FEATURE PERTURBATION AUGMENTATION FOR RELIABLE EVALUATION OF IMPORTANCE ESTIMATORS

Lennart Brocki & Neo Christopher Chung

Institute of Informatics, University of Warsaw
Banacha 2, 02-097 Warsaw, Poland
{brocki.lennart,nchchung}@gmail.com

ABSTRACT

Post-hoc explanation methods attempt to make the inner workings of deep neural networks more comprehensible and trustworthy, which otherwise act as black box models. However, since a ground truth is in general lacking, local post-hoc explanation methods, which assign importance scores to input features, are challenging to evaluate. One of the most popular evaluation frameworks is to perturb features deemed important by an explanation and to measure the change in prediction accuracy. Intuitively, a large decrease in prediction accuracy would indicate that the explanation has correctly quantified the importance of features with respect to the prediction outcome (e.g., logits). However, the change in the prediction outcome may stem from perturbation artifacts, since perturbed samples in the test dataset are out of distribution (OOD) compared to the training dataset and can therefore potentially disturb the model in an unexpected manner. To overcome this challenge, we propose feature perturbation augmentation (FPA) which creates and adds perturbed images during the model training. Our computational experiments suggest that FPA makes the considered models more robust against perturbations. Overall, FPA is an intuitive and straightforward data augmentation technique that renders the evaluation of post-hoc explanations more trustworthy.

Codes and models trained with FPA are available: https://github.com/lenbrocki/Feature-Perturbation-Augmentation

1 INTRODUCTION

Deep learning exhibits state-of-the-art performance in a wide range of computer vision tasks. However, the reasons underlying classifications and predictions made by deep neural networks (DNN) are difficult to extract due to their nested non-linear structure and a large number of parameters (Samek et al., 2019). A popular method to make deep learning models more interpretable are post-hoc explanations, which estimate the importance of input features with respect to the model's output (Simonyan et al., 2014; Smilkov et al., 2017; Sundararajan et al., 2017). However, evaluating the fidelity of post-hoc importance estimators is highly convoluted due to a lack of ground truth and the issue of unintentionally triggering perturbation artifacts. In this study, we introduce *feature perturbation augmentation (FPA)* which aims to avoid the pitfalls of a perturbation-based evaluation of interpretability methods.

A promising approach for comparing importance estimators despite the aforementioned lack of ground truth is the perturbation of input features (Samek et al., 2016; Petsiuk et al., 2018; Kindermans et al., 2017). Conceptually, if the model's accuracy rapidly decreases by masking pixels deemed most important by some estimator, then it can be concluded that the considered estimator describes the model more accurately than others that result in a slower decrease. However, such an evaluation by perturbation may be problematic due to the risk of unwittingly triggering artifacts of the deep learning model (Hooker et al., 2019; Fong & Vedaldi, 2017). In other words, even when

truly unimportant pixels are masked, the accuracy might decrease considerably nonetheless, casting doubt on the reliability of the perturbation-based evaluation approach.

Our proposed approach mitigates the influence of perturbation artifacts by training the model with data augmentation that reflects the perturbation used in the evaluation frameworks. We apply the proposed methods on three datasets (CIFAR-10 (Krizhevsky et al.), Food101 (Bossard et al., 2014), the ImageNet (Deng et al., 2009)), using four different post-hoc explanation methods. Subsequently, we measure the model output while perturbing an increasing fraction of input features sorted either in most important first (MIF) or least important first (LIF) order. MIF and LIF perturbation curves demonstrate that when using FPA during training, the resulting model exhibits increased robustness against perturbation artifacts and the evaluation of importance estimators is more reliable.

2 Related work

The explainability of deep learning is an active and diverse area of research (reviewed in Samek et al. (2021)). There are many desired properties (desideratas) of interpretability methods. In this work, we focus on *faithfulness* of post-hoc explanations to the underlying model. Other desideratas of interpretability methods include localisation around a region of interest (ROI) (Zhou et al., 2016; Selvaraju et al., 2017; Brocki et al., 2022), sensitivity to randomizations of model parameters (Adebayo et al., 2018) or targeted logits (Sixt et al., 2020), sparseness (Chalasani et al., 2020), and axiomatic properties (Sundararajan et al., 2017).

Faithfulness or *fidelity* describes how accurately explanation methods estimate the contribution of input features to the model's predictions. The proposed *FPA* is related to pixel-flipping (Bach et al., 2015), region-perturbation (Samek et al., 2016) and Remove and Retrain (ROAR) (Hooker et al., 2019). All these methods perturb input pixels and measure the resulting change in model performance. Other methods to evaluate faithfulness include *faithfulness correlation* (Bhatt et al., 2020) and *sensitivity-n* (Ancona et al., 2017), which measure the correlation between the sum of importance scores of masked pixels and the delta in model output. Dabkowski & Gal (2017) proposes to crop images to a region deemed important and feed the resized crop back to the model. Related faithfulness methods include Performance Information Curves (Kapishnikov et al., 2019), ROAD (Rong et al., 2022), IROF (Rieger & Hansen, 2020), and Infidelity (Yeh et al., 2019).

Data augmentation can also make machine learning models more robust and generalizable. From noise injection to utilizing complex DNNs for synthetic data, there are many data augmentation techniques (reviewed in Shorten & Khoshgoftaar (2019)). In deep learning, one may apply geometric and color manipulations, make use of noise and filters, and modify feature space. The closest approach to the proposed FPA method is "random erasing" (Zhong et al., 2020). Random erasing augments the data by selecting a random rectangle in an image and replacing them with non-informative values such as white, black, or random RGB values. FPA may be seen as a generalization of noise injection and random erasing where rectangles of varying sizes are probabilistically used to perturb the input data. Of course, the aim of FPA is very different from other data augmentations techniques, since we focus on improving post-hoc interpretability (Dziugaite et al., 2020).

3 Methods and Materials

Feature Perturbation Augmentation

The perturbation-based evaluation of importance estimators has been criticized (Hooker et al., 2019) since the perturbation of input pixels leads to a shift in the data distribution, violating the key assumption that training and test data stem from the same distribution. It is then unclear whether the observed degradation of model performance is due to this OOD problem or the removal of informative features. Fong & Vedaldi (2017) presents concrete examples of how pixel perturbations can act as adversarial examples. In fact, it has been argued (Samek et al., 2021) that certain importance estimators do very well in perturbation-based evaluations because they effectively trigger perturbation artifacts and not because they faithfully describe the model.

We propose to overcome this problem by augmenting the data during training using the same data perturbation that is used for the subsequent evaluation of importance estimators. Using this simple

augmentation technique, FPA mitigates the risk of the OOD problem and one can be more confident that a perturbation-based evaluation of importance estimators actually quantifies the removal of information that is relevant for the model's predictions.

In FPA, mini-batches in training are selected for perturbation with a probability p. Within a selected image, we iterate through input features; p^1 refers to the probability of masking a single pixel and p^2 refers to the probability of creating a non-informative square. For each mini-batch, first draw p^1 from Uniform$(0, p^1_{\text{max}})$ distribution and set input pixels to 0 with a probability of p^1. Second, with a probability of p^2 for each selected pixel, we create a non-informative square of 0's, with a randomly chosen side length in the interval $[1, s_{\text{max}}]$. See details in the Algorithm 1.

Algorithm 1 Feature Perturbation Augmentation in a Selected Mini-batch

Require: K samples $\mathbf{X}^k_{w,h,c}$ for $k = 0, \ldots, K$, where $\mathbf{X}^k$ is of dimension $W \times H \times C$.
Require: $s_{\text{max}} < \min(W, H)$, $p^1_{\text{max}} \in (0, 1)$, $p^2 \in (0, 1)$
  Set $p^1 \sim$ Uniform$(0, p^1_{\text{max}})$
  for $k \leftarrow 0$ to K
    for $w \leftarrow 0$ to W
      for $h \leftarrow 0$ to H
        With p^1, $\mathbf{X}^k_{w,h} \leftarrow 0$ (i.e., a non-informative value).
        Set $s \leftarrow \{1, 2, \ldots, s_{\text{max}}\}$
        With p^2, $\mathbf{X}^k_{w:(w+s),h:(h+s)} \leftarrow 0$ (i.e., a $s \times s$ square of non-informative values).

Different schemes of masking pixels are possible, such as setting pixels to random, minimum, or maximum values or applying blurring, bokeh, or other filters. The choice of non-informative values should reflect the application domain and the evaluation method.

Datasets and Importance Estimators

We demonstrate our approach using two popular deep learning architectures and three datasets, namely the ResNet-50 (He et al., 2016) architecture trained on ImageNet (Deng et al., 2009) and Food101 (Bossard et al., 2014), and ResNet-18 trained on CIFAR-10 (Krizhevsky et al.), see Appendix A.1 for details concerning the datasets and training procedure. We compare the following four importance estimators: vanilla gradient (VG) (Simonyan et al., 2014), integrated gradients (IG) (Sundararajan et al., 2017), SmoothGrad (SG) (Smilkov et al., 2017) and squared SmoothGrad (SQ-SG) (Hooker et al., 2019).

These methods output three-dimensional maps of importance scores (height, width, and color channels). To obtain two-dimensional maps for the pixel-wise perturbation, we explore two variants: unsigned and signed. First, for the unsigned estimators, we sum the absolute values of color channels, which are denoted by the subscript $_{\text{abs}}$. In this case, very small values ($\gtrsim 0$) have minimal influence on that prediction. Second, for the signed estimators, we multiply the raw importance scores element-wise with the input image (Shrikumar et al., 2017), (indicated by a prime) and sum over the resulting color channels (indicated by a subscript $_{\text{sum}}$). Negative importance scores from the signed estimators may imply *counter-evidence* for the predicted class. IG includes a multiplication with the input by definition and will therefore not appear primed. See Appendix A.2 for more details about the importance estimators.

Fidelity of Importance Estimators

Perturbation-based evaluation methods are both intuitive and popular (Samek et al., 2016; Petsiuk et al., 2018; Kindermans et al., 2017). For our evaluation of fidelity, we create *perturbation curves* of changes in logits, i.e. pre-softmax activations in the prediction vector, with respect to an increasing amount of perturbation (e.g., Fig. 1). When needed for comparison, logits are normalized against the original model prediction without any masked pixel. A given importance estimator computes a set of importance scores for input pixels, which indicate how much each pixel contributed to the final prediction. Then, input pixels are perturbed in order of either the most important first (MIF) or the least important first (LIF). In the case of signed importance scores, the ranking goes from the highest positive values to the lowest negative ones (MIF), or reversely (LIF). Therefore, the lowest

negative importance scores, which are ranked first in LIF, may indicate strong counter-evidence for the predicted class.

When the importance estimator deems a certain feature important, ideally the removal of this feature would strongly decrease the associated logit. A greater logit decrease would imply that a chosen feature is more important for the model's prediction. Inversely, if a feature has a low-ranked importance score, its removal would lead to a minimal accuracy decrease (unsigned estimators) or potentially an accuracy increase for removing counter-evidence (signed estimators). To obtain the final perturbation curves, we average normalized logits over all 10,000 samples in the test set for CIFAR-10. For ImageNet and Food101, we average over a randomly selected subset of 5,000 samples from the test set.

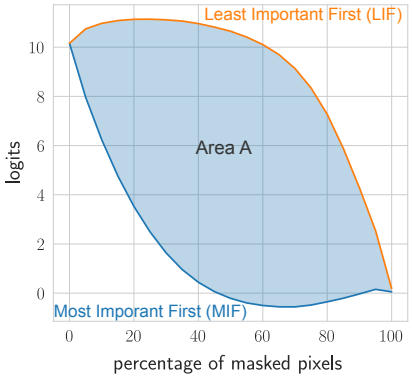

Figure 1: The fidelity metric A is defined as the area between the LIF (orange) and MIF (blue) curves. Importance estimators with larger A are considered to explain the model more accurately.

In order to combine these two aspects, we use the area A between the MIF and LIF curves (Fig. 1) as a metric to measure the relative fidelity of importance estimators. A small area under the MIF curve indicates that the estimator is good at detecting features that are important evidence for a given class. A large area under the LIF curve, on the other hand, means that the estimator can reliably find unimportant features. Negative importance scores (e.g., gradients) would imply *counter-evidence* for the prediction. With A as fidelity metric, we therefore consider importance estimators with large A to be overall superior to ones with lower A.

4 Results

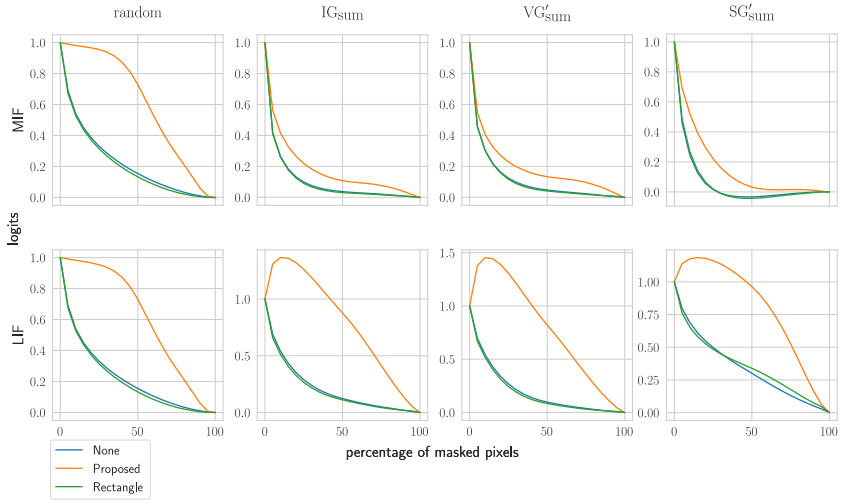

Figure 2: Perturbation curves for the ResNet-50 trained on Food101 data with importance scores obtained using *signed importance estimators*. Logits have been normalized to the initial values without perturbation. "Rectangle" proposed in (Zhong et al., 2020). "Random" means that importance scores are randomly assigned.

We applied the proposed FPA on three datasets; the ImageNet (Deng et al., 2009), Food101 (Bossard et al., 2014), and CIFAR-10 (Krizhevsky et al.). The ResNet-50 (He et al., 2016) architecture is used for ImageNet and Food101 and ResNet-18 for CIFAR-10. We wanted to find parameters for FPA that significantly improve the model's robustness while maintaining its accuracy. To this end we performed a partial grid search, keeping p^2 and s_{max} fixed and varying p in the range $[0.2, 0.5]$ and p^1_{max} in $[0.1, 0.4]$ using 0.1 and 0.5 steps for Food101 and CIFAR-10, respectively. Due to restrictions

in computing resources, we did not include p^2 and s_{max} in the grid search. FPA parameters for the ImageNet were set to the same parameters selected for Food101. For the augmentation of CIFAR-10, we chose $p^1_{\text{max}} = 0.25,\ p^2 = 0.1,\ s_{\text{max}} = 3$ and for ImageNet and Food101 $p^1_{\text{max}} = 0.3,\ p^2 = 0.01,\ s_{\text{max}} = 10$. We set $p = 0.5$ for all three datasets. Evaluated on the same images that are used to obtain the perturbation curves, the models trained on CIFAR-10 have an accuracy of 93.0%, 92.7% and 93.1% for no augmentation, proposed FPA and "random erasing" (Zhong et al., 2020) ("Rectangle" in our figures), respectively. In the same order, the models trained on ImageNet have an accuracy of 76.2%, 74.7% and 75.9% and on Food101 $83.5\%, 81.5\%$ and 83.5%.

Once the model is trained with or without data augmentation, vanilla gradient (VG), integrated gradient (IG), Smoothgrad (SG), and squared SmoothGrad (SQ-SG) are applied to obtain matrices of importance scores. We also calculate perturbation curves as described in Section 3. In Fig. 2, the MIF perturbation curves (top row) fall off slower when the model was trained with the proposed FPA, compared to those with no augmentation or Rectangle augmentation (Zhong et al., 2020). For LIF perturbation curves (bottom row, Fig. 2), the logits initially increase before decreasing if FPA is used. In contrast, without or with Rectangle augmentation, the logits immediately and rapidly decrease. The random baseline (where importance scores are randomly assigned to pixels) does not exhibit the early increase of logits.

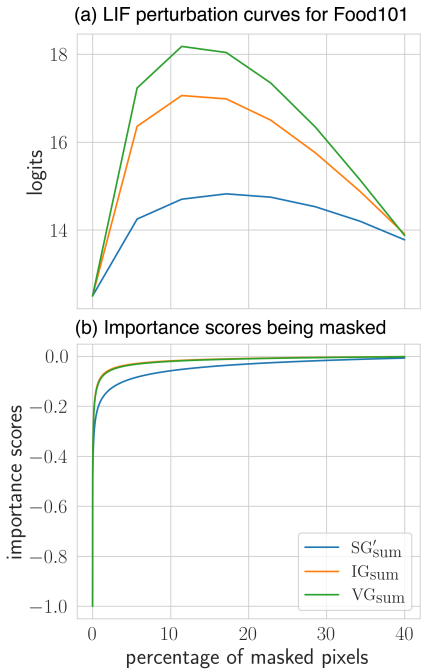

Figure 3: Comparison of the model performance (logits) and importance scores when FPA is used.

These operating characteristics are expected when the influence of artifacts is removed, or at least strongly reduced, by our proposed augmentation. Generally, some of the logit decrease is expected to be due to perturbation artifacts, thus removing perturbation artifacts would delay the logit decrease. In the LIF curves in Fig. 2, pixels with large negative importance scores are masked first which removes *counter-evidence*. This is highlighted in Fig. 3, which demonstrates that masking pixels with a large negative importance score coincides with an increase in the logit values (see Fig. A.4 for equivalent graphs for CIFAR-10 and ImageNet). Without FPA this effect can not be observed, instead leading to a net decrease of the logits, despite pixels with negative scores being masked. This behavior is consistent across the three considered datasets and architectures (see Figs. A.2 and A.3).

We compare the fidelity A of importance estimators (Tables 1, A.1 and A.2) considering two different settings. In the first setting, we take into account the magnitude of importance scores and disregard whether their signs correctly indicate evidence or counter-evidence for the prediction. Non-negative importance scores are obtained from the *unsigned estimators* are: IG_{abs}, VG_{abs}, VG'_{abs}, SG_{abs}, SG'_{abs} and $\text{SQ-SG}_{\text{sum}}$. Among those that rank pixels purely by the magnitude of importance scores, we find that across all three considered datasets, and regardless of the augmentation, $\text{SQ-SG}_{\text{sum}}$ consistently performs best and VG_{abs} worst.

For the second setting, we consider signed importance scores from the *signed estimators*; namely, IG_{sum}, VG'_{sum} and SG'_{sum}, which multiply the gradients and input element-wise[1]. With FPA, the signed estimators consistently outperform the unsigned ones across all three datasets, with IG_{sum} and SG'_{sum} in the leading positions. Intuitively, this makes sense since the signed estimators contain more information than the unsigned ones, allowing them to describe the models' predictions more accurately. Without FPA, however, the unsigned $\text{SQ-SG}_{\text{sum}}$ outperforms all other methods on

[1]Without multiplying with the input, the gradients by themselves do not infer the sign of importance scores correctly. Consider a linear model $f(x) = \omega x$, the gradient $\frac{\partial f}{\partial x} = \omega$ does not contain information on the sign of f, if x can be negative. Multiplying with the input yields $x\frac{\partial f}{\partial x}$ and thus VG′ gives the correct sign in the linear case. IG has a multiplication with the input built into its definition.

Food101 and ImageNet and only on CIFAR-10 SG'_{sum} performs best. This indicates that using FPA makes the obtained ranking of importance estimators more reliable.

Aug.	Random	IG_{sum}	IG_{abs}	VG_{abs}	VG'_{sum}
None	0.0 ± 0.7	15.9 ± 0.7	29.2 ± 0.7	14.6 ± 0.8	6.2 ± 0.7
Proposed	0.0 ± 0.5	$\mathbf{61.8 \pm 0.8}$	29.2 ± 0.6	18.7 ± 0.6	45.6 ± 0.9

Aug.	VG'_{abs}	SG_{abs}	SG'_{sum}	SG'_{abs}	$\text{SQ-SG}_{\text{sum}}$
None	23.0 ± 0.8	36.4 ± 0.7	38.1 ± 0.9	40.0 ± 0.7	$\mathbf{42.7 \pm 0.6}$
Proposed	23.7 ± 0.6	31.0 ± 0.6	57.3 ± 0.8	33.9 ± 0.6	36.2 ± 0.6

Table 1: The fidelity of importance estimators A (the area between LIF and MIF perturbation curves), measured on the ResNet-50 trained on ImageNet with 95% confidence intervals. See the main text for difference between the unsigned and signed importance estimators.

5 Discussion

As this work has been focused on the interpretability and trustworthiness of deep learning models, there still is room for performance improvement. Training the models with FPA leads to a slight decrease in performance compared to training them without augmentation. This might be due to "augment ambiguity" Wei et al. (2020), which occurs when the annotated class is not recognizable anymore as a result of an augmentation. On the other hand, random erasing, which is closely related to FPA, reported an increase in the model performance in certain settings (Zhong et al., 2020). In future work, we plan to explore variations of FPA and training schemes to mitigate the performance loss, but also point out that trading some accuracy for increased interpretability can be worthwhile.

Although FPA has been introduced here for the case of perturbing pixels with constant values it is applicable to any perturbation scheme, e.g. blurring. The only requirement is that the perturbation performed to augment images needs to reflect the perturbation applied during evaluation of interpretability methods.

Since FPA randomly selects pixels to construct augmented training samples it can be expected to only increase a model's robustness against such random perturbations. To increase the robustness against adversarial perturbations, we plan to extend our work by additionally performing adversarial training Goodfellow et al. (2014); Madry et al., which would allow to rule out perturbation artifacts during evaluation with even higher confidence.

Lastly, training with FPA elucidates counter-evidence associated with negative importance scores. Often only the absolute values of importance scores are considered in practice. In contrast, if perturbation artifacts are accounted for, the sign of importance scores can help us understand the model characteristics.

Acknowledgments

This work was funded by the ERA-Net CHIST-ERA grant [CHIST-ERA-19-XAI-007] long term challenges in ICT project INFORM (ID: 93603), by the National Science Centre (NCN) of Poland [2020/02/Y/ST6/00071]. This research was carried out with the support of the Interdisciplinary Centre for Mathematical and Computational Modelling University of Warsaw (ICM UW) under computational allocation no GDM-3540; the NVIDIA Corporation's GPU grant; and the Google Cloud Research Innovators program.

References

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.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

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):e0130140, 2015.

Umang Bhatt, Adrian Weller, and José MF Moura. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European conference on computer vision*, pp. 446–461. Springer, 2014.

Lennart Brocki, Wistan Marchadour, Jonas Maison, Bogdan Badic, Panagiotis Papadimitroulas, Mathieu Hatt, Franck Vermet, and Neo Christopher Chung. Evaluation of importance estimators in deep learning classifiers for computed tomography. In *International Workshop on Explainable, Transparent Autonomous Agents and Multi-Agent Systems*, pp. 3–18. Springer, 2022.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pp. 1383–1391. PMLR, 2020.

Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. *Advances in neural information processing systems*, 30, 2017.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Gintare Karolina Dziugaite, Shai Ben-David, and Daniel M Roy. Enforcing interpretability and its statistical impacts: Trade-offs between accuracy and interpretability. *arXiv preprint arXiv:2010.13764*, 2020.

Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pp. 3429–3437, 2017.

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

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*, pp. 770–778, 2016.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. *Advances in neural information processing systems*, 32, 2019.

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*, pp. 4948–4957, 2019.

Pieter-Jan Kindermans, Kristof T Schütt, Maximilian Alber, Klaus-Robert Müller, Dumitru Erhan, Been Kim, and Sven Dähne. Learning how to explain neural networks: Patternnet and patternattribution. *arXiv preprint arXiv:1705.05598*, 2017.

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

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*.

Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. *arXiv preprint arXiv:2003.08747*, 2020.

Yao Rong, Tobias Leemann, Vadim Borisov, Gjergji Kasneci, and Enkelejda Kasneci. A consistent and efficient evaluation strategy for attribution methods. In *International Conference on Machine Learning*, pp. 18770–18795. PMLR, 2022.

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, 2016.

Wojciech Samek, Grégoire Montavon, Andrea Vedaldi, Lars Kai Hansen, and Klaus-Robert Müller. *Explainable AI: interpreting, explaining and visualizing deep learning*, volume 11700. Springer Nature, 2019.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

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. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pp. 3145–3153. PMLR, 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *In Workshop at International Conference on Learning Representations*. Citeseer, 2014.

Leon Sixt, Maximilian Granz, and Tim Landgraf. When explanations lie: Why many modified bp attributions fail. In *International Conference on Machine Learning*, pp. 9046–9057. PMLR, 2020.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

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

Longhui Wei, An Xiao, Lingxi Xie, Xiaopeng Zhang, Xin Chen, and Qi Tian. Circumventing outliers of autoaugment with knowledge distillation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III*, pp. 608–625. Springer, 2020.

Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 13001–13008, 2020.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

A APPENDIX

A.1 DATASETS AND TRAINING PROCEDURE

We adapted the ResNet-18 architecture to be suitable for the smaller input dimensions of CIFAR-10. The ResNet-50 model was trained on ImageNet for 90 epochs using the SGD optimizer with momentum 0.9, weight decay 10^{-4} and initial learning rate 0.1, which we reduced by a factor of 10 on epochs 30 and 60. On Food101, Resnet-50 was trained for 68 epochs with momentum set to 0.9, weight decay 5×10^{-4} and an initial learning rate of 0.1, where a cosine annealing schedule was used to continuously reduce the learning rate to zero. The ResNet-18 model was trained for 40 epochs using the SGD optimizer with momentum 0.9, weight decay 5×10^{-4} and initial learning rate 0.01, which we reduced by a factor of 10 on epoch 30. For CIFAR-10 and ImageNet, we scaled input images to the range $[-1, 1]$ and for Food101 we performed a z-score normalization with mean and standard deviation from the training set. In all three cases, we performed a horizontal flip with a probability of 0.5 to augment the data.

A.2 IMPORTANCE ESTIMATORS

We compare the following four importance estimators:

Vanilla gradient (VG) (Simonyan et al., 2014): Gradients of the class score S_c with respect to input pixels x_i

$$\mathbf{e} = \frac{\partial S_c}{\partial x}$$

The class score S_c is the activation of the neuron for the predicted class c.

Integrated gradient (IG) (Sundararajan et al., 2017): Average over gradients obtained from inputs interpolated between a reference point x^0 and input x

$$\mathbf{e} = \left(x - x^0\right) \times \sum_{k=1}^{m} \frac{\partial S_c\left(x^0 + \frac{k}{m}\left(x - x^0\right)\right)}{\partial x} \times \frac{1}{m},$$

where x^0 is chosen to be a black image and $m = 200$.

SmoothGrad (SG) (Smilkov et al., 2017): Average over gradients obtained from inputs with injected noise

$$\mathbf{e} = \frac{1}{n} \sum_{1}^{n} \hat{\mathbf{e}}\left(x + \mathcal{N}\left(0, \sigma^2\right)\right),$$

where $\mathcal{N}\left(0, \sigma^2\right)$ is Gaussian noise, $\hat{\mathbf{e}}$ is obtained using vanilla gradient, and $n = 15$.

Squared SmoothGrad (SQ-SG) (Hooker et al., 2019): Variant of SmoothGrad that squares $\hat{\mathbf{e}}$

$$\mathbf{e} = \frac{1}{n} \sum_{1}^{n} \hat{\mathbf{e}}\left(x + \mathcal{N}\left(0, \sigma^2\right)\right)^2.$$

A.3 SUPPLEMENTARY FIGURES AND TABLES

Figure A.1: Examples of feature perturbation augmentation (FPA) applied on *Left:* CIFAR-10 and *Right:* ImageNet.

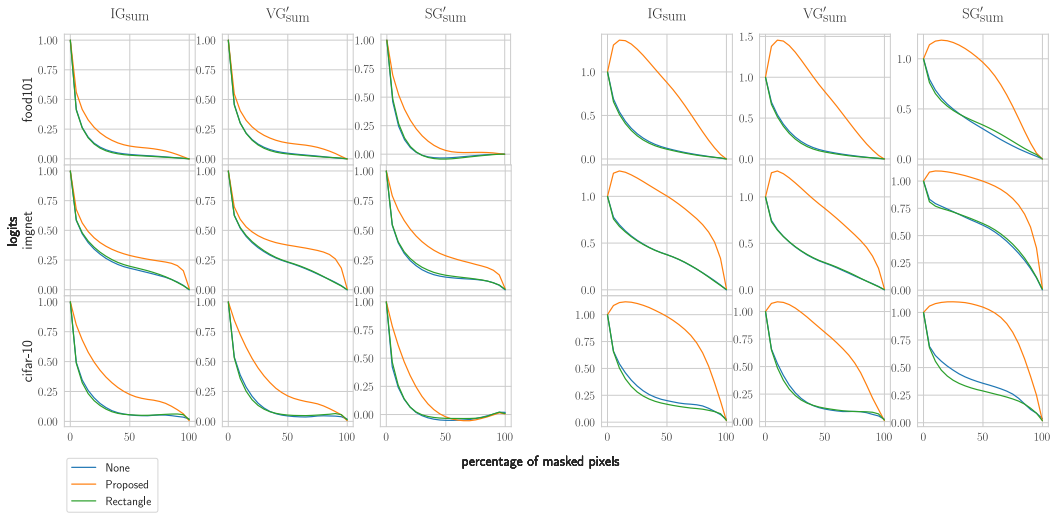

Figure A.2: Overview of perturbation curves for signed importance estimators for all three considered datasets. The raw importance scores are multiplied element-wise with the input image (Shrikumar et al., 2017), indicated by a prime, and then followed by summing over the resulting color channels. By definition, IG includes a multiplication with the input already. The change in normalized logits is measured as a percentage of pixels that are masked according to the Most Important First (MIF; *Left*) and the Least Important First (LIF; *Right*). Note that masking pixels with negative importance scores can increase logits, as counter-evidence for the predicted class is removed from the input image.

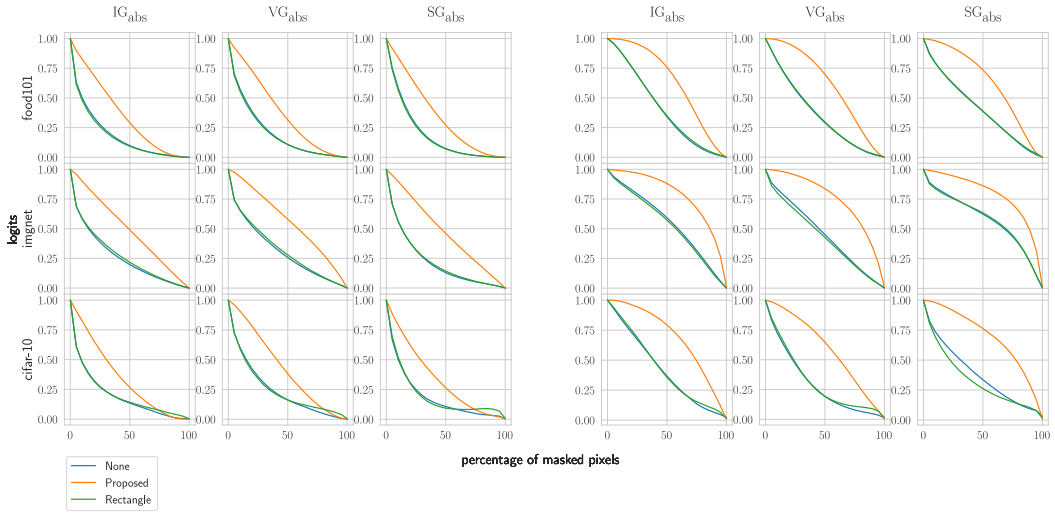

Figure A.3: Overview of perturbation curves for unsigned importance estimators for all three considered datasets. Sums of the absolute values of color channels are used. The change in normalized logits is plotted as a percentage of pixels that are masked according to the Most Important First (MIF; *Left*) and the Least Important First (LIF; *Right*). Notice that in comparison to the LIF perturbation curves from the signed estimators (Figure A.2), the initial increase for LIF perturbation does not occur for the unsigned estimators.

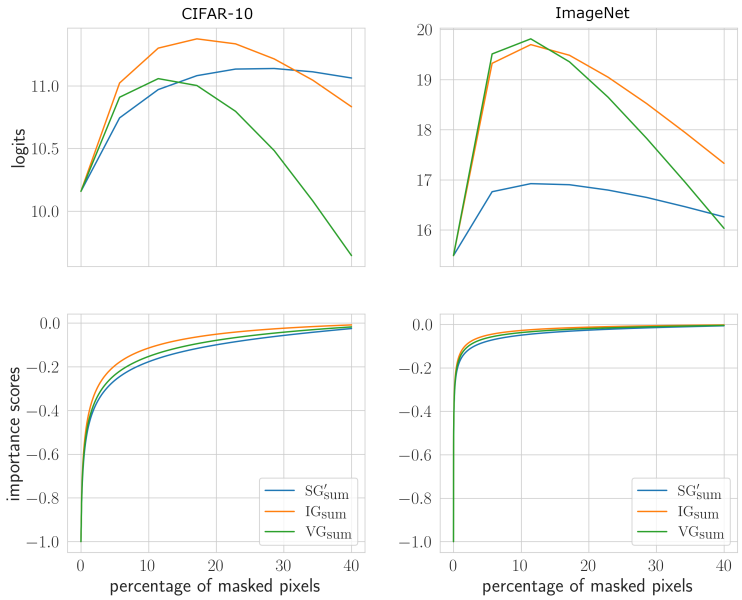

Figure A.4: *Top row:* LIF perturbation curves averaged over CIFAR-10 and ImageNet test samples. *Bottom row:* The curve has been obtained by flattening the heat map and plotting the importance scores in LIF order, with the x-axis indicating their position in the ranking. At each point on the x-axis, one can compare the importance scores of a group of pixels (y-axis labeled 'importance scores') and the influence of their masking on the model output (y-axis labeled 'logits').

Aug.	Random	IG_{sum}	IG_{abs}	VG_{abs}	VG'_{sum}
None	0.0 ± 0.7	14.0 ± 0.8	21.9 ± 0.6	5.6 ± 0.6	8.0 ± 0.7
Proposed	0.0 ± 0.5	59.6 ± 0.7	34.9 ± 0.5	16.9 ± 0.5	47.0 ± 0.7

Aug.	VG'_{abs}	SG_{abs}	SG'_{sum}	SG'_{abs}	$\text{SQ-SG}_{\text{sum}}$
None	16.1 ± 0.6	18.5 ± 0.6	$\mathbf{34.6 \pm 0.8}$	27.0 ± 0.6	27.9 ± 0.6
Proposed	25.8 ± 0.5	35.4 ± 0.5	$\mathbf{80.5 \pm 0.6}$	39.2 ± 0.5	41.1 ± 0.5

Table A.1: The fidelity of importance estimators A (the area between LIF and MIF perturbation curves), measured on the ResNet-18 trained on CIFAR-10 with 95% confidence intervals.

Aug.	Random	IG_{sum}	IG_{abs}	VG_{abs}	VG'_{sum}
None	0.0 ± 0.8	11.6 ± 0.7	22.6 ± 0.8	15.5 ± 0.8	8.7 ± 0.6
Proposed	0.0 ± 0.6	67.7 ± 1.0	29.4 ± 0.6	24.8 ± 0.7	66.0 ± 1.0

Aug.	VG'_{abs}	SG_{abs}	SG'_{sum}	SG'_{abs}	$\text{SQ-SG}_{\text{sum}}$
None	15.5 ± 0.8	22.2 ± 0.8	30.3 ± 0.8	23.7 ± 0.7	$\mathbf{31.2 \pm 0.8}$
Proposed	24.8 ± 0.7	28.9 ± 0.7	$\mathbf{69.4 \pm 0.9}$	27.0 ± 0.6	31.3 ± 0.7

Table A.2: The fidelity of importance estimators A (the area between LIF and MIF perturbation curves), measured on the ResNet-50 trained on Food101 with 95% confidence intervals.