t_{\mathrm{A}^{-1}}(\sigma(\mathbf{P})) \ . \tag{7.13}$$

However, the Bernoulli distribution cannot be sampled in a differentiable manner. We will therefore use an approximation, the Gumbel-Softmax (Jang et al., 2016) or Concrete (Maddison et al., 2016) distribution, which can be differentiably sampled using the reparameterization trick. The Concrete distribution is modulated with the

temperature parameter, τ, such that

$$\lim_{\tau \to 0} \mathit{Concrete}(p, \tau) = \mathit{Bern}(p) \ . \tag{7.14}$$

We can then construct the canvas iteratively with the expression

$$\begin{aligned} \mathbf{C}_N =& \mathbf{C}_{N-1} \odot (1 - \mathrm{B}) + t_{\mathrm{A}^{-1}}(\sigma(\mathbf{U}_N)) \odot \mathrm{B} \ , \\ \mathrm{B} \sim& \mathit{Concrete}(t_{\mathrm{A}^{-1}}(\sigma(\mathbf{P})), \tau) \ , \end{aligned} \tag{7.15}$$

where $\mathbf{C}_0 \in \{0\}^{H_i \times W_i}$ and $\odot$ is the elementwise multiplication operator. Note that this is a simplified *over* operator from alpha compositing where we assume that objects already drawn have an alpha value of one (Porter and Duff, 1984).

Learning a Sketch Space Although similar in its objective, our approach differs from the DRAW model from Gregor et al. (2015) in that we have a sketch sequence that is generated independently of the positional information that is jointly required to construct the canvas. An important consequence of this is that it is not possible to use STAWM as a fully generative model since you would need to not only generate a sequence of sketches but also the sequence of correlated sketch positions. We can, however, use STAWM as a generative model in the sketch space, rather than the image space if we constrain the sketches to derive from a variational latent space.

A common approach in reconstructive models is to model the latent space as a distribution whose shape over a mini-batch is constrained using a Kullback-Liebler (KL) divergence against a (typically Gaussian) prior distribution. We employ this variational approach, as shown in Figure 7.4, for two key reasons. The first is to enable sampling of the latent space as discussed. The second is to impose an information bottleneck on each glimpse to encourage representation over multiple glimpses. Using the natural logarithm, the KL term here measures precisely the number of nats required to encode the latent factors over the prior (Cover and Thomas, 2012). By minimising this, we implicitly limit the amount of information that can be conveyed with each sketch. As with Variational Auto-Encoders (VAEs) (Kingma and Welling, 2013), the latent space is modelled as the variance ($\boldsymbol{\sigma}^2$) and mean ($\boldsymbol{\mu}$) of a multivariate Gaussian with K components and diagonal covariance. For input $\mathbf{x} \in \mathbb{R}^{H_i \times W_i}$ and some latent factors $\mathbf{z} \in \mathbb{R}^K$, the β-VAE (Higgins et al., 2016) uses the objective

$$\mathcal{L} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x}) q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x})} \left[\log p_{\boldsymbol{\theta}}(\mathbf{x} \,|\, \mathbf{z}) \right] - \beta \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})} \left[D_{\mathrm{KL}} \left(q_{\boldsymbol{\phi}}(\mathbf{z} \,|\, \mathbf{x}) \,\|\, p(\mathbf{z}) \right) \right] \ , \tag{7.16}$$

where $p_{\mathcal{D}}(\mathbf{x})$ is the underlying probability of observing an $\mathbf{x}$ from the dataset $\mathcal{D}$.

For our model, we do not have a single latent space but a sequence of glimpse sub-spaces

$$Z = (\mathbf{z}^{(n)})_{n=0}^{N_g}, \ \ \mathbf{z}^{(n)} \in \mathbb{R}^K \ . \tag{7.17}$$

We can derive a term for the KL divergence between the posterior and the prior for the joint distribution of the glimpses

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(0)}\dots\mathbf{z}^{(N_g)}\,|\,\mathbf{x})\,\|\,p(\mathbf{z}^{(0)}\dots\mathbf{z}^{(N_g)})\right)\ . \tag{7.18}$$

Assuming that elements of Z are conditionally independent, we have

$$q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x}) = \prod_{n=0}^{N_g} q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x}),\ p(Z) = \prod_{n=0}^{N_g} p(\mathbf{z}^{(n)})\ . \tag{7.19}$$

We can therefore re-write the joint KL term and simplify

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})\right)\|\,p(Z)) = \mathbb{E}_{q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})}{p(Z)}\right]\ , \tag{7.20}$$

$$= \mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(0)}\,|\,\mathbf{x})}\dots\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(N_g)}\,|\,\mathbf{x})}\left[\log\prod_{n=0}^{N_g}\frac{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}{p(\mathbf{z}^{(n)})}\right]\ , \tag{7.21}$$

$$= \sum_{n=0}^{N_g}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(0)}\,|\,\mathbf{x})}\dots\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(N_g)}\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}{p(\mathbf{z}^{(n)})}\right]\ . \tag{7.22}$$

Since $\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(i)}\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}{p(\mathbf{z}^{(n)})}\right] = \log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}{p(\mathbf{z}^{(n)})}$, $\forall\ \mathbf{z}^{(i)}\in Z\setminus\{\mathbf{z}^{(n)}\}$, we have

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})\right)\|\,p(Z)) = \sum_{n=0}^{N_g}\mathbb{E}_{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}\left[\log\frac{q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})}{p(\mathbf{z}^{(n)})}\right]\ , \tag{7.23}$$

$$= \sum_{n=0}^{N_g} D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})\,\|\,p(\mathbf{z}^{(n)})\right)\ . \tag{7.24}$$

Finally, by the linearity of expectation, we can write

$$\mathcal{L} = \mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})}\left[\log p_{\boldsymbol{\theta}}(\mathbf{x}\,|\,Z)\right] - \beta\sum_{n=0}^{N_g}\mathbb{E}_{p_{\mathcal{D}}(\mathbf{x})}\left[D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}^{(n)}\,|\,\mathbf{x})\,\|\,p(\mathbf{z}^{(n)})\right)\right]\ . \tag{7.25}$$

For our experiments we set the prior to be an isotropic unit Gaussian ($p(\mathbf{z}) = \mathcal{N}(\mathbf{0},\mathrm{I})$) yielding the KL term

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(\mathbf{z}\,|\,\mathbf{x})\,\|\,\mathcal{N}(\mathbf{0},\mathrm{I})\right) = -\frac{1}{2}\left(1+\log\boldsymbol{\sigma}^2-\boldsymbol{\mu}^2-\boldsymbol{\sigma}^2\right)\ , \tag{7.26}$$

and thus the joint KL term

$$D_{\mathrm{KL}}\left(q_{\boldsymbol{\phi}}(Z\,|\,\mathbf{x})\,\|\,\mathcal{N}(\mathbf{0},\mathrm{I})\right) = -\frac{1}{2}\left(N_g+\sum_{n=0}^{N_g}\left[\log\boldsymbol{\sigma}_{(n)}^2-\boldsymbol{\mu}_{(n)}^2-\boldsymbol{\sigma}_{(n)}^2\right]\right)\ , \tag{7.27}$$

where $(\boldsymbol{\mu}_{(n)},\boldsymbol{\sigma}_{(n)}) = \mathbf{z}^{(n)}$.

TABLE 7.1: Supervised classification performance of our model on the MNIST dataset. Mean and standard deviation reported from 5 trials.

Model	Error
RAM, $S_g = 8$, $N_g = 5$	1.34%
RAM, $S_g = 8$, $N_g = 6$	1.12%
RAM, $S_g = 8$, $N_g = 7$	1.07%
STAWM, $S_g = 8$, $N_g = 8$	$0.41\%_{\pm 0.03}$
STAWM, $S_g = 28$, $N_g = 10$	$0.35\%_{\pm 0.02}$

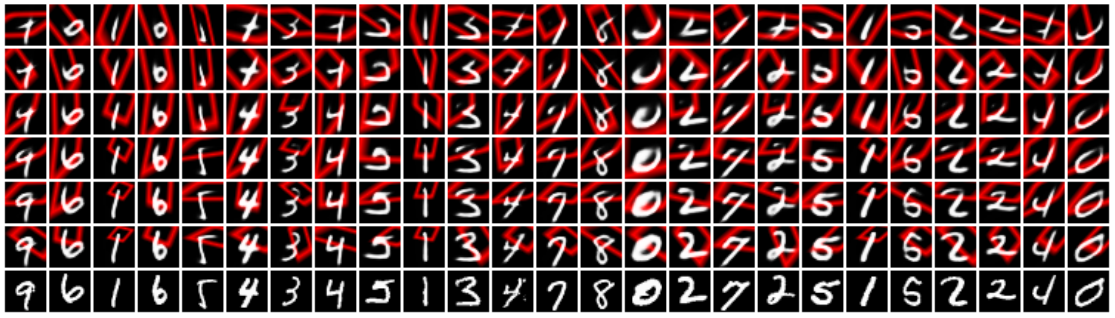
7.2 Experiments

In this section we discuss the results that have been obtained using the STAWM model. Our code is implemented in PyTorch (Paszke et al., 2017) with torchbearer (Harris et al., 2018).

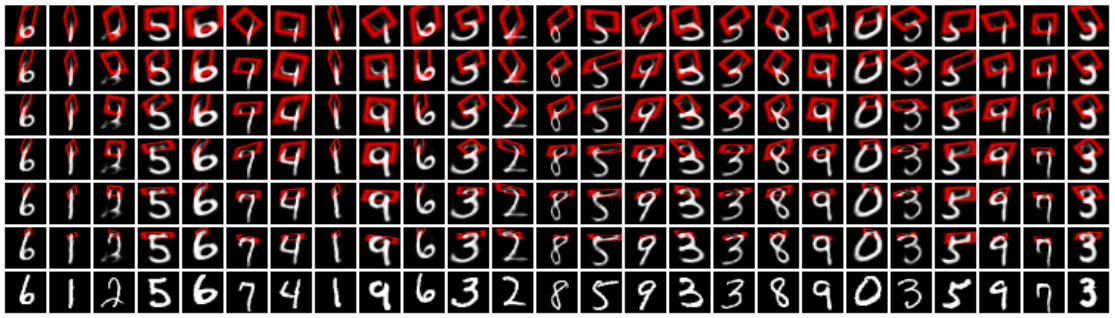
7.2.1 Classification

We first perform classification experiments on handwritten digits from the MNIST dataset (LeCun, 1998) using the model in Figure 7.3. We perform some experiments with $S_g = 8$ in order to be comparable to previous results in the literature. We also perform experiments with $S_g = 28$ where each glimpse can communicate the whole image subjected to a different transform. The MNIST results are reported in Table 7.1 and show that STAWM obtains superior classification performance on MNIST compared to the RAM model. It can also be seen that the over-complete strategy obtains performance that is competitive with the state of the art of around 0.25% for a single model (Sabour et al., 2017), with the best STAWM model from the 5 trials obtaining a test error of 0.31%. This suggests an alternative view of visual attention as enabling the model to learn a more powerful representation of the image. That is, by subjecting the image to series of affine distortions and subsequently observing salient features, the model can learn a more robust representation.

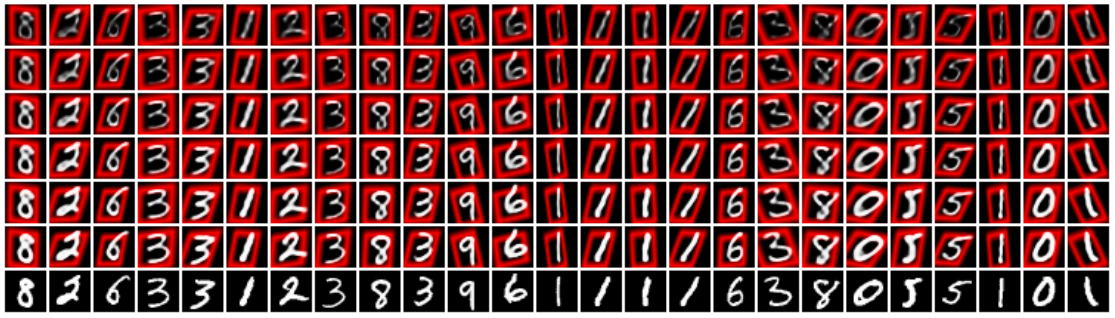
We additionally experimented with classification on CIFAR-10 (Krizhevsky, 2009) but found that the choice of glimpse CNN was the dominating factor in performance, rather than the number of glimpses or size of the memory. For example, using MobileNetV2 (Sandler et al., 2018) as the glimpse CNN we obtained a single run accuracy of 93.05%, whereas performance with the three layer CNN used for the MNIST experiments did not exceed 80%.



(A) $S_g = 4$, line drawing



(B) $S_g = 6$, parts based



(C) $S_g = 8$, compression

FIGURE 7.5: Canvas updates for the drawing model on MNIST with $S_g = 8$, $S_g = 6$ and $S_g = 4$, $N_g = 12$. The first 6 rows give the drawing sequence (we omit the first 6 steps for brevity) and the bottom row shows the target image. Best viewed in colour.

7.2.2 Drawing - Addition

For our second experiment, we report results using our model as an autoencoder trained to reconstruct the input. Some example reconstruction sketch sequences for models with different glimpse sizes are given in Figure 7.5. The figure shows that there are three types of drawing policy that can be learned using the addition method. First, the model can simply compress the image into a square equal to the glimpse size and decompress later. Second, the model can learn to trace lines such that all of the notion of object is contained in the pose information instead of the features. Finally, the model can learn a pose invariant, parts based representation.

The most significant control of the type of sketching policy learned is the glimpse size. At $S_g = 8$ or above, enough information can be stored to make a simple compression

FIGURE 7.6: CIFAR-10 reconstructions for the baseline β-VAE and the drawing model with $S_g = 16$, $N_g = 8$. Top: baseline results. Middle: drawn results. Bottom: target images. Best viewed in colour.

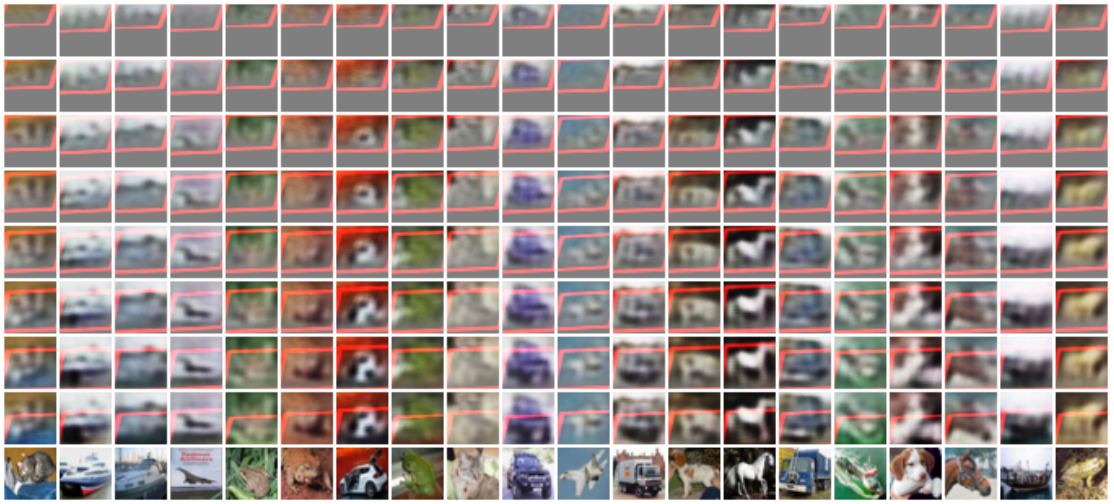


FIGURE 7.7: Canvas updates for the drawing model on CIFAR-10 with $S_g = 16$, $N_g = 8$. The first 8 rows give the drawing sequence and the bottom row shows the target image. Best viewed in colour.

the most successful option. Conversely, at $S_g = 4$ or below, the model is forced to simply draw lines. When $S_g = 6$ with $N_g = 12$ we obtain an appropriate balance between the two extreme and the model learns to reconstruct the image as a sequence of object parts. In this way the model has learned an implicit notion of class since each deviate of a particular character is drawn with the same sequence of parts, with the sketch transform being use to cater the parts to the individual example.

We also report results painting images from CIFAR-10. To establish a baseline we show reconstructions from a reimplementation of β-VAE (Higgins et al., 2016). To be as fair as possible, our baseline uses the same CNN architecture and latent space size as STAWM. This is still only an indicative comparison as the two models operate in fundamentally different ways. Autoencoding CIFAR-10 is a much harder task due to the large diversity in the training set for relatively few images (Recht et al., 2018; Gregor et al., 2015). However, our model marginally outperforms the baseline with a terminal mean squared error of $0.0083_{\pm 0.0006}$ vs $0.0113_{\pm 0.0001}$ for the VAE.

On inspection of the glimpse sequence given in Figure 7.7 we can see that although STAWM has not learned the kind of parts-based representation we had hoped for, it

has learned to scan the image vertically and produce a series of slices. The reason for this seems clear; any 'edges' in the output where one sketch overlaps another would have values that are scaled away from the target, resulting in a constant pressure for the sketch space to expand. This is not the case in MNIST, where overlapping values will only saturate the sigmoid and still produce valid images.

7.2.3 Drawing - Masking

An approach that may enable the learning of meaningful sketch policies on complex data is the masking method discussed in Section 7.1.3. To test our model in this setting we use the CelebA dataset (Liu et al., 2015), which contains a large number of pictures of celebrity faces. The model is trained with the auto-encoding objective and the addition of a KL divergence for the joint distribution of the glimpse spaces with a Gaussian prior given in Equation 7.27 to encourage the model to learn a space of sketches.

Figure 7.8 shows the learned sketch sequence for images from the CelebA data set. The model begins by drawing backgrounds, then moving on to the hair and ears, before finally drawing the face. The sequence is repeated for every image, with the last sketch consistently painting the face to the canvas.

An advantage of the masking method is that we have an explicit mask for the regions drawn at each step. Since the face region is drawn at a consistent point in the sequence, we can use the mask as an unsupervised segmentation mask. Figure 7.9 shows the result of the drawing model on CelebA along with the target image and the learned mask from the final glimpse elementwise multiplied with the ground truth. Here, STAWM has learned to separate the salient face region from the background without explicit supervision.

As a consequence of the variational objective, we can sample the sketch space and treat the decoder as a generative model. This allows us to produce new images from the specific part of the space which contains faces, as shown in Figure 7.10. These imaginary faces are pose-normalised since the transform parameters are given independently of the latent factors by the attention policy. The faces are cleanly segmented since the alpha mask is included with the sketch. This allows us to multiply by the mask as is usually done when producing sketches for reconstructions. That is, we show only the pixels which have a high probability of being in the output, given by the mask probabilities $\sigma(\mathbf{P})$.

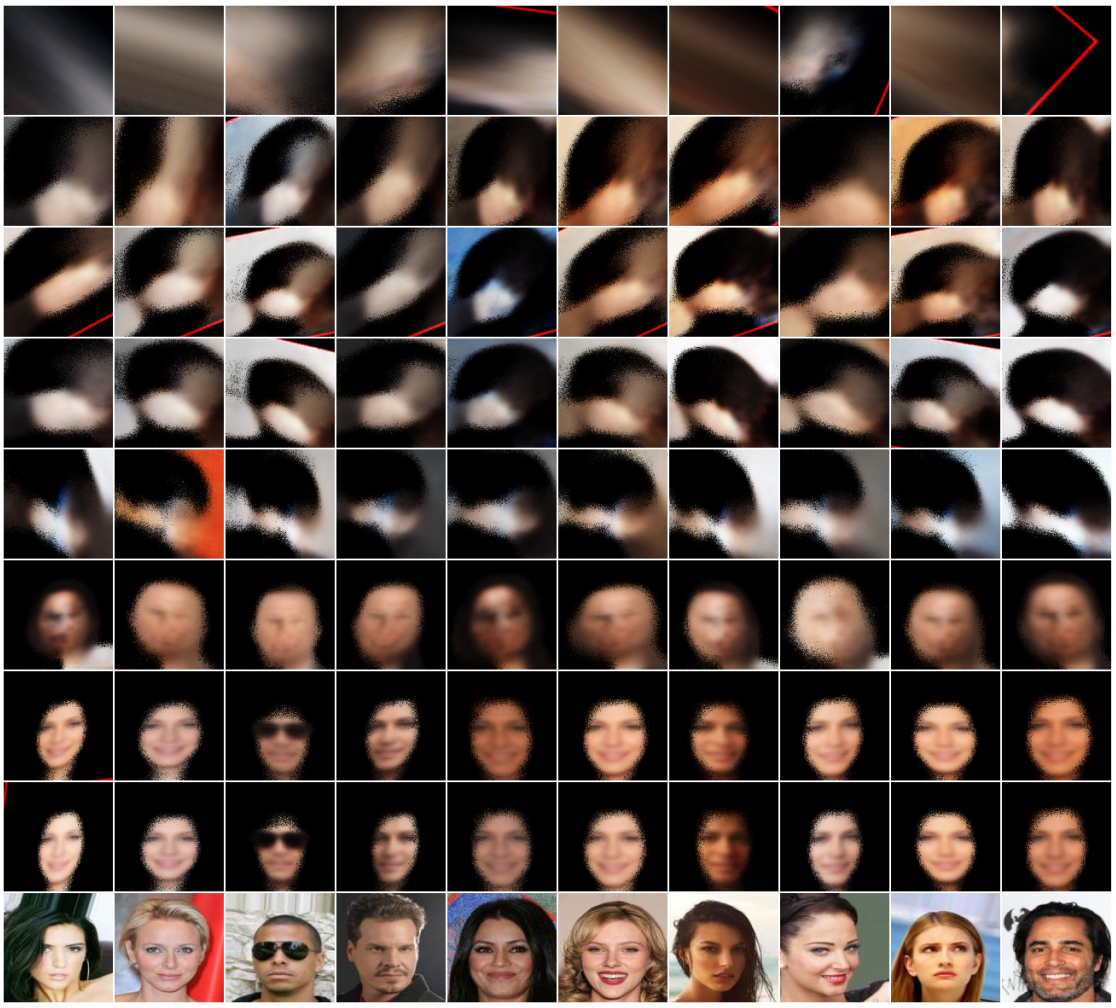


FIGURE 7.8: Canvas updates for the drawing model on CelebA with $S_g = 64$, $N_g = 8$. The first 8 rows give the drawing sequence and the bottom row shows the target image. Best viewed in colour.

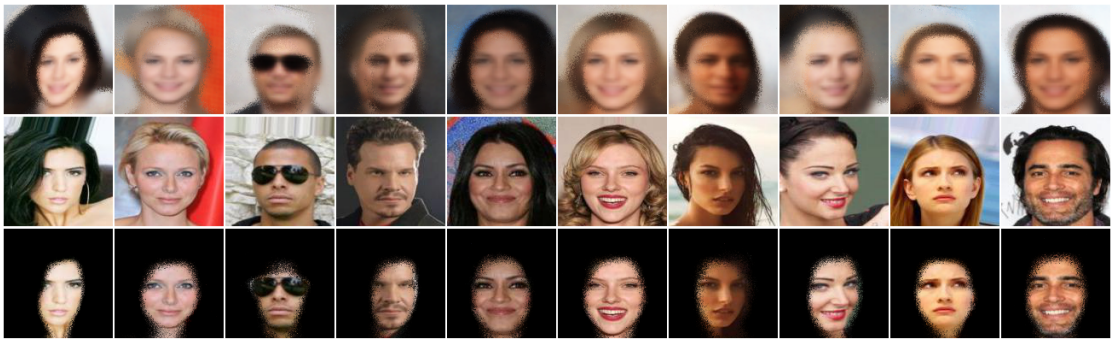


FIGURE 7.9: Output from the drawing model for CelebA with $S_g = 64$, $N_g = 8$. Top: the drawn results. Middle: the target images. Bottom: the learned mask from the last glimpse, pointwise multiplied with the target image. Best viewed in colour.

FIGURE 7.10: Imaginary faces produced by randomly sampling the latent face subspace of the trained CelebA model with $S_g = 64$, $N_g = 8$. Best viewed in colour.

TABLE 7.2: Self-supervised classification performance of our model on the MNIST dataset.

Model	Error
DBM, Dropout (Srivastava et al., 2014)	0.79%
Adversarial (Goodfellow et al., 2014)	0.78%
Virtual Adversarial (Miyato et al., 2015)	0.64%
Ladder (Rasmus et al., 2015)	0.57%
STAWM, $S_g = 6$, $N_g = 12$	0.77%

TABLE 7.3: Self-supervised classification performance of our model on the CIFAR-10 dataset. Mean and standard deviation reported from 5 trials.

Model	Error
Baseline β-VAE	$63.44\%_{\pm 0.31}$
STAWM, $S_g = 16$, $N_g = 8$	$55.40\%_{\pm 0.63}$

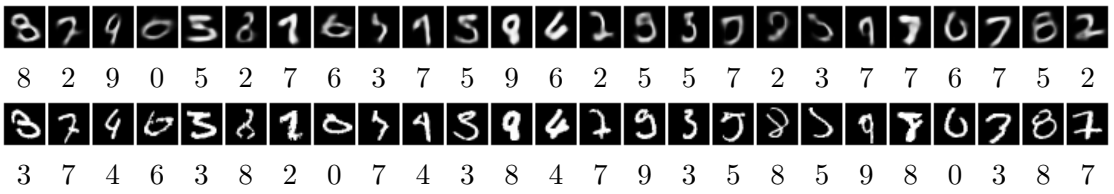


FIGURE 7.11: Top: sketchpad results and associated predictions for a sample of mis-classifications. Bottom: associated input images and target classes.

7.2.4 Self-Supervised Classification

In the previous experiments we have shown that a parts-based representation can be induced in the auto-encoding setting through control of the glimpse size. One way to view this representation is that it separates object from pose. To validate this, we now experiment with classifying from the memory representation with frozen weights pre-trained on the auto-encoding task. In this case, the only learnable parameters are the weights of the two linear layers of the classification head.

We first report results on MNIST in Table 7.2. The table shows that classification performance with the parts-based representation is competitive with contemporary self-supervised approaches. We additionally report self-supervised results on CIFAR-10 in Table 7.3. Since the sketching model did not learn a parts-based representation of CIFAR-10, we would not naturally expect competitive self-supervised performance. Indeed, the results show that, although providing a small improvement over the baseline VAE, the performance is not competitive.

7.2.5 Interpretable Classification

One of the interesting properties of the STAWM model is the ability to project different things through the memory to obtain different views on the hidden state. We can therefore use both the drawing network and the classification network in tandem by using separate linear layers to obtain independent queries and simply summing the reconstruction and classification losses. By deriving the classification and the sketch from the same multi-modal representation, it is hoped that the sketch will provide insight on any mis-classifications by the model.

For this experiment, with $S_g = 8$, the terminal classification error for the model is 1.0%. We show the terminal state of the canvas for a sample of mis-classifications in Figure 7.11. Here, the drawing gives an interesting reflection of what the model 'sees' and in many cases the drawn result looks closer to the predicted class than to the target. For example, in the rightmost image, the model has drawn and predicted a '2' despite attending to a '7'. More generally, it can be seen that the model is over-sensitive to the peak at the top of a '0', often confusing it for a '6'. The opposite

is also true, where a '6' with a suppressed crest is mistakenly drawn and classified as a '0'. Yet another example is given by the '8' whose circles have collapsed to lines causing it to instead be drawn and classified as a '7'.

7.3 Summary

In this chapter we have extended deep models of visual attention to include a Hebbian short-term 'memory' mechanism and the ability to sketch their input. We suggest that this sketch combined with the short-term memory constitutes a working memory model. This model learns a representational hierarchy with the hidden state of the memory giving rise to a series of sketches for which we have joint pose information that is then integrated in the canvas. Under the right conditions, and through careful choices of hyper-parameters, the model can learn a parts-based representation that is valuable as a self-supervised feature for classification.

Regarding sketch policies, we have shown that a model equipped with a masking mechanism can learn to sketch human faces. In particular, the model learns to employ a consistent sketching policy where the final sketch represents the face. As a result, we are able to reliably obtain a mask the can be used to segment the faces in the images. This unsupervised segmentation behaviour is not induced in the model through a particular loss or construction, and is instead an emergent property of the system.

In a final experiment, we have shown that the Hebbian mechanism can act as a multi-modal representation to jointly address both classification and reconstruction objectives. In this setting, the model produces sketches that accord with its predictions. As such, we are able to interpret mis-classifications by the model through analysis of the corresponding sketch.

Chapter 8

Conclusions and Future Work

In this thesis we have taken inspiration from biology in order to study deep networks. In one vein, we have used biological analyses as the basis for new approaches to the interpretation of the function learned by trained models. Tangentially, we have considered novel architectures inspired by human anatomy and psychology to expand the generalisation capabilities of modern deep learning architectures. Combined these efforts provide a clear mandate for further exploration of ideas built on the foundations of biological inspiration. We now briefly summarise the key findings of this work, before detailing our suggested directions for future work.

Chapter 3 introduced a method for determining whether CNNs use colour when classifying images. We first reviewed and re-implemented approaches to the generation of high-quality super-stimuli. Next, we introduced the notion of a trait group; a set of classes that are all uniquely defined by a particular trait. We defined the'fruits' and 'animals' trait groups that are well characterised by their colour and by a combination of colour and pattern / texture respectively. Our results presented super-stimuli for these classes for a range of CNN architectures. Analysing these results, we made a few observations First, we noted that CNNs generally do make use of colour, although this can change dramatically between architectures. Second, we observed that the characteristic shapes of the classes were, in many cases, recovered by the super-stimuli, suggesting that CNNs may not be as biased against shape as previously thought.

In Chapter 4 we introduced techniques for the interpretation of convolutional networks in terms of how they process colour. We showed that colour and spatial opponency of convolutional cells can be characterised using long established techniques from biology. We further showed that the emergence of opponent representations mirrors what is found in nature. We subsequently demonstrated that the characteristics of this colour processing are affected by the chosen architecture of the convolutional network. Through a comprehensive set of studies, we detailed this connection and shed light on the causal relationship between network architecture and the emergence of opponency.

Chapter 5 built on the findings of Chapter 4 to study the impact of opponency on performant models trained on the large scale ImageNet dataset. In particular, Chapter 5 questioned whether an increase in the percentage of opponent cells gives a corresponding increase in shape bias or adversarial robustness. We additionally expanded on the analyses introduced in Chapter 4 to include a notion of luminance opponency. Our results showed that an increase in opponency does not have a significant impact on adversarial robustness. Although this was a negative result, it demonstrated the limitations of an approach based purely on distributions of opponent cells. Although such analyses are valuable for characterising the nature of learned functions, they are not necessarily good indicators for more complex metrics such as generalisation performance.

In Chapter 6 we introduced a convolutional model of foveation. In our first experiment, we studied a failure of spatial transformer networks to localise their input. We showed that the foveated convolution improves localisation performance effectively solving the observed localisation problem. In our second experiment, we analysed the receptive fields of cells in the foveated convolution layer of networks tasked with both localization and classification of the input. Our analyses showed that cells in the foveated networks are able to modulate the size of their receptive field based in the input in order to support both tasks with a single representation. This mirrors similar findings of multi-modal cells in macaque inferior temporal cortex. For our final experiment, we extended the foveated convolution to allow for multiple layers of foveation in order to study how opponency varies with eccentricity. Our results showed a reduction in opponency with eccentricity, consistent with observations from foveated trichromatic vision in nature. We further developed a method for establishing the stimulus preference (Red-Green or Blue-Yellow) of cells. Analysis of these results showed a reduction in Red-Green opponency in the periphery. This again accords with observations from trichromatic primates. By comparison with results from locally connected networks (which allow spatially variant function without a scale-space) we showed that the characteristic distribution of Red-Green preference is induced by spatially variant function whereas the peripheral reduction in opponency requires a scale-space.

With the objective of exploring mechanisms for visual memory in deep networks, Chapter 7 introduced a visual attention network equipped with a visual memory capable of reconstructing a scene by producing a series of sketches. Through a series of experiments, we demonstrated various emergent properties of this model. First, we showed that by controlling the glimpse size, we can control the drawing mode; transitioning from line drawing to parts-based reconstruction and finally iterative refinement. Next, we demonstrated that the model learns a repeatable sketching procedure when trained on images of faces. The procedure is sufficiently repeatable that we were able to use the alpha masks from the ‘face’ sketch to extract the faces

from the images. Finally, we showed that a sketching model trained in tandem with a classification objective can learn to draw what it sees, giving the ability to visually asses and interpret misclassifications. These experiments demonstrate the potential for emergent phenomena in the largely untapped realm of deep network architectures inspired by nature.

Combined, our efforts have encompassed two themes: interpretation and generalisation. Regarding interpretation, we have presented clear evidence for the potential of interpretation techniques based on approaches from biology. Regarding generalisation, we have seen the value of biological inspiration for expanding the horizons of deep networks, imbuing them with enhanced capabilities, and giving rise to complex emergent phenomena.

8.1 Directions for Future Work

Naturally, this thesis leaves a number of open questions and potential directions for future study. In this section, we present examples of such directions and provide a recommendation for future work.

8.1.1 Transferring Opponent Representations

Following the findings of Chapter 4, we suggest that opponent cells may be of greater utility in transfer learning. Transfer learning is the practice of using some or all of a pre-trained network as a fixed feature extractor for a new task. Our contention is that opponent representations constitute a more general form of visual feature than those typically learned by a CNN. On this basis we suppose that opponent representations may be more transferable to novel settings. Specifically, one can envisage a scenario where the pre-bottleneck weights are fixed, and the post-bottleneck weights are updated to fit a new data set.

8.1.2 Cascading Opponent Representations

The finding from Chapter 4 that the penultimate layer exhibits a spike in opponency may provide insight to the efficacy of layer-wise training procedures such as deep cascade learning (Marquez et al., 2018). These approaches progressively increase the depth of the network freezing all but the last convolutional layer each time. Based on the evidence presented in this work that the number of opponent cells is related to the distance to the output layer, one might speculate that cascade learning increases the number of opponent cells. That said, whether the opponent cells in later layers inherit the same properties as opponent cells in earlier layers remains to be determined. Note

that cascade learning has been found by Du et al. (2019) to work well with transfer learning, potentially also tying in to the suggestions in Section 8.1.1.

8.1.3 Quantifying Representations of Shape in Super-Stimuli

Chapter 3 considered super-stimuli as an approach for determining the nature of information used by a network to classify its input. Contrary to our expectations, the super-stimuli effectively recovered a notion of shape regarding each of the classes in our proposed trait groups. However, the absence of a metric to quantify this shape representation makes it difficult to determine how repeatable these findings are and to make meaningful comparisons between different models. In light of this fact, an important direction for future work is to try to quantify the extent of shape, colour, and texture in these super-stimuli or at least in the representations that they exploit. One possible avenue through which such a metric may be constructed is information theory. In particular, mutual information is a compelling concept as a measure of the extent to which one representation reduces uncertainty about another. However, mutual information is notoriously difficult to measure, although recent approaches such as mutual information neural estimation from Belghazi et al. (2018) may improve its viability.

8.1.4 Simulating a Retina at the Front of V1Net

In this work we have provided evidence for the emergence of opponent representations in learning machines. We have not found evidence for improved adversarial robustness as a result of these opponent representations. This contrasts with the findings from Dapello et al. (2020) that fixed Gabor filter representations in the first layer have a dramatic impact on robustness. We suspect that this is because the cells in our networks are learnable floating point functions and as such can be manipulated by minute changes in the input.

To test this theory, we suggest implementing a fixed bank of cone opponent centre-surround filters that mimic the filters learned by our models. This filter bank would then be included at the front of a convolutional network and evaluated for its' impact on adversarial robustness. Furthermore, we would expect this model and the V1Net model from Dapello et al. (2020) to complement each other such that they can be used in conjunction for even further improvements in robustness.

8.1.5 Foveation, Opponency, and a Scale-Space

In Chapter 6 we explored a convolutional model of foveation and showed how the presence of a scale-space representation induces a reduction in peripheral opponency.

This experiment starts to provide insight in to the emergence of opponency, however, further examination is required to unpack it fully. We first propose a control experiment to study the number of opponent cells in a series of models trained with progressively increased scale factors. If the results show an increase in opponency as the scale factor is increased then this experiment will prove that opponency presentation can be controlled by the input scale.

In addition to a control study, we propose further investigation into the underlying reason that an increase in scale corresponds to a reduction in opponency. One possible explanation is that opponent representations are more valuable in object-centric vision where spatial frequencies are higher. To test this theory, we could construct a data set of landscape classes (or any other group of natural image classes that are not object-centric) and assess whether the peripheral reduction in opponency remains.

References

Artsiom Ablavatski, Shijian Lu, and Jianfei Cai. Enriched deep recurrent visual attention model for multiple object recognition. In *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*, pages 971–978. IEEE, 2017.

E. D. Adrian and Rachel Matthews. The action of light on the eye. *The Journal of Physiology*, 65(3):273–298, 1928. . URL https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1928.sp002475.

Emre Akbas and Miguel P Eckstein. Object detection through search with a foveated visual system. *PLoS computational biology*, 13(10):e1005743, 2017.

Jimmy Ba, Volodymyr Mnih, and Koray Kavukcuoglu. Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*, 2014.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems*, pages 4331–4339, 2016.

Alan D Baddeley and Graham Hitch. Working memory. In *Psychology of learning and motivation*, volume 8, pages 47–89. Elsevier, 1974.

H. B. Barlow. Summation and inhibition in the frog's retina. *The Journal of Physiology*, 119(1):69–88, 1953. . URL https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1953.sp004829.

R.E. Bedford and Günter W. Wyszecki. Wavelength discrimination for point sources. *JOSA*, 48(2):129–135, 1958.

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R Devon Hjelm. Mine: mutual information neural estimation. *arXiv preprint arXiv:1801.04062*, 2018.

Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327 – 3338, 1997. ISSN 0042-6989. . URL http://www.sciencedirect.com/science/article/pii/S0042698997001211.

J Bilotta and ISRAEL Abramov. Spatial properties of goldfish ganglion cells. *The Journal of general physiology*, 93(6):1147–1169, 1989.

HD Block. The Perceptron: A Model for Brain Functioning. I. *Reviews of Modern Physics*, 34(1):123, 1962. .

HD Block, BW Knight Jr, and Frank Rosenblatt. Analysis of a Four-Layer Series-Coupled Perceptron. II. *Reviews of Modern Physics*, 34(1):135, 1962. .

Léon Bottou, Corinna Cortes, John S. Denker, Harris Drucker, Isabelle Guyon, Larry D. Jackel, Yann LeCun, Urs A. Müller, Eduard Säckinger, Patrice Y. Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *International conference on pattern recognition*, pages 77–77. IEEE Computer Society Press, IEEE, 1994.

Geoffrey M Boynton. Color vision: How the cortex represents color. *Current Biology*, 12(24):R838 – R840, 2002. ISSN 0960-9822. . URL http://www.sciencedirect.com/science/article/pii/S0960982202013477.

Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, and Alexander S. Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *bioRxiv*, 2017. . URL https://www.biorxiv.org/content/early/2017/10/11/201764.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. Thread: Circuits. *Distill*, 2020. . https://distill.pub/2020/circuits.

Nick Cammarata, Gabriel Goh, Shan Carter, Chelsea Voss, Ludwig Schubert, and Chris Olah. Curve circuits. *Distill*, 2021. . https://distill.pub/2020/circuits/curve-circuits.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Soumya Chatterjee and Edward M Callaway. Parallel colour-opponent pathways to primary visual cortex. *Nature*, 426(6967):668–671, 2003.

Brian Cheung, Eric Weiss, and Bruno A. Olshausen. Emergence of foveal image sampling from learning to attend in visual scenes. In *International Conference on Learning Representations*, 2017.

Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.

Francis Crick. Function of the thalamic reticular complex: the searchlight hypothesis. *Proceedings of the National Academy of Sciences*, 81(14):4586–4590, 1984.

Christine A Curcio and Kimberly A Allen. Topography of ganglion cells in human retina. *Journal of comparative Neurology*, 300(1):5–25, 1990.

Christine A Curcio, Kenneth R Sloan, Robert E Kalina, and Anita E Hendrickson. Human photoreceptor topography. *Journal of comparative neurology*, 292(4): 497–523, 1990.

Dennis M Dacey. The mosaic of midget ganglion cells in the human retina. *Journal of Neuroscience*, 13(12):5334–5355, 1993.

Dennis M Dacey and Michael R Petersen. Dendritic field size and morphology of midget and parasol ganglion cells of the human retina. *Proceedings of the National Academy of sciences*, 89(20):9666–9670, 1992.

Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David D Cox, and James J DiCarlo. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. *BioRxiv preprint*, 2020.

Nigel W. Daw. Goldfish retina: organization for simultaneous color contrast. *Science*, 158(3803):942–944, 1967.

Peter Dayan and Laurence F Abbott. *Theoretical neuroscience*, volume 806. Cambridge, MA: MIT Press, 2001.

Russell L. De Valois, Ch J Smith, AJ Karoly, and ST Kitai. Electrical responses of primate visual system: I. different layers of macaque lateral geniculate nucleus. *Journal of comparative and physiological psychology*, 51(6):662, 1958a.

Russell L. De Valois, C.J. Smith, Stephen T. Kitai, and A.J. Karoly. Response of single cells in monkey lateral geniculate nucleus to monochromatic light. *Science*, 1958b.

Russell L. De Valois, Israel Abramov, and Gerald H. Jacobs. Analysis of response patterns of lgn cells∗. *J. Opt. Soc. Am.*, 56(7):966–977, Jul 1966. . URL http://www.osapublishing.org/abstract.cfm?URI=josa-56-7-966.

Russell L. De Valois, Duane G Albrecht, and Lisa G Thorell. Spatial frequency selectivity of cells in macaque visual cortex. *Vision research*, 22(5):545–559, 1982.

Andrew M Derrington, John Krauskopf, and Peter Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):241–265, 1984.

Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020a.

Arturo Deza and Talia Konkle. Foveation induces robustness to scene occlusion in deep neural networks. *Journal of Vision*, 20(11):442–442, 2020b.

Xin Du, Katayoun Farrahi, and Mahesan Niranjan. Transfer learning across human activities using a cascade neural network architecture. In *Proceedings of the 23rd International Symposium on Wearable Computers*, pages 35–44, 2019.

S. Engel, X. Zhang, and B. Wandell. Colour tuning in human visual cortex measured with functional magnetic resonance imaging. *Nature*, 388(6637):68–71, 1997. . URL https://www.scopus.com/inward/record.uri?eid=2-s2.0-0030744384&doi=10.1038%2f40398&partnerID=40&md5=d42786f821eb1652af0ad297996452d6. cited By 266.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, and Pascal Vincent. Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1, 2009.

O Estévez. On the fundamental data-base of normal and dichromatic colour vision. *Ph. D. Thesis, University of Amsterdam*, 1981.

William T Freeman and Michal Roth. Orientation histograms for hand gesture recognition. In *International workshop on automatic face and gesture recognition*, volume 12, pages 296–301, 1995.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bygh9j09KX.

Masoud Ghodrati, Seyed-Mahdi Khaligh-Razavi, and Sidney R Lehky. Towards building a more complex view of the lateral geniculate nucleus: recent advances in understanding its role. *Progress in Neurobiology*, 156:214–255, 2017.

James J Gibson. *The perception of the visual world.* Houghton Mifflin, 1950.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

Johann Wolfgang von Goethe. *Theory of colours*, volume 3. MIT Press, 1840.

Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, and Marcelo Bertalmio. Convolutional Neural Networks Can Be Deceived by Visual Illusions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Melvyn A Goodale and A David Milner. Separate visual pathways for perception and action. *Trends in neurosciences*, 15(1):20–25, 1992.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL http://arxiv.org/abs/1412.6572.

Ian J Goodfellow, Yaroslav Bulatov, Julian Ibarz, Sacha Arnoud, and Vinay Shet. Multi-digit number recognition from street view imagery using deep convolutional neural networks. *arXiv preprint arXiv:1312.6082*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Jimenez Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. *arXiv preprint arXiv:1502.04623*, 2015.

Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

Thorsten Hansen, Lars Pracejus, and Karl R Gegenfurtner. Color perception in the intermediate periphery of the visual field. *Journal of vision*, 9(4):26–26, 2009.

Ethan Harris, Matthew Painter, and Jonathon Hare. Torchbearer: A model fitting library for pytorch. *arXiv preprint arXiv:1809.03363*, 2018.

Ethan Harris, Daniela Mihai, and Jonathon Hare. Spatial and Colour Opponency in Anatomically Constrained Deep Networks. In *Workshop on Shared Visual Representations in Humans and Machines (SVRHM)*, 2019a. URL https://arxiv.org/abs/1910.11086.

Ethan Harris, Mahesan Niranjan, and Jonathon Hare. Foveated Convolutions: Improving Spatial Transformer Networks by Modelling the Retina. In *Workshop on Shared Visual Representations in Humans and Machines (SVRHM)*, 2019b.

Ethan Harris, Daniela Mihai, and Jonathon Hare. Anatomically Constrained ResNets Exhibit Opponent Receptive Fields; So What? In *Workshop on Shared Visual Representations in Humans and Machines (SVRHM)*, 2020a.

Ethan Harris, Daniela Mihai, and Jonathon Hare. How Convolutional Neural Network Architecture Biases Learned Opponency and Colour Tuning. *Neural Computation*, 2020b. URL https://arxiv.org/abs/2010.02634.

H. K. Hartline. The response of single optic nerve fibers of the vertebrate eye to illumination of the retina. *American Journal of Physiology-Legacy Content*, 121(2): 400–415, 1938. .

H. K. Hartline. The nerve messages in the fibers of the visual pathway∗. *J. Opt. Soc. Am.*, 30(6):239–247, Jun 1940. . URL http://www.osapublishing.org/abstract.cfm?URI=josa-30-6-239.

H. Keffer Hartline, Henry G. Wagner, and E. F. Macnichol. The peripheral origin of nervous activity in the visual system. *Cold Spring Harbor symposia on quantitative biology*, 17:125–41, 1952.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. New York: Wiley, 1949.

Richard Held and Alan Hein. Movement-produced stimulation in the development of visually guided behavior. *Journal of comparative and physiological psychology*, 56(5): 872, 1963.

Hermann von Helmholtz. LXXXI. On the theory of compound colours. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(28): 519–534, 1852.

Olivier J Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, SM Eslami, and Aaron van den Oord. Data-efficient image recognition with contrastive predictive coding. *arXiv preprint arXiv:1905.09272*, 2019.

Stewart HC Hendry and R Clay Reid. The koniocellular pathway in primate vision. *Annual review of neuroscience*, 23(1):127–153, 2000.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *arXiv preprint arXiv:1907.07174*, 2019.

Ewald Hering. *Grundzüge der Lehre vom Lichtsinn*. Springer, 1920.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011. .

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Q Weinberger. Deep networks with stochastic depth. In *European conference on computer vision*, pages 646–661. Springer, Springer International Publishing, 2016.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1): 106–154, 1962.

David H Hubel and Torsten N Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

David H. Hubel and Torsten N. Wiesel. *Brain and visual perception: the story of a 25-year collaboration*. Oxford University Press, 2004.

RWG Hunt and MR Pointer. A colour-appearance transform for the cie 1931 standard colorimetric observer. *Color Research & Application*, 10(3):165–179, 1985.

Intel. Intel scene classification challenge, 2018. URL https://datahack.analyticsvidhya.com/contest/practice-problem-intel-scene-classification-challe/.

Gerald H Jacobs. Single cells in squirrel monkey lateral geniculate nucleus with broad spectral sensitivity. *Vision Research*, 4(3-4):221–IN3, 1964.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Elizabeth N. Johnson, Michael J. Hawken, and Robert Shapley. The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nature neuroscience*, 4(4):409, 2001.

Elizabeth N. Johnson, Michael J. Hawken, and Robert Shapley. The orientation selectivity of color-responsive neurons in macaque v1. *Journal of Neuroscience*, 28 (32):8096–8106, 2008.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Andreas Kleinschmidt, Barry B Lee, Martin Requardt, and Jens Frahm. Functional mapping of color processing by magnetic resonance imaging of responses to selective p-and m-pathway stimulation. *Experimental Brain Research*, 110(2):279–288, 1996.

Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 43(1):59–69, 1982.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, volume 25, pages 1097–1105, 2012.

Stephen W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *ICLR Workshop*, 2017. URL https://arxiv.org/abs/1607.02533.

Hugo Larochelle and Geoffrey E Hinton. Learning to combine foveal glimpses with a third-order Boltzmann machine. In *Advances in neural information processing systems*, pages 1243–1251, 2010.

Quoc V Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y Ng. Ica with reconstruction cost for efficient overcomplete feature learning. In *Advances in neural information processing systems*, volume 24, pages 1017–1025, 2011.

Yann Le Cun, Ofer Matan, Bernhard Boser, John S. Denker, Don Henderson, Richard E. Howard, Wayne Hubbard, L.D. Jackel, and Henry S. Baird. Handwritten zip code recognition with multilayer networks. In *Proc. 10th International Conference on Pattern Recognition*, volume 2, pages 35–40. IEEE, 1990.

Yann Le Cun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Sidney R Lehky and Terrence J Sejnowski. Network model of shape-from-shading: Neural function arises from both receptive and projective fields. *Nature*, 333(6172): 452–454, 1988.

Sidney R Lehky and Terrence J Sejnowski. Neural model of stereoacuity and depth interpolation based on a distributed representation of stereo disparity. *Journal of Neuroscience*, 10(7):2281–2299, 1990.

Sidney R Lehky and Terrence J Sejnowski. Seeing white: Qualia in the context of decoding population codes. *Neural computation*, 11(6):1261–1280, 1999.

Peter Lennie, John Krauskopf, and Gary Sclar. Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience*, 10(2):649–669, 1990.

J. Y. Lettvin, H. R. Maturana, W. S. McCulloch, and W. H. Pitts. What the frog's eye tells the frog's brain. *Proceedings of the IRE*, 47(11):1940–1951, 1959.

WR Levick and LN Thibos. Analysis of orientation bias in cat retina. *The Journal of Physiology*, 329:243, 1982.

Jack Lindsey, Samuel A Ocko, Surya Ganguli, and Stephane Deny. A Unified Theory of Early Visual Representations from Retina to Cortex through Anatomically Constrained Deep CNNs. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=S1xq3oR5tQ.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015.

Margaret S Livingstone and David H Hubel. Anatomy and physiology of a color system in the primate visual cortex. *Journal of Neuroscience*, 4(1):309–356, 1984.

Fuhui Long, Zhiyong Yang, and Dale Purves. Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proceedings of the National Academy of Sciences*, 103(15):6013–6018, 2006.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

Peter W Lucas, Nathaniel J Dominy, Pablo Riba-Hernandez, Kathryn E Stoner, Nayuta Yamashita, Esteban Lorí Calderön, Wanda Petersen-Pereira, Yahaira Rojas-DurÁN, Ruth Salas-Pena, Silvia Solis-Madrigal, et al. Evolution and function of routine trichromatic vision in primates. *Evolution*, 57(11):2636–2643, 2003.

Chris J Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Elian Malkin, Arturo Deza, and Tomaso Poggio. Cuda-optimized real-time rendering of a foveated visual system. *arXiv preprint arXiv:2012.08655*, 2020.

Thomas Mansencal, Michael Mauderer, and Michael Parsons. Colour 0.3.16. URL https://doi.org/10.5281/zenodo.3757045.

Enrique S Marquez, Jonathon S Hare, and Mahesan Niranjan. Deep cascade learning. *IEEE transactions on neural networks and learning systems*, 29(11):5475–5485, 2018.

David Marr. *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press, 1982.

David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 207(1167):187–217, 1980.

Luis M Martinez and Jose-Manuel Alonso. Complex receptive fields in primary visual cortex. *The neuroscientist*, 9(5):317–331, 2003.

Kevin Matzen and Noah Snavely. Bubblenet: Foveated imaging for visual discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1931–1939, 2015.

James Clerk Maxwell. IV. On the theory of compound colours, and the relations of the colours of the spectrum. *Philosophical Transactions of the Royal Society of London*, (150):57–84, 1860.

Kerry McAlonan, James Cavanaugh, and Robert H Wurtz. Guarding the gateway to cortex with attention in visual thalamus. *Nature*, 456(7220):391–394, 2008.

Robert K McConnell. Method of and apparatus for pattern recognition, 1986. US Patent 4,567,610.

Thomas Miconi, Jeff Clune, and Kenneth O Stanley. Differentiable plasticity: training plastic neural networks with backpropagation. *arXiv preprint arXiv:1804.02464*, 2018.

Takeru Miyato, Shin Maeda, Masanori Koyama, Ken Nakae, and Shin Ishii. Distributional Smoothing with Virtual Adversarial Training. *stat*, 1050:2, 2015.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212. IEEE, may 2014. .

S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2574–2582, 2016.

Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. 2015.

Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 3(7):e12, 2018.

Nathan Moroney, Mark D Fairchild, Robert WG Hunt, Changjun Li, M Ronnier Luo, and Todd Newman. The CIECAM02 color appearance model. In *Color and Imaging Conference*, volume 2002, pages 23–27. Society for Imaging Science and Technology, 2002.

Kathy T Mullen. Colour vision as a post-receptoral specialization of the central visual field. *Vision research*, 31(1):119–130, 1991.

Kathy T Mullen and Frederick A A Kingdom. Differential distributions of red-green and blue-yellow cone opponency across the visual field. *Visual neuroscience*, 19(1): 109, 2002.

Kathy T Mullen, Masato Sakurai, and William Chu. Does l/m cone opponency disappear in human periphery? *Perception*, 34(8):951–959, 2005.

IJ Murray, NRA Parry, and DJ McKeefry. Cone opponency in the near peripheral retina. *Visual neuroscience*, 23(3-4):503, 2006.

K.I. Naka and William A.H. Rushton. S-potentials from colour units in the retina of fish (cyprinidae). *The Journal of physiology*, 185(3):536–555, 1966.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

Mark S Nixon and Alberto S Aguado. *Feature extraction & image processing for computer vision*. Academic Press, 2012.

Erkki Oja. Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, 15(3):267–273, 1982.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. Feature visualization. *Distill*, 2(11):e7, 2017.

Chris Olah, Arvind Satyanarayan, Ian Johnson, Shan Carter, Ludwig Schubert, Katherine Ye, and Alexander Mordvintsev. The building blocks of interpretability. *Distill*, 3(3):e10, 2018.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. An overview of early vision in inceptionv1. *Distill*, 5(4):e00024–002, 2020a.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020b.

Bruno A Olshausen and David J Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Jonathan Peirce, Jeremy R Gray, Sol Simpson, Michael MacAskill, Richard Höchenberger, Hiroyuki Sogo, Erik Kastman, and Jonas Kristoffer Lindeløv. Psychopy2: Experiments in behavior made easy. *Behavior research methods*, 51(1): 195–203, 2019.

Thomas Porter and Tom Duff. Compositing digital images. In *ACM Siggraph Computer Graphics*, volume 18, pages 253–259. ACM, 1984.

Ralph W. Pridmore. Theory of corresponding colors as complementary sets. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(5):371–381, 2005.

Ralph W. Pridmore. Complementary colors theory of color vision: Physiology, color mixture, color constancy and color perception. *Color Research & Application*, 36(6): 394–412, 2011.

Dale Purves and R Beau Lotto. *Why we see what we do: An empirical theory of vision.* Sinauer Associates, 2003.

Dale Ed Purves, George J Augustine, David Ed Fitzpatrick, Lawrence C Katz, et al. *Neuroscience.* Sinauer Associates, 1997. .

Ivet Rafegas and Maria Vanrell. Color encoding in biologically-inspired convolutional neural networks. *Vision research*, 151:7–17, 2018.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554, 2015.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL http://arxiv.org/abs/1707.04131.

Jonas Rauber, Roland Zimmermann, Matthias Bethge, and Wieland Brendel. Foolbox native: Fast adversarial attacks to benchmark the robustness of machine learning models in pytorch, tensorflow, and jax. *Journal of Open Source Software*, 5(53):2607, 2020. . URL https://doi.org/10.21105/joss.02607.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do cifar-10 classifiers generalize to cifar-10? *arXiv preprint arXiv:1806.00451*, 2018.

Edmund T Rolls and Gustavo Deco. *Computational neuroscience of vision*. Oxford university press, 2002.

Edmund T Rolls, Nicholas C Aggelopoulos, and Fashan Zheng. The receptive fields of inferior temporal cortex neurons in natural scenes. *Journal of Neuroscience*, 23(1): 339–348, 2003.

Frank Rosenblatt. *Principles of neurodynamics*. Spartan Books, 1962. .

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic Routing Between Capsules. In *Advances in Neural Information Processing Systems*, pages 3857–3867, 2017.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

Denis Schluppeck and Stephen A Engel. Color opponent neurons in v1: a review and model reconciling results from imaging and single-unit recording. *Journal of vision*, 2(6):5–5, 2002.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018.

K. J. Seymour, M. A. Williams, and A. N. Rich. The Representation of Color across the Human Visual Cortex: Distinguishing Chromatic Signals Contributing to Object Form Versus Surface Color. *Cerebral Cortex*, 26(5):1997–2005, 02 2015. ISSN 1047-3211. . URL https://doi.org/10.1093/cercor/bhv021.

Honghao Shan, Lingyun Zhang, and Garrison W. Cottrell. Recursive ica. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1273–1280. MIT Press, 2007. URL http://papers.nips.cc/paper/3018-recursive-ica.pdf.

Robert Shapley and Michael J. Hawken. Color in the Cortex: single- and double-opponent cells. *Vision Research*, 51(7):701 – 717, 2011. ISSN 0042-6989. . URL

http://www.sciencedirect.com/science/article/pii/S0042698911000526. Vision Research 50th Anniversary Issue: Part 1.

Steven K. Shevell and Paul R. Martin. Color opponency: tutorial. *J. Opt. Soc. Am. A*, 34(7):1099–1108, Jul 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

VIVIANNE C Smith, BB Lee, JOEL Pokorny, PR Martin, and A Valberg. Responses of macaque ganglion cells to the relative phase of heterochromatically modulated lights. *The Journal of Physiology*, 458(1):191–221, 1992.

Søren Kaae Sønderby, Casper Kaae Sønderby, Lars Maaløe, and Ole Winther. Recurrent spatial transformer networks. *arXiv preprint arXiv:1509.05329*, 2015.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

Kenneth O Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines*, 8(2):131–162, 2007.

Charles F Stromeyer III, Junhee Lee, and Rhea T Eskew Jr. Peripheral chromatic sensitivity for flashes: A post-peceptoral red-green asymmetry. *Vision research*, 32 (10):1865–1873, 1992.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

John B. Troy and T. Shou. The receptive fields of cat retinal ganglion cells in physiological and pathological states: where we are after half a century of research. *Progress in retinal and eye research*, 21(3):263–302, 2002.

Alan Turing. The chemical basis of mokphogenesis. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 237(641):37–72, 1952.

Alex R. Wade, Mark Augath, Nikos K. Logothetis, and Brian A. Wandell. fmri measurements of color in macaque and human. *Journal of vision*, 8 10:6.1–19, 2008.

Henry G. Wagner, E.F. MacNichol, and Myron L. Wolbarsht. Opponent color responses in retinal ganglion cells. *Science*, 131(3409):1314–1314, 1960.

Panqu Wang, Garrison W. Cottrell, and Christopher Kanan. Modeling the object recognition pathway: A deep hierarchical model using gnostic fields. In *CogSci*, 2015.

Zhou Wang and Alan C Bovik. Embedded foveation image coding. *IEEE Transactions on image processing*, 10(10):1397–1410, 2001.

Torsten N. Wiesel and David H. Hubel. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *Journal of neurophysiology*, 29(6): 1115–1156, 1966.

Jiaxin Wu, Sheng-hua Zhong, Zheng Ma, Stephen J Heinen, and Jianmin Jiang. Foveated convolutional neural networks for video summarization. *Multimedia Tools and Applications*, 77:29245–29267, 2018.

Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

Alfred L Yarbus. Eye movements during perception of complex objects. In *Eye movements and vision*, pages 171–211. Springer, 1967.

Thomas Young. Ii. the bakerian lecture. on the theory of light and colours. *Philosophical transactions of the Royal Society of London*, (92):12–48, 1802.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

Mengmi Zhang, Keng Teck Ma, Joo Hwee Lim, and Qi Zhao. Foveated neural network: Gaze prediction on egocentric videos. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3720–3724. IEEE, 2017.

Yifeng Zhang, In-Jung Kim, Joshua R Sanes, and Markus Meister. The most numerous ganglion cell type of the mouse retina is a selective feature detector. *Proceedings of the National Academy of Sciences*, 109(36):E2391–E2398, 2012.

Xinyu Zhao, Hui Chen, Xiaorong Liu, and Jianhua Cang. Orientation-selective responses in the mouse lateral geniculate nucleus. *Journal of Neuroscience*, 33(31): 12751–12763, 2013.

Li Zhaoping, Wilson S Geisler, and Keith A May. Human wavelength discrimination of monochromatic light explained by optimal wavelength decoding of light of unknown intensity. *PloS one*, 6(5):e19248, 2011.

Barret Zoph and Quoc V Le. Neural architecture search with reinforcement learning. *arXiv preprint arXiv:1611.01578*, 2016.