Robustness of Visual Explanations to Common Data Augmentation Methods

Lenka Tětková lenhy@dtu.dk
Lars Kai Hansen lkai@dtu.dk
Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, 321, 2800 Kgs. Lyngby, Denmark

Abstract

As the use of deep neural networks continues to grow, understanding their behaviour has become more crucial than ever. Post-hoc explainability methods are a potential solution, but their reliability is being called into question. Our research investigates the response of post-hoc visual explanations to naturally occurring transformations, often referred to as augmentations. We anticipate explanations to be invariant under certain transformations, such as changes to the colour map while responding in an equivariant manner to transformations like translation, object scaling, and rotation. We have found remarkable differences in robustness depending on the type of transformation, with some explainability methods (such as LRP composites and Guided Backprop) being more stable than others. We also explore the role of training with data augmentation. We provide evidence that explanations are typically less robust to augmentation than classification performance, regardless of whether data augmentation is used in training or not.

1. Introduction

Convolutional neural networks (CNNs) are commonly used in computer vision. However, CNNs are fragile to adversarial attacks [11]. It has been shown that explanation methods are fragile as well and that attackers can manipulate the explanations arbitrarily [9, 10].

To be trusted, explanations need to show common-sense behaviour. In this work, we investigate one such basic behaviour: *If a transformation of an image does not change the target class, the explanation should assign importance to the same part of the object as in the untransformed image*[1]. If the explainability method does not preserve the explanations of the perturbed images, we lose trust in it. We believe that it is even more concerning than adversarial attacks since perturbations such as *e.g.*, object rotation, are omnipresent and happen spontaneously.

In this work, we investigate how perturbations of an image influence visual post-hoc explanations. To understand the role of augmentation during training, we train CNNs from scratch on both augmented and non-augmented data. We examine the robustness of the models and compare the explanations. We pose the questions: Are visual explanations as robust to augmenting the input image as the predictions? Are there differences among various explainability methods and model architectures? Does training with augmented images improve the robustness? Which explainability methods are the best both in robustness to small augmentations and in faithfulness measured by the pixel-flipping test?

Related work The feasibility of adversarial attacks [2] is well-known. It has been shown [9, 10] that explanation methods are fragile as well and that attackers can manipulate the explanations arbitrarily. In this paper, we focus on the fragility of the explanations in the case of more naturally occurring (often unintentional) disruptions.

Data augmentation techniques [7, 34] have been used to improve the generalization of the image classifiers (*e.g.*, [24, 31]). Rebuffi *et al.* [21] found that using data augmentations helps to improve the robustness against adversarial attacks. Very recent work by Won *et al.* [33] found that data augmentation used under model training has an impact on model interpretability, however, they do not consider stability under test time augmentation as in the present work.

Wang and Wang [32] built a model with transformation invariant interpretations. However, this self-interpretable model violates one of the desiderata for explanations [29]: low construction overhead. We explore whether we could get similar robustness with available post-hoc explainability methods. Moreover, we broaden the set of considered transformations.

Although explainability is important for understanding

[1] We do not consider cases where the transformation of an image would change the ground-truth label.

neural networks, the existing methods differ in the quality of produced explanations and many saliency methods have been criticized (*e.g.*, [1, 15, 20]). Therefore, metrics to evaluate the quality have been developed (*e.g.*, [5, 6, 22]). Quantus [13] is a toolkit that collects many of those metrics. Our experiments shed further light on the stability of explainability methods.

2. Methods

Augmentation methods Here we divide augmentation techniques into two groups: invariant and equivariant methods. For invariant techniques, the explanation of the augmented image should be the same as the explanation of the original image. In the case of the equivariant techniques, the explanation of the augmented image should be the same as the augmented explanation of the original image.

We chose three invariant (change of brightness, hue and saturation) and three equivariant techniques (rotation, translation and scaling). When using equivariant methods, the background of each image was padded with black pixels to match the original image format if necessary. In preliminary experiments, we studied the influence of various background padding methods and differences were negligible.

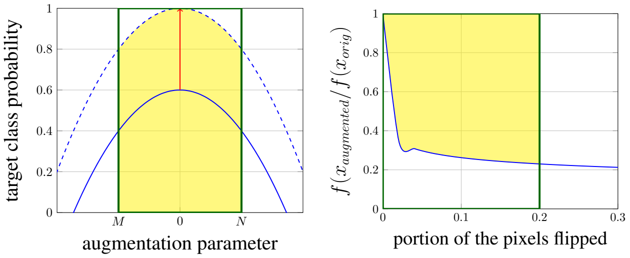
We used the library ImgAug [14] for augmenting images. For each method, we chose an interval of values so that classification performance was reduced by 10%. A table showing the chosen intervals for each method can be found in Appendix B and Fig. 3 displays one image augmented by values within these intervals for changing brightness and rotation. The figures for the rest of the methods are in Appendix B.

The experiments were performed on the ImageNet [8] dataset. For comparing explanations, 500 images across all ImageNet classes were randomly selected. Analyses were done on the correctly classified images. For every augmentation method and for each image, the interval of possible values of the augmentation method parameter was divided into equidistant units and augmented versions of the image were created, one for each of these values. Each image was passed through the networks to get the probability of the target class and we got explanations for post-hoc explainability methods. We computed the Pearson correlation between the explanations of the augmented images and the explanation of the original image (augmented explanation in the case of the equivariant methods) and top-1000 intersection (intersection of 1000 most important pixels in the explanation). We compared only the area of the original image – hence, in equivariant methods, we computed the correlation and top-1000 intersection only on the parts that were present in both the original and the augmented image and mask the rest.

Metrics We can plot the probability of the target class and all its augmented versions with the augmentation parameter on the x-axis and the probability on the y-axis. We call this relation a *probability curve*. In the same way, we plot the correlations between the original and the augmented images (call it *correlation curve*) and the top-1000 intersections (*top-1000 curve*). These curves can be visualised as in Fig. 2. To compare explainability methods in a fair way, we score relative to classification certainty. For a fixed range $[M, N]$, we compute a normalized area under the response curve for $x \in [M, N]$, or, more precisely, the portion of this area out of a rectangle with corners $[M, 0]$, $[N, 0]$, $[N, 1]$, $[M, 1]$. Moreover, to be able to compare the scores of different curves and let the score depend only on the shape of the curve, we ensure that the point on the curve corresponding to the zero-augmented image takes a value of 1 by shifting the response curve. Figure 1a illustrates how the score is computed. For each curve, we get a number between 0 and 1 and higher values indicate a more stable response. Finally, since we want to compare the robustness of the model's predictions and its explanations, we divide the score for the correlation (or top-1000) curve of explanations by the score for the probability curve and denote it as *S(corelation, probability)* (or *S(top-1000, probability)*). If S(·, probability) is smaller than 1, it means that the predictions are more stable than the explanations, whereas values higher than 1 entails more robust explanations. The intervals for augmentation parameters are chosen such that the probability of the target class drops on average by at least 10% at one of the endpoints (in comparison to the original image).

Apart from comparing the robustness of the explainability methods, we are interested in the overall quality of explanations. One method for evaluating the quality is pixel flipping [5]. We consider only the original and correctly classified images. We flip the most relevant pixels first and replace them with black pixels. For each perturbed image, we divide its probability of the target class by the original image's probability of the target class and plot these values as a curve by linear interpolation. We compute the normalized area over the curve (up to 1) from zero to the first 20% pixels flipped and average these numbers across all images. Figure 1b visualizes how the pixel flipping score is computed. A similar definition has been given by Samek *et al.* [23]. Our definition differs in dividing the probabilities instead of subtracting them. The fractions better capture the relative decline of the probability and can take all values in $[0, 1]$.

Networks We study three convolutional networks (ResNet50 [12], VGG16 [26], and EfficientNetV2 small [30]). Because of space constraints, we present in this paper only the results for ResNet50. However, the results for VGG16 and EfficientNet V2 small show



(a) Robustness score

(b) Pixel flipping

Figure 1. Visualization of the metrics defined in Sec. 2. In both cases, we compute the portion of the yellow part in the green rectangle.

similar tendencies. Since we wanted to explore the role of augmenting images during training, we trained each model architecture with two different settings. Models trained with fully augmented data (denoted "full aug" in the following) were trained with Trivial Augment wide [19] strategy. Models trained with limited data augmentation (denoted "lim aug") used only random resized cropping, random horizontal flipping and random erasing [36] (only EfficientNet V2 and ResNet50). Details on training can be found in Appendix A.

Explanation methods We investigated the following explanation methods: Gradients [25], Input x Gradients [25], Integrated Gradients [28], Guided Backpropagation [27], Deconvolution [35] and three variants of Layer-wise Relevance Propagation [4, 16] composites: EpsilonPlusFlat (LRP-ε-rule for dense layers, LRP-α, β ($\alpha = 1, \beta = 0$), also called ZPlus rule, for convolutional layers, and the flat rule for the first linear layer), EpsilonGammaBox (LRP-ε-rule for dense layers, the LRP-γ-rule ($\gamma = 0.25$) for convolutional layers, and the LRP-Z^B-rule (or box-rule) for the first layer) and EpsilonAlpha2Beta1Flat (LRP-ε-rule for dense layers, LRP-α, β ($\alpha = 2, \beta = 1$) for convolutional layers and the flat rule for the first linear layer) [18]. We used Zennit [3] to generate LRP explanations and Captum [17] for the rest of the explainability methods.

The code and hyperparameters for reproducing the experiments can be found in the project repository [2].

3. Results

Figure 2 shows the probability and correlation curves for rotation and "ResNet50 full aug". It shows that, although the predictions do not change much for increasing magnitudes of augmentation, the drop in correlation is huge. Table 1 shows S(corelation, probability) for all augmentation and explainability methods tested on "ResNet50 full aug". We observe that the explanations are in most cases less stable than the predictions. Moreover, the robustness of explanations depends on the augmentation method – for some of them, the explanations are more robust than for others. Specifically, explanations of images augmented by invariant methods are more stable than the ones augmented by equivariant methods. The variance in robustness across explainability methods was an unexpected finding. The most stable ones, the composites of LRP and Guided Backprop, indicate a certain degree of stability, whereas the least stable ones,

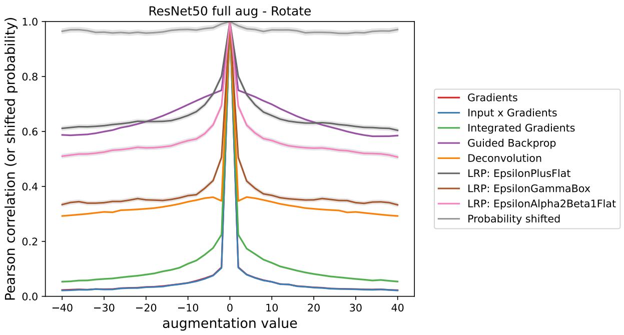
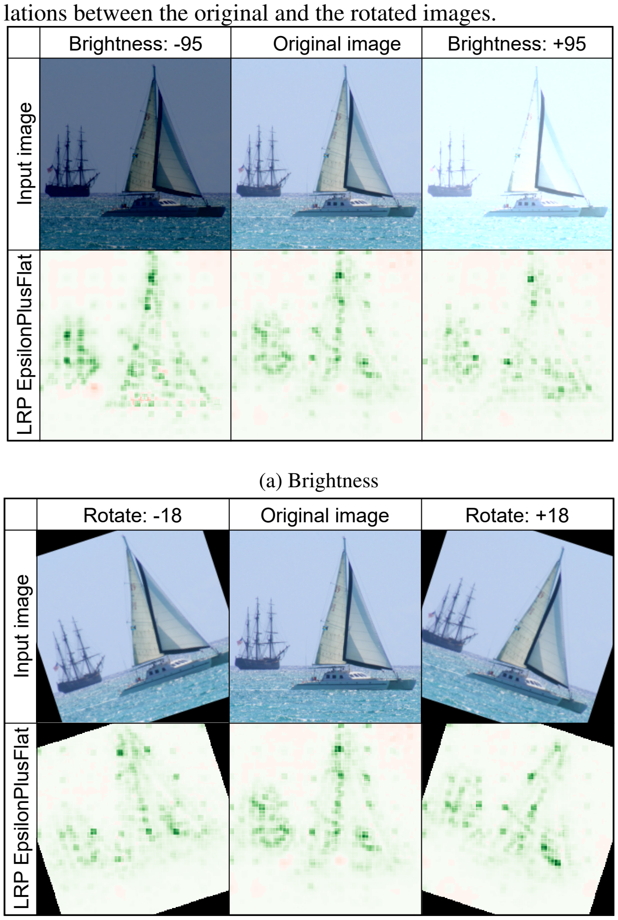


Figure 2. Example of curves showing the probabilities and correlations between the original and the rotated images.



(a) Brightness

(b) Rotate

Figure 3. Examples of the augmented images and their explanations.

[2] https://github.com/LenkaTetkova/robustness-of-explanations.git

	Brightness	Hue	Saturation	Rotate	Scale	Translate
Gradients	0.468	0.442	0.354	0.127	0.122	0.246
Input x Gradients	0.330	0.443	0.343	0.126	0.120	0.245
Integrated Gradients	0.478	0.636	0.546	0.209	0.229	0.327
Guided Backprop	**1.005**	1.028	0.994	**0.819**	**0.866**	**0.875**
Deconvolution	0.975	1.014	0.975	0.434	0.437	0.449
LRP: EpsilonPlusFlat	0.923	**1.053**	**1.038**	0.796	0.834	0.792
LRP: EpsilonGammaBox	0.632	0.856	0.832	0.480	0.512	0.532
LRP: EpsilonAlpha2Beta1Flat	0.662	1.006	0.972	0.691	0.722	0.706

Table 1. Results of S(correlation, probability) for ”ResNet50 full aug”, computed on 391 (correctly classified) images. All numbers are with uncertainty (standard error of the mean) at most ±0.007. Highlighted are the highest values for each augmentation.

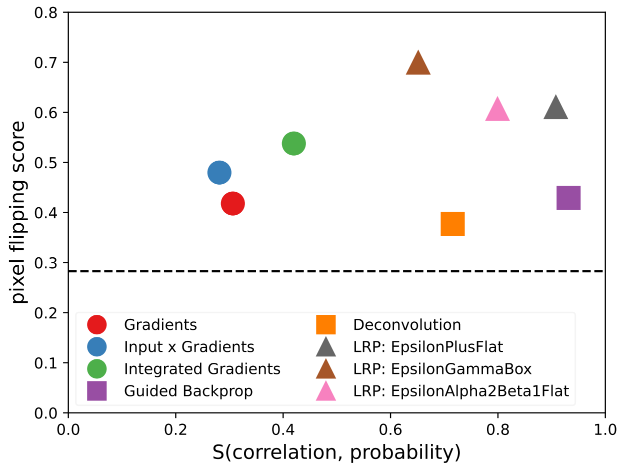


Figure 4. Comparison of S(corelation, probability) and pixel flipping score for ”ResNet50 full aug”. The scores are defined in Sec. 2. The x-axis shows the average of the S(corelation, probability) for all six augmentation methods used in this paper. The dashed line corresponds to a baseline pixel-flipping score computed with random sorting of the pixels. The best methods are in the top right corner.

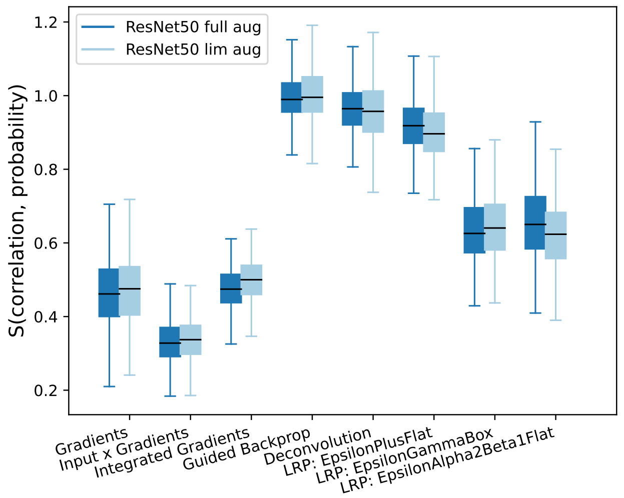


Figure 5. Comparison of ”ResNet50 full aug” (391 images) and ”ResNet50 lim aug” (385 images) for each explainability method. We plot S(corelation, probability) for changes in brightness (AddToBrightness from -95 to 95). Boxes show the quartiles and medians, and whiskers extend to the most extreme, non-outlier data points.

Gradients and Gradients x Inputs, show a steep decrease in the similarity of explanations even for small perturbations. Figure 5 depicts the comparison of ResNet50 trained with full and limited augmentations evaluated on the changes in brightness. We can observe negligible differences between both networks. Therefore, it indicates that training with data augmentations does not diminish this problem. Additional plots for other augmentation methods can be found in Appendix C.

However, stability is not the only desired property of explainability methods. We need to consider also their overall quality. In our experiments, we measured faithfulness, specifically pixel flipping score. Figure 4 shows S(corelation, probability) against the pixel-flipping scores. We can observe that all the LRP composites lie in the top-right corner. On the other hand, Guided Backprop and Deconvolution attain low pixel-flipping scores in comparison to other methods. This is not surprising because Nie *et al.* [20] showed that these two methods do not depend much on the tested model but rather perform a (partial) image recovery.

We consider the instability of explanations to be a serious problem that is relevant for many domains where computer vision tasks are solved using neural networks. Our study contributes additional evidence that current explainability methods cannot be trusted to deliver a reliable justification of the outputs of a model. Many of the tested perturbations may occur unintentionally when taking images under different light conditions, from a different angle or by domain shift and variability of the data. Unless more stability of explainability methods is ensured, explanations cannot be trusted and used as a foundation for authorizing neural networks with important tasks with significant impact.

4. Conclusion

We investigated the robustness of post-hoc explainability methods under natural perturbations of the input images. We found out that LRP composites and Guided Backprop produce the most stable explanations and Gradients and Input x Gradients are the least stable ones. When perturbing with the invariant methods (*e.g*., changing brightness, hue and saturation), the explanations are more stable than when perturbing with equivariant methods (*e.g*., rotation, scaling and translation). Training with data augmentation does not reduce this problem.

5. Acknowledgements

This work was supported by the DIREC Bridge project Deep Learning and Automation of Imaging-Based Quality of Seeds and Grains, Innovation Fund Denmark grant number 9142-00001B and by the Danish Pioneer Centre for AI, DNRF grant number P1. We acknowledge EuroHPC Joint Undertaking for awarding us access to Karolina at IT4Innovations, Czech Republic.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 1

[3] Christopher J. Anders, David Neumann, Wojciech Samek, Klaus-Robert Müller, and Sebastian Lapuschkin. Software for dataset-wide xai: From local explanations to global insights with Zennit, CoRelAy, and ViRelAy. *CoRR*, abs/2106.13200, 2021. 3

[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):e0130140, 2015. 3

[5] 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. 2

[6] Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Somesh Jha, and Xi Wu. Concise explanations of neural networks using adversarial training. (arXiv:1810.06583), Jul 2020. arXiv:1810.06583 [cs, stat]. 2

[7] Dan Claudiu Cireşan, Ueli Meier, Luca Maria Gambardella, and Jürgen Schmidhuber. Deep, big, simple neural nets for handwritten digit recognition. *Neural Computation*, 22(12):3207–3220, dec 2010. 1

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 2

[9] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. Explanations can be manipulated and geometry is to blame. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[10] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3681–3688, 2019. 1

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[12] 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. 2

[13] Anna Hedström, Leander Weber, Daniel Krakowczyk, Dilyara Bareeva, Franz Motzkus, Wojciech Samek, Sebastian Lapuschkin, and Marina Marina M.-C. Höhne. Quantus: An explainable ai toolkit for responsible evaluation of neural network explanations and beyond. *Journal of Machine Learning Research*, 24(34):1–11, 2023. 2

[14] Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. imgaug. https://github.com/aleju/imgaug, 2020. Online; accessed 01-Feb-2020. 2

[15] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un)reliability of saliency methods. (arXiv:1711.00867), Nov 2017. arXiv:1711.00867 [cs, stat]. 2

[16] Maximilian Kohlbrenner, Alexander Bauer, Shinichi Nakajima, Alexander Binder, Wojciech Samek, and Sebastian Lapuschkin. Towards best practice in explaining neural network decisions with lrp. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020. 3

[17] Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. Captum: A unified and generic model interpretability library for pytorch, 2020. 3

[18] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. *Layer-Wise Relevance Propagation: An Overview*, page 193–209. Sep 2019. journalAbbreviation: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 3

[19] Samuel G. Müller and Frank Hutter. Trivialaugment: Tuning-free yet state-of-the-art data augmentation. page 754–762. IEEE Computer Society, Oct 2021. 3

[20] Weili Nie, Yang Zhang, and Ankit Patel. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. (arXiv:1805.07039), Feb 2020. arXiv:1805.07039 [cs]. 2, 4

[21] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021. 1

[22] Laura Rieger and Lars Kai Hansen. Irof: a low resource evaluation metric for explanation methods. (arXiv:2003.08747), Mar 2020. arXiv:2003.08747 [cs]. 2

[23] Wojciech Samek, Alexander Binder, Gregoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Muller. Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2660–2673, Nov 2017. 2

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

[25] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2

[27] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 3

[28] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017. 3

[29] William Swartout and Johanna Moore. Explanation in second generation expert systems. page 543–585, Jan 1993. 1

[30] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International Conference on Machine Learning*, pages 10096–10106. PMLR, 2021. 2

[31] Jason Wang, Luis Perez, et al. The effectiveness of data augmentation in image classification using deep learning. *Convolutional Neural Networks Vis. Recognit*, 11(2017):1–8, 2017. 1

[32] Yipei Wang and Xiaoqian Wang. Self-interpretable model with transformation equivariant interpretation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 1

[33] Soyoun Won, Sung-Ho Bae, and Seong Tae Kim. Analyzing effects of mixed sample data augmentation on model interpretability. (arXiv:2303.14608), Mar 2023. arXiv:2303.14608 [cs]. 1

[34] Larry Yaeger, Richard Lyon, and Brandyn Webb. Effective training of a neural network character classifier for word recognition. In M.C. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. 1

[35] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3

[36] 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, pages 13001–13008, 2020. 3