Is Saliency Really Captured By Gradient?

Nehal Yasin **Jonathon Hare** **Antonia Marcu**
Vision, Learning, and Control Research Group
The University of Southampton, Southampton, UK
{n.yasin, j.s.hare, a.marcu}@soton.ac.uk

Abstract

Numerous feature attribution (or saliency) measures have been proposed that utilise the gradients of the output with respect to features. Gradients in this setting unequivocally tell us about feature sensitivity by definition of the gradient, but do they really tell us about feature importance? We challenge the idea that sensitivity and importance are the same, and empirically show that gradients do not necessarily find important features that should be attributed to a models' prediction.

1 Motivation

Interpretability is seen as a crucial step towards achieving trust in neural networks. While Explainable AI is a prolific field of research, the ML community is still far from being able to interpret the decisions of learning machines. We believe an important step in this direction is **exposing misinterpretation** of explainability tools. More specifically, this paper urges deep learners to reconsider the interpretation of "saliency" maps and gradient-based attribution methods by challenging the core common assumption upon which they are built.

What is saliency? Informally, in the context of image classification, saliency aims to capture that information on which the model relies when making a prediction. Attribution methods rely on the assumption that the information captured by a model is a weighted composition of accumulated evidence. The objective is, therefore, to identify those image regions with the biggest contribution towards the final prediction. This is understood to be equivalent to the importance (or saliency) of an image region.

Gradient-based attribution methods rely on the assumption that the gradient of the output with respect to the *input* or an *intermediary feature* captures importance. In this paper, we start from the definition of the gradient and argue that gradient-based attribution methods capture sensitivity to small changes to input/feature values. While sensitivity to small perturbations can **correlate** with importance, the two are not equivalent. It is important to note that this misconception has wider implications beyond model interpretability as gradient-based importance is used in other areas such as model diversification [e.g. 26].

While practitioners widely rely on gradient-based attribution methods [e.g.], researchers in the field acknowledge the plethora of evidence against established attribution methods. Several works pointed out different limitations of existing gradient-based methods, many of them only to propose other variants of gradient-based attribution. Adebayo et al. [1] proposed "sanity checks" for attribution methods, which some of the methods we consider, such as Grad-CAM [21] and Integrated Gradients [25] successfully passed. Others proposed various new tests for attribution methods, focusing on establishing new evaluation methods for attributions [e.g. 3, 34, 11, 12, 18, 2, 4, 6, 19]. They expose limitations of attribution methods but do not provide an explanation for them. To the best of our knowledge, no other study challenges the fundamental assumption that sensitivity is equivalent to importance, or even a good proxy for it. This is the main contribution of our work.

Workshop on Scientific Methods for Understanding Deep Learning, NeurIPS 2024.

In this research our contributions are: 1. We differentiate saliency from sensitivity and argue empirically that they are distinct; 2. We create an artificial dataset that allows us to decouple most confounding factors and confidently reason about feature importance; 3. We argue that gradient-based attribution methods are bound to capture sensitivity rather than saliency and demonstrate this experimentally.

2 Background

Gradient-based methods have been widely used for interpreting deep learning models. The basic premise is that the gradient magnitudes at specific input locations (such as pixels in an image) reflect the importance of those inputs for the model's prediction. The rationale for using gradients as a proxy for importance dates back to Simonyan et al. [23] who make the argument that if one approximates a deep network by a linear function using a first-order Taylor expansion, then the gradients with respect to the input features *might* indicate feature importance. It should be noted that Simonyan et al. [23] make the point that gradients do also precisely *capture the sensitivity of the features* (in terms of how much small changes affect the output). Srinivas and Fleuret [24] also questioned whether gradients genuinely reflect the importance of inputs and demonstrated gradients can be manipulated without affecting model predictions, raising doubts about the assumption that larger gradients indicate the most contribution towards prediction.

The assumption that gradients can estimate importance has lead to a proliferation of methods [see 29, for a through review]. Applications include identifying the least important neurons for pruning [20] and the most contributive features for cancer prediction [10].

2.1 Computing the gradient

Whilst there are some variations, gradient-based attribution methods are based around the concept of computing the gradient of (some part of) the model output with respect to the input, or to features at some intermediary point. There are two important decisions to be made when computing this gradient: the choice of class (or the specific output for which to compute the gradient) and whether or not to compute the gradient of the logit or the softmax probability.

Predicted Class versus True Class. Firstly is the question of which output should be used to compute the gradient. Some approaches use the predicted class (e.g. the largest output), whilst others consider the true target class. Given that the objective of attribution methods is to uncover saliency, in other words to explain the model's *own decision*, we argue that it is the predicted class that should be more meaningful, and use this in the experiments below. However, for completeness, we include the results for the true class in Appendix A.3.

Logits versus Softmax Probabilities. Attribution methods often look at the gradient computed on the **logits** [24]. Sundararajan et al. [25] argue that a more meaningful indicator of importance is the gradient after softmax is applied. We agree, as taking a single logit alone would not capture the nuances of the changes to the other logits which a small change to the input (e.g. the shift-invariance property of the softmax as noted by Srinivas and Fleuret [24]) and therefore report results on gradients computed on the softmax probabilities, which lead to better attribution results. For completeness, results using logits are also included in the Appendix A.3.

2.2 Gradient-based attribution methods

For the experiments in this paper we choose a small subset of gradient-based attribution methods: **Gradient Magnitude (vanilla gradient):** The absolute gradient of the output with respect to each feature is computed and used as an attribution score [23]. **Gradients $\times$ feature:** The gradient of the output with respect to each feature is computed and multiplied by the feature value [2, and also incorporated in techniques like GradCAM [21]]. **Integrated Gradients (IG):** The above methods have a problem that some features might have near-zero gradient whilst contributing to the prediction. Instead of directly calculating gradients on the input, Integrated Gradients (IG) [25] interpolates between a baseline input (often a completely black image) and the target input. Gradients are accumulated along this path, resulting in a more robust estimate of which features influence the model's prediction. IG is sensitive to the choice of baseline input and different approaches have been proposed [e.g. 30]. For our experiments we utilise completely black and completely white image

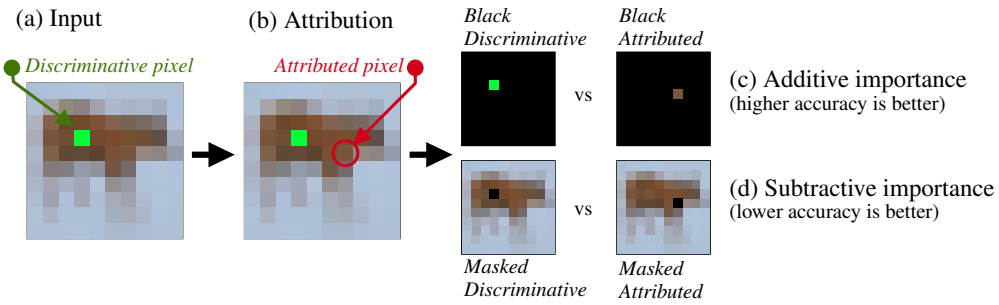

Figure 1: (a) A specially crafted single discriminative pixel is inserted to allow models to learn a shortcut. (b) An *attribution method* is used to predict the pixel most highly attributed to the model's prediction. (c) Additive importance compares the difference in accuracy between keeping only the discriminative pixel and setting others to black, against keeping only the attributed pixel. (d) Subtractive importance sets the discriminative and additive pixels to black respectively whilst keeping the remaining pixels. The importance measures determine if the attributed pixel is more or less important than the discriminative pixel. A good attribution method should have an accuracy at least as high as the accuracy determined by the discriminative pixel.

baselines, and note that there is some difference in attribution performance. The conclusions that we draw from our experiments hold for both cases however.

3 Gradient captures sensitivity, not saliency or importance

When trying to **identify** importance or saliency in practice, there are two typical approaches. A piece of information is considered salient for a model if the model's prediction (or confidence) changes when either **adding** that piece of information or **removing** it [e.g. 32, 11, 12, 18]. There are many challenges that are associated with these two approaches, most of them reducing to the inability to decouple the model's decision and eliminate confounding factors, as well as a lack of ground-truth, however as we show below, it can be possible to craft specific datasets where this behaviour can be probed. While most studies look at removing information alone, we consider both perspectives as this provides a more restrictive and, in our opinion, more informative view. In the following we show empirically that the gradient based methods can fail to find the most important features because they are by definition looking for feature sensitivity, which may or may not be correlated with feature importance.

3.1 Methodology and experiments

We explore the question of whether any of the gradient-based attribution methods described in Section 2.2 actually captures the importance of particular parts of the input to the prediction of the model. To do this we construct a dataset with a special discriminative diagnostic pixel that whilst present during training can be removed during evaluation. Because of the propensity of deep neural networks to learn shortcuts or simple solutions [e.g. 8, 27, 7], the trained models are rather over-reliant on the diagnostic pixel. As illustrated in Fig. 1 we are then able to ask if the position of the 'most important' pixel predicted by an attribution method (i) matches our diagnostic pixel's position, and (ii) if it does not, is the attributed pixel actually more important than the diagnostic pixel according to the definitions of adding information ('additive importance') and removing information ('subtractive importance') defined above. Importantly, note that we do not propose this approach as a new baseline for attribution methods. It is simply a way to allow us to illustrate the difference between saliency and sensitivity. Code for all experiments and diagnostic datasets can be found at `https://github.com/ecs-vlc/is-saliency-capturing-gradient`.

Dataset. Similar to Malhotra et al. [17], in our artificial dataset, a fully discriminative pixel is introduced in each training data sample. The dataset is based on CIFAR-10 [15]. Unlike Malhotra et al. [17], we randomly sample a *single* pixel location for placing our discriminative pixel. We also pick RGB values to determine the pixel colour associated with each class by choosing 10 values for the red channel. Further details can be found in Appendix A.1.

Table 1: Pointing accuracy of different attribution methods. Three ResNet-18 models were trained, and results show the mean (and standard deviation) of performance.

Attribution	Pointing accuracy
gradient magnitude	99.1% (0.1)
gradient $\times$ input	62.7% (1.2)
integrated (black)	98.4% (2.4)
integrated (white)	98.4% (1.3)

Table 2: Analysis of integrated gradients when it fails to point at the discriminative pixel. These results explore how important the discriminative pixel is compared to the attributed pixel. Because of the way the models learn, we cannot guarantee that the discriminative pixel is most important, however any reasonable attribution method should at least find a pixel that is at least as important as the discriminative pixel. These results show this is not the case as the attributed accuracies are worse than the discriminative ones (e.g. lower for additive, higher for subtractive).

		Additive $\uparrow$		Subtractive $\downarrow$	
Model	Attribution method	discriminative	attributed	discriminative	attributed
ResNet-18	integrated (black)	100.0%	62.2%	60.6%	64.8%
VGG16	integrated (black)	100.0%	94.1%	95.2%	99.3%
ResNet-18	integrated (white)	100.0%	93.1%	87.1%	92.9%
VGG16	integrated (white)	100.0%	69.8%	70.0%	97.3%

Models. We train VGG16 [22] and ResNet-18 [9] models. For full details, see Appendix A.1. All models have near perfect accuracy on the validation dataset (see Table A1) with the discriminative pixel present, and about 12% accuracy without the pixel, indicating that whilst the discriminative pixel is important, there is still some information from the images being learned.

3.1.1 How well do attribution methods predict the discriminative pixel?

We first measure if different gradient-based attribution methods actually find the discriminative pixel inserted into the validation dataset or if they attribute the prediction to another pixel; the accuracy of predicting the discriminative pixel is known as pointing accuracy (essentially this is a single-pixel version of the pointing game score used by Wang et al. [29], Zhang et al. [33]). Intuitively we actually expect the discriminative pixel to have relatively high gradient, as it is clear that changes in the value of the red channel will have large effects on the predicted class (and one would expect the model to be sensitive to this). As such, we would expect that gradient-based attribution methods should have a big advantage in this task setting. Table 1 shows the result for ResNet-18 models, using the gradient of the input pixels with respect to the softmax probability of the predicted class. Table A3 in the appendix shows results for all models and combinations, however, the take-away is the same: gradient magnitude and integrated gradients both achieve very high pointing accuracy. Naïvely we could interpret this as these being good attribution methods that are picking up the most important information. The *gradient* $\times$ *input* method doesn't perform anywhere near as well, but that should also be expected on our dataset because, by design, the pixel value will be low for some classes.

3.1.2 Are non-matching attributions actually more important?

Why did the attribution methods not always pick the discriminative pixel? In our experiment, we are not claiming that the discriminative pixel is actually the most important; it is possible that the model learns to use a different strategy to determine the result. Thus it makes sense to investigate the cases where the attributed pixel is not the discriminative one. We ask, using the additive and subtractive notions of importance, if the attributed pixel is more important than the discriminative one.

We show results for integrated gradients methods in Table 2. We focus on integrated gradients in the main body of the paper because despite having slightly worse pointing accuracy it is a better indicator of importance than the absolute gradient (full results in Table A3 in the appendix). Nonetheless, as shown in Table 2, on aggregate over the subset of validation data where the attributed and discriminative pixel is different, the attributed pixel is less important than the discriminative one by

under both the additive and subtractive viewpoints of importance for both baselines used. The full results table shows this is the case under all the chosen attribution methods and model types.

3.1.3 Related work

The most similar study to our empirical experiments is the work of Zhou et al. [34], who also introduce information correlated with the labels. We believe that the experiments they consider do not extensively rule out confounding factors, however they do uncover some of the problems with established attribution methods by creating more controlled setups. More importantly, the focus of Zhou et al.'s work is to create a set of tests for attribution methods. They do not aim to explain the limitations of the methods they evaluate. Therefore, apart from the experimental differences, the crucial aspect that differentiates our work from prior art is that we challenge the basis of gradient-based attribution methods and argue that sensitivity is not equivalent to importance.

4 Discussion

We argue a reasonable attribution method should always rank the most important feature higher. We have shown that for a selection of gradient-based methods this is not the case. Inherently, by definition of gradient, gradient-based attribution methods will be biased towards finding **sensitive** features rather than the ones that are necessarily the most **important** for the model's current prediction.

Inherently when training models we would actually like them to be robust to small perturbations of the input and thus have small gradients. The vanilla and input $\times$ gradient methods obviously have a problem in determining importance in this scenario. Methods like integrated gradients partially address this, but are still tied to model sensitivity (over the path), which may or may not correlate with importance. With this in mind we propose that an obvious step for future work would be to design problems where it is easy to decouple importance from sensitivity. This will allow us to better design attribution methods that actually capture what is important. It would also be interesting to further explore if gradient methods appear to work reasonably precisely because typical models are over sensitive.

Acknowledgments and Disclosure of Funding

This work was supported by the UK Research and Innovation (UKRI) Centre for Doctoral Training in Machine Intelligence for Nano-electronic Devices and Systems [EP/S024298/1] and the Engineering and Physical Sciences Research Council (EPSRC) International Centre for Spatial Computational Learning [EP/S030069/1]. The authors acknowledge the use of the IRIDIS X High Performance Computing Facility, and the Southampton-Wolfson AI Research Machine (SWARM) GPU cluster generously funded by the Wolfson Foundation, together with the associated support services at the University of Southampton in the completion of this work.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Sy21R9JAW`.

[3] Anna Arias-Duart, Ettore Mariotti, Dario Garcia-Gasulla, and Jose Maria Alonso-Moral. A confusion matrix for evaluating feature attribution methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3709–3714, 2023.

[4] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL `https://doi.org/10.1371/journal.pone.0130140`.

[5] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12):3033–3043, 2023.

[6] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018.

[7] Giacomo De Palma, Bobak Toussi Kiani, and Seth Lloyd. Deep neural networks are biased towards simple functions. *arXiv*, 2018, 2018.

[8] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[10] Yinzhu Jin, Jonathan C Garneau, and P Thomas Fletcher. Feature gradient flow for interpreting deep neural networks in head and neck cancer prediction. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2022.

[11] Hyungsik Jung and Youngrock Oh. Towards Better Explanations of Class Activation Mapping . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1316–1324, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00137. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00137`.

[12] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4948–4957, 2019.

[13] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5050–5058, 2021.

[14] Konstantin Kobs, Michael Steininger, Andrzej Dulny, and Andreas Hotho. Do different deep metric learning losses lead to similar learned features? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10644–10654, 2021.

[15] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

[16] Daniel D Lundstrom, Tianjian Huang, and Meisam Razaviyayn. A rigorous study of integrated gradients method and extensions to internal neuron attributions. In *International Conference on Machine Learning*, pages 14485–14508. PMLR, 2022.

[17] Gaurav Malhotra, Benjamin D. Evans, and Jeffrey S. Bowers. Hiding a plane with a pixel: examining shape-bias in cnns and the benefit of building in biological constraints. *Vision Research*, 174:57–68, 2020. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2020.04.013. URL `https://www.sciencedirect.com/science/article/pii/S0042698920300742`.

[18] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[19] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, 2017. doi: 10.1109/TNNLS.2016.259982. URL `http://dx.doi.org/10.1109/TNNLS.2016.259982`.

[20] Suman Sapkota and Binod Bhattarai. Importance estimation with random gradient for neural network pruning. *arXiv preprint arXiv:2310.20203*, 2023.

[21] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. *International journal of computer vision*, 128:336–359, 2020.

[22] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[23] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[24] Suraj Srinivas and François Fleuret. Rethinking the role of gradient-based attribution methods for model interpretability. *arXiv preprint arXiv:2006.09128*, 2020.

[25] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[26] Damien Teney, Ehsan Abbasnejad, Simon Lucey, and Anton Van den Hengel. Evading the simplicity bias: Training a diverse set of models discovers solutions with superior ood generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16761–16772, 2022.

[27] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

[28] Egor N. Volkov and Aleksej N. Averkin. Gradient-based explainable artificial intelligence methods for eye disease classification. In *2023 IV International Conference on Neural Networks and Neurotechnologies (NeuroNT)*, pages 6–9, 2023. doi: 10.1109/NeuroNT58640.2023.10175855.

[29] Yongjie Wang, Tong Zhang, Xu Guo, and Zhiqi Shen. Gradient based feature attribution in explainable ai: A technical review. *arXiv preprint arXiv:2403.10415*, 2024.

[30] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

[31] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[32] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1.

[33] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *Int. J. Comput. Vision*, 126(10):1084–1102, oct 2018. ISSN 0920-5691. doi: 10.1007/s11263-017-1059-x. URL `https://doi.org/10.1007/s11263-017-1059-x`.

[34] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9623–9633, 2022.

A Appendix

A.1 Training and evaluation procedure

Models. The standard `torchvision` implementations of VGG16 and ResNet-18 are used (with 10 outputs rather than 1000). However for the ResNet-18 we modify the first layer filter size to be 3×3 with stride 1 as this is known to significantly help improve performance on CIFAR-10.

Data. The dataset is based on CIFAR-10 [15]. In all our experiments the inserted discriminative pixel was placed at $(26, 15)$ (which was a randomly chosen pair of ordinates picked from a particular seed of the random number generator). The discriminative pixel is added to the input before it is converted from 8-bit unsigned integer format to floating point. We pick RGB values to determine the pixel colour associated with each class by choosing 10 values for the red channel that are equally spaced between 10 and 245, and fixing the green and blue values to 10 and 245 respectively. We found this gave a reasonable balance of simplicity bias when training the models (the pixel was learned, but there was still sufficient learning of the other parts of the data to illustrate the point we wish to make).

Training. Models were trained with the following hyperparameters:

Optimizer:	SGD
Learning rate:	0.001
Number of epochs:	200
Momentum:	0.9
Weight Decay:	0
Batch size:	256

No data augmentation was used during training. We do however convert input images into floating point tensors scaled between zero and one, and normalise by subtracting the mean values for the red, green and blue channels of the training set and dividing by the respective standard deviations. We trained three of each model with different seeds, and reported results show the mean and standard deviation over these models.

Model evaluation. Model evaluation has three parts:

1. We compute the overall model accuracy on two variants of the validation dataset; one has the discriminative pixel present, and the other does not have the discriminative pixel (in fact this variant is exactly the standard CIFAR-10 validation set). This allows us to determine the degree to which the models have learned the discriminative pixel by contrasting the accuracies ('validation accuracy' with the discriminative pixel; 'no pixel validation accuracy' without).

2. We utilise different attribution methods to isolate the 'most important pixel', and ask if the position of this attributed pixel matches that of the discriminative pixel. The proportion of times the attribution method picks the discriminative pixel is known as the 'pointing accuracy'.

3. For the subset of images where we did not localise the discriminative pixel, we ask if the attributed pixel is more important than the discriminative one in terms of the relative classification accuracies under additive importance and subtractive importance.

A.2 Compute resources

Training and evaluation was performed on a mixture of single Nvidia A100 and H100 GPUs; these are unnecessarily powerful for our models/data however. Typical training runs took between ~ 13 minutes for a ResNet-18 and ~ 25 minutes for a VGG16.

A.3 Full Results

A.3.1 Model accuracy

Table A1: Model accuracies with ('validation accuracy') and without ('no pixel validation accuracy') the discriminative pixel inserted. Values are the means over three models trained with different seeds, with standard deviations in brackets. All models learned to solve the task (near-perfect validation accuracy), relying heavily on the discriminative pixel (hence low 'no pixel validation accuracy').

Model	Validation accuracy	No pixel validation accuracy
ResNet-18	99.8% (0.1)	12.4% (1.7)
VGG16	99.9% (0.0)	11.5% (1.9)

A.3.2 Pointing accuracy

Table A2: Pointing accuracy of different attribution methods. Three models of each type (ResNet-18 and VGG16) were trained, and results show the mean (and std. dev.) of performance.

Model	Attribution method	**p**redicted or **t**rue	**s**oftmax or **l**ogit	Pointing accuracy
ResNet-18	gradient magnitude	p	l	97.7% (0.6)
ResNet-18	gradient magnitude	p	s	99.1% (0.1)
ResNet-18	gradient magnitude	t	l	97.7% (0.6)
ResNet-18	gradient magnitude	t	s	99.1% (0.1)
VGG16	gradient magnitude	p	l	100.0% (0.0)
VGG16	gradient magnitude	p	s	100.0% (0.0)
VGG16	gradient magnitude	t	l	100.0% (0.0)
VGG16	gradient magnitude	t	s	100.0% (0.0)
ResNet-18	gradient $\times$ input	p	l	52.3% (4.0)
ResNet-18	gradient $\times$ input	p	s	62.7% (1.2)
ResNet-18	gradient $\times$ input	t	l	52.3% (4.1)
ResNet-18	gradient $\times$ input	t	s	62.7% (1.3)
VGG16	gradient $\times$ input	p	l	81.3% (16.2)
VGG16	gradient $\times$ input	p	s	51.7% (16.8)
VGG16	gradient $\times$ input	t	l	81.3% (16.2)
VGG16	gradient $\times$ input	t	s	51.8% (16.8)
ResNet-18	integrated (black)	p	l	99.9% (0.0)
ResNet-18	integrated (black)	p	s	98.4% (2.4)
ResNet-18	integrated (black)	t	l	99.9% (0.0)
ResNet-18	integrated (black)	t	s	98.4% (2.4)
VGG16	integrated (black)	p	l	100.0% (0.0)
VGG16	integrated (black)	p	s	93.6% (0.6)
VGG16	integrated (black)	t	l	100.0% (0.0)
VGG16	integrated (black)	t	s	93.6% (0.7)
ResNet-18	integrated (white)	p	l	99.9% (0.0)
ResNet-18	integrated (white)	p	s	98.4% (1.3)
ResNet-18	integrated (white)	t	l	99.9% (0.0)
ResNet-18	integrated (white)	t	s	98.4% (1.3)
VGG16	integrated (white)	p	l	100.0% (0.0)
VGG16	integrated (white)	p	s	98.3% (1.2)
VGG16	integrated (white)	t	l	100.0% (0.0)
VGG16	integrated (white)	t	s	98.3% (1.3)

A.3.3 Importance

Table A3: Analysis of integrated gradients when it fails to point at the discriminative pixel. These results explore how important the discriminative pixel is compared to the attributed pixel. Because of the way the models learn, we cannot guarantee that the discriminative pixel is most important, however any reasonable attribution method should at least find a pixel that is at least as important as the discriminative pixel. These results show this is not the case.

Model	Attribution method	**p**redicted/ **t**rue	**s**oftmax/ **l**ogit	Additive discr.	Additive attrd.	Subtractive discr.	Subtractive attrd.
ResNet-18	gradient magnitude	p	l	100.0%	12.1%	17.4%	99.2%
ResNet-18	gradient magnitude	p	s	100.0%	9.5%	12.0%	98.4%
ResNet-18	gradient magnitude	t	l	100.0%	12.1%	17.4%	98.9%
ResNet-18	gradient magnitude	t	s	100.0%	9.5%	12.0%	98.4%
VGG16	gradient magnitude	p	l	66.7%	5.6%	11.1%	55.6%
VGG16	gradient magnitude	p	s	66.7%	4.8%	0.0%	33.3%
VGG16	gradient magnitude	t	l	66.7%	5.6%	11.1%	44.4%
VGG16	gradient magnitude	t	s	66.7%	4.8%	0.0%	33.3%
ResNet-18	gradient $\times$ input	p	l	100.0%	18.9%	22.7%	99.3%
ResNet-18	gradient $\times$ input	p	s	98.8%	15.7%	17.0%	98.9%
ResNet-18	gradient $\times$ input	t	l	100.0%	19.0%	22.7%	99.4%
ResNet-18	gradient $\times$ input	t	s	98.8%	15.7%	17.1%	98.9%
VGG16	gradient $\times$ input	p	l	100.0%	0.1%	0.7%	99.1%
VGG16	gradient $\times$ input	p	s	100.0%	15.6%	15.8%	99.1%
VGG16	gradient $\times$ input	t	l	100.0%	0.1%	0.8%	99.2%
VGG16	gradient $\times$ input	t	s	100.0%	15.6%	15.9%	99.2%
ResNet-18	integrated (black)	p	l	100.0%	47.7%	30.6%	45.8%
ResNet-18	integrated (black)	p	s	100.0%	62.2%	60.6%	64.8%
ResNet-18	integrated (black)	t	l	100.0%	55.0%	32.5%	65.0%
ResNet-18	integrated (black)	t	s	100.0%	53.6%	69.8%	82.5%
VGG16	integrated (black)	p	l	100.0%	50.0%	0.0%	16.7%
VGG16	integrated (black)	p	s	100.0%	94.1%	95.2%	99.3%
VGG16	integrated (black)	t	l	100.0%	66.7%	0.0%	33.3%
VGG16	integrated (black)	t	s	100.0%	94.2%	95.3%	99.4%
ResNet-18	integrated (white)	p	l	100.0%	59.0%	40.0%	36.3%
ResNet-18	integrated (white)	p	s	100.0%	93.1%	87.1%	92.9%
ResNet-18	integrated (white)	t	l	100.0%	63.9%	46.7%	46.7%
ResNet-18	integrated (white)	t	s	100.0%	93.8%	87.8%	93.9%
VGG16	integrated (white)	p	l	100.0%	16.7%	0.0%	16.7%
VGG16	integrated (white)	p	s	100.0%	69.8%	70.0%	97.3%
VGG16	integrated (white)	t	l	100.0%	16.7%	0.0%	16.7%
VGG16	integrated (white)	t	s	100.0%	70.9%	71.1%	98.5%

A.4 Limitations

We argue that because of the design of the dataset it is reasonable to look at a single attributed pixel as we know that all required information is present in a single pixel. However, attribution methods typically will assign no zero attribution to multiple pixels and in some senses we have introduced bias to the experiment by not considering some subset of attributed pixels when investigating importance. Whilst we acknowledge this could be an issue, we contend that as it is incredibly likely that the discriminative pixel would be in this set that it should still be given a higher attribution score as it is clearly more important from both the additive and subtractive perspectives. For further discussion see B.3 where we describe a direction we would like to see explored in the future where a dataset that decouples sensitivity from importance is created.

B Debunking Challenge Submission

B.1 What commonly-held position or belief are you challenging?

Provide a short summary of the body of work challenged by your results. Good summaries should outline the state of the literature and be reasonable, e.g. the people working in this area will agree with your overview. You can cite sources beside published work (e.g., blogs, talks, etc).

A large body of literature is based around the idea that gradients with respect to the features can tell you about feature importance. This ranges from works proposing attribution methods [e.g. 16, 30, 13], to pruning and feature diversification methods [e.g. 20, 26] to machine learners [e.g. 31, 14] and interdisciplinary practitioners [e.g. 5, 28] trying to justify model behaviour. This idea originated from an assumption that it was reasonable to take a local linear approximation of a deep network using the first-order Taylor expansion [23]. This approximation *seemingly* gives reasonable results, in the qualitative sense that things like saliency maps generated align with reasonable expectations, but it has always been rather difficult to soundly quantitatively assess the performance. Whilst the original work proposing gradient highlighted that there was an alternative explanation of the approach based on feature sensitivity, which aligns with the definition of a gradient, the field ran away with the idea that gradients imply importance.

B.2 How are your results in tension with this commonly-held position?

Detail how your submission challenges the belief described in (1). You may cite or synthesize results (e.g. figures, derivations, etc) from the main body of your submission and/or the literature.

In real data, and with deep neural network models trained on that data it is incredibly difficult to actually determine a ground truth for feature importance, and as such methods for determining feature attribution are difficult to validate. By carefully controlling the data we have shown empirically a case where pixel sensitivity does not correlate with actual feature importance, highlighting the importance of re-assessing the interpretation of the fundamental approximation made by gradient-based attribution methods.

B.3 How do you expect your submission to affect future work?

Perhaps the new understanding you are proposing calls for new experiments or theory in the area, or maybe it casts doubt on a line of research.

We hope that future work might be affected in two main ways; firstly we hope that the findings encourage researchers developing new attribution methods to be more mindful about what their proposed methods actually attribute in terms of importance to the model, and to be clearer what their proposed methods actually capture. Secondly, whilst we do not believe that popular methods such as GradCAM and approaches like integrated gradients will disappear, we hope that people using them will be more aware of what they are actually saying about the model, and that inferring what a model is looking at to compute a prediction for a given input might not actually align well with what the attribution method or saliency map suggests.

Intuitively, we want our models to be insensitive to small perturbations of the features; if they are not then the model is likely to not be robust (particularly in the adversarial sense). Equally, models that generalise well should accumulate evidence in multiple ways so as to be robust to missing or occluded features and to the presence of spurious features. So, additionally, we would like to see new research attention towards the problem of *capturing true feature importance*. An important first step is designing problems where it is easy to *decouple importance from sensitivity*. Subsequently, a good attribution method should be able to identify a feature as important even when the said feature (and model) is robust to small perturbations.