Metrics for Saliency Map Evaluation of Deep Learning Explanation Methods

Tristan Gomez[1](✉), Thomas Fréour[2], and Harold Mouchère[1]

[1] Nantes Université, Centrale Nantes, CNRS, LS2N, 44000 Nantes, France
{tristan.gomez,harold.mouchere}@univ-nantes.fr

[2] CRTI, Inserm UMR 1064, Nantes University Hospital Inserm, 44000 Nantes, France
thomas.freour@chu-nantes.fr

Abstract. Due to the black-box nature of deep learning models, there is a recent development of solutions for visual explanations of CNNs. Given the high cost of user studies, metrics are necessary to compare and evaluate these different methods. In this paper, we critically analyze the Deletion Area Under Curve (DAUC) and Insertion Area Under Curve (IAUC) metrics proposed by Petsiuk et al. (2018). These metrics were designed to evaluate the faithfulness of saliency maps generated by generic methods such as Grad-CAM or RISE. First, we show that the actual saliency score values given by the saliency map are ignored as only the ranking of the scores is taken into account. This shows that these metrics are insufficient by themselves, as the visual appearance of a saliency map can change significantly without the ranking of the scores being modified. Secondly, we argue that during the computation of DAUC and IAUC, the model is presented with images that are out of the training distribution which might lead to unexpected behavior of the model being explained. To complement DAUC/IAUC, we propose new metrics that quantify the sparsity and the calibration of explanation methods, two previously unstudied properties. Finally, we give general remarks about the metrics studied in this paper and discuss how to evaluate them in a user study.

Keywords: Interpretable machine learning · Objective evaluation · Saliency maps

1 Introduction

Recent years have seen a surge of interest in interpretable machine learning, as many state-of-the-art learning models currently are deep models and suffer from their lack of interpretability due to their black-box nature. In image classification, many generic approaches have been proposed to explain a model's decision by generating saliency maps that highlight the important areas of the image concerning the task at hand [1,3,14,16,20,21,23,27]. The community of

Supported by Nantes Excellence Trajectory (NExT).

M. El Yacoubi et al. (Eds.): ICPRAI 2022, LNCS 13363, pp. 84–95, 2022.
https://doi.org/10.1007/978-3-031-09037-0_8

interpretable deep learning has yet to find a consensus about how to evaluate these methods, the main difficulty residing in the ambiguity of the concept of interpretability. Indeed, depending on the application context, the users' requirements in terms of interpretability may vary a lot, making it difficult to find a universal evaluation protocol.

This has started a trend in literature where authors confront users with models' decisions along with explanations to determine the users' preference on a particular application [2,24,25]. The main issues of this approach are its financial cost and the difficulty to establish a correct protocol, which mainly comes from the requirement to design an experiment whose results will help understand the users' needs and also from the fact that most machine learning researchers are not used to run experiments involving humans.

Because of these issues, another trend proposes to design objective metrics to evaluate generic explanation methods [3,14,20]. In this paper, we chose to follow this trend, by proposing three new metrics.

We focus our work on the DAUC and IAUC metrics proposed by [20]. First, we study several aspects of these metrics and we show that the actual saliency score values given by the saliency map are ignored as they only take into account the ranking of the scores. This shows that these metrics are insufficient by themselves, as the visual appearance of a saliency map can change significantly without the ranking of the scores being modified. We also argue that during the computation of DAUC and IAUC, the model is presented with images that are out of the training distribution which might lead to unexpected behavior of the model and of the method used to generate the saliency maps. We then introduce a new metric called Sparsity, which quantifies the sparsity of a saliency map, a property that is ignored by previous work. Another property that was not studied until now is the calibration of the saliency maps. Given it could be a useful property for interpretability, we also propose two new metrics to quantify it, namely Deletion Correlation (DC) and Insertion Correlation (IC). Finally, we give general remarks about all the metrics studied in this paper and discuss how to evaluate these metrics in a user study.

2 Existing Metrics

Various metrics have been proposed to automatically evaluate saliency maps generated by explanation methods [3,14,20]. These metrics consist to add or remove the important areas according to the saliency map and measure the impact on the initially predicted class score. For example, Chattopadhay et al. proposed "increase in confidence" (IIC) and "average drop" (AD) [3]. These metrics consist to multiply the input image with an explanation map to mask the non-relevant areas and to measure the class score variation. Jung et al. proposed a variant of AD where the salient areas are masked instead of the non-salient, called Average Drop in Deletion (ADD) [14]. In parallel, Petsiuk et al. proposed DAUC and IAUC which study the score variation while progressively masking/revealing the image instead of applying the saliency map once [20].

Given the similarity of these metrics, we will focus our study on DAUC and IAUC, which we will now describe.

2.1 DAUC and IAUC

To evaluate the reliability of the proposed attention mechanism, Petsiuk et al. proposed the Deletion Area Under Curve (DAUC) and Integration Area Under Curve (IAUC) metrics [20]. These metrics evaluate the reliability of the saliency maps by progressively masking/revealing the image starting with the most important areas according to the saliency map and finishing with the least important.

The input image is a 3D tensor $I \in \mathbb{R}^{H \times W \times 3}$ and the saliency map is a 2D matrix $S \in \mathbb{R}^{H' \times W'}$ with a lower resolution, $H' < H$ and $W' < W$. First, S is sorted and parsed from the highest element to its lowest element. At each element $S_{i'j'}$, we mask the corresponding area of I by multiplying it by a mask $M^k \in \mathbb{R}^{H \times W}$, where

$$M^k_{ij} = \begin{cases} 0, & \text{if } i' + r < i < i' + 2r \text{ and } j' + r < j < j' + 2r \\ 1, & \text{otherwise,} \end{cases} \tag{1}$$

where $r = H/H' = W/W'$. After each masking operation, the model m runs an inference with the updated version of I, and the score of the initially predicted class is updated, producing a new score c_k :

$$c_k = m(I \cdot \prod_{\tilde{k}=1}^{\tilde{k}=k} M^{\tilde{k}}), \tag{2}$$

where $k \in \{1, ..., H' \times W'\}$. Examples of input images obtained during this operation can be seen in Fig. 1. Secondly, once the whole image has been masked, the scores c_k are normalized by dividing them by the maximum $\underset{k}{max}\ c_k$ and then plotted as a function of the proportion p_k of the image that is masked. The DAUC is finally obtained by computing the area under the curve (AUC) of this function. The intuition behind this is that if a saliency map highlights the areas that are relevant to the decision, masking them will result in a large decrease of the initially predicted class score, which in turn will minimize the AUC. Therefore, minimizing this metric corresponds to an improvement.

Instead of progressively masking the image, the IAUC metric starts from a blurred image and then progressively unblurs it by starting from the most important areas according to the saliency map. Similarly, if the areas highlighted by the map are relevant for predicting the correct category, the score of the corresponding class (obtained using the partially unblurred image) is supposed to increase rapidly. Conversely, maximizing this metric corresponds to an improvement.

2.2 Limitations

DAUC and IAUC Generate Out of Distribution (OOD) Images. When progressively masking/unblurring the input image, the model is presented with samples that can be considered out of the training distribution, as shown in Fig. 1.

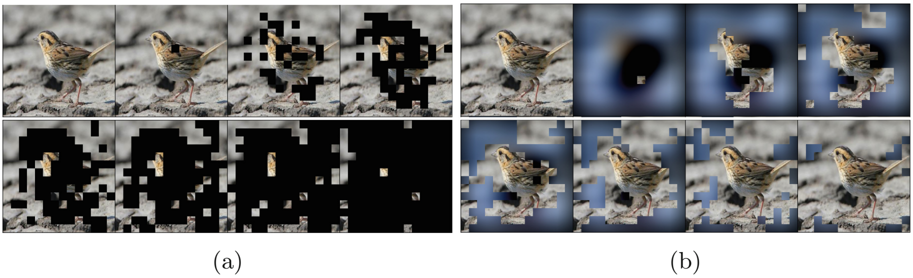
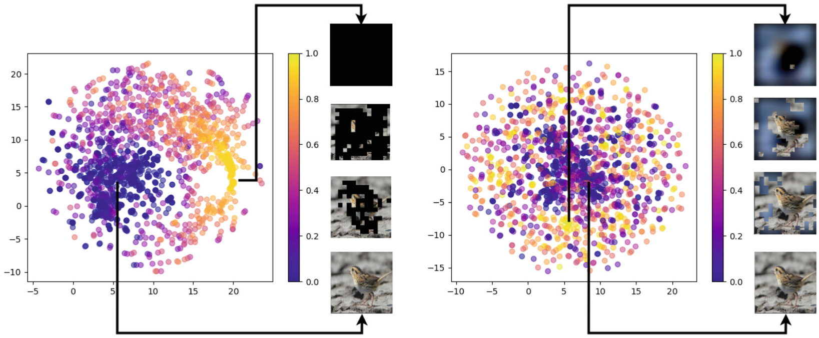


Fig. 1. Examples of images passed to the model during the computation of (a) DAUC and (b) IAUC. Masking and blurring the input images probably lead to OOD samples.

Indeed, the kind of distortions produced by the masking/blurring operations do not exist naturally in the dataset and are different from the kind produced by the standard data augmentations like random crop, horizontal flip, and color jitter, meaning that the model has not learned to process images with such distortions. Therefore, the distribution of the images presented to the model is different from the one met during training. However, it has been documented that CNNs and more generally deep learning models have poor generalization outside of the training distribution [8]. This shows that DAUC and IAUC may not reflect the faithfulness of explanation methods as they are based on a behavior of the model that is different from that encountered when facing training distribution (e.g. during the test phase).

To verify this hypothesis we visualize the UMAP [17] projections of the representations of 100 masked/blurred samples obtained during the computation of DAUC and IAUC on the CUB-200-2011 dataset [26]. We also added the representation of 500 unmodified test images (in blue) to visualize the training distribution. The model used is a ResNet50 [11] on which we applied Grad-CAM++ [3]. Figure 2 shows that, during computation of DAUC, the representations gradually converge towards a unique point, which is not surprising as, at the end of the computation, all images are fully masked, i.e. plain black. However even when only a proportion of 0.4 of the image is masked, the corresponding representation is distant from the blue point cloud indicating the training distribution. A similar phenomenon happens with IAUC, where blurring the image causes the representation to move away from the training distribution. This experiment demonstrates that the DAUC and IAUC metrics indeed present OOD samples, which might lead to unexpected behavior of the model and of the method used to generate the explanation maps. However, as suspected by [20], the blurring operation seems to create samples that are less far from the training distribution compared to the masking operation, probably because a blurred image still contains the low-frequency parts of the original image. Another explanation is that most current classification models are designed with the assumption that an input image contains an object to recognize, which is in contradiction with DAUC and IAUC as they consist to remove the object to recognize from the image. This suggests that modifying these metrics in such a way as to always leave an object to recognize in the input image would solve this issue.

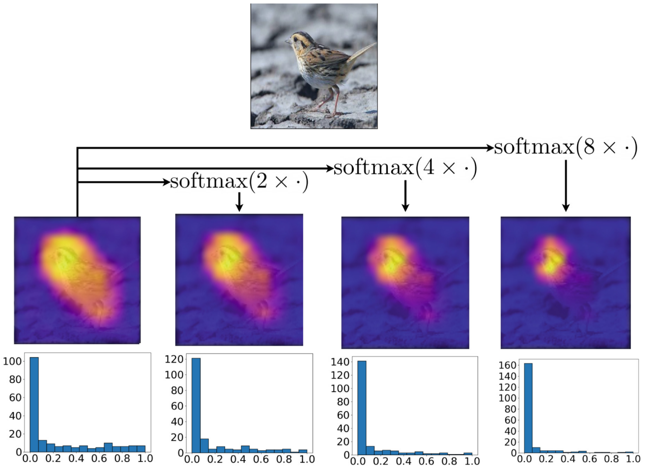


(a) (b)

Fig. 2. UMAP projection of representations obtained while computing (a) DAUC and (b) IAUC on 100 images. The color indicates the proportion of the image that is masked/unblurred. The model used is a ResNet50 on which we applied Grad-CAM++ on the CUB-200-2011 dataset. We also plotted representations from 500 points of the test set to visualize the training distribution (in blue). By gradually masking the image, the representations converge towards a point (in yellow) that is distant from the points corresponding to unmasked images (in blue). Similarly, blurring the image causes the representation to move away from the training distribution. This shows that masking/blurring indeed creates OOD samples. (Color figure online)

DAUC and IAUC Only Take the Pixel Score Rank into Account. When computing DAUC and IAUC, the saliency map is used only to determine in which order to mask/reveal the input image. Hence, only the ranking of the saliency scores S_{ij} is used to determine in which order to mask the image, leaving the actual values of the scores ignored.

However, pixel ranking is not the only characteristic that should be taken into account, as the visual appearance can vastly vary between two attention maps without changing the ranking. Figure 3 shows examples of a saliency map with various score distributions artificially modified. We used a saliency map produced by the Score-CAM [27] explanation method and altered its score distribution by multiplying all values by a coefficient followed by the application of a softmax function. By increasing the coefficient we alter the visual appearance of the maps, without changing the pixel ranking, which maintains the same DAUC and IAUC scores. This illustrates the fact that DAUC and IAUC ignore the score dynamic of the saliency map, which can vastly affect the visual appearance. To complement DAUC and IAUC, we propose new metrics that take the score values into account in the following section.



(a)

Model	Viz. Method	Transformation	DAUC	IAUC	Sparsity	DC	IC
ResNet50	Score-CAM	None	0.012	0.52	3.98	0.187	0.22
		softmax(2 × ·)			5.96	0.24	0.19
		softmax(4 × ·)			8.96	0.29	0.14
		softmax(8 × ·)			16.52	0.38	0.04

(b)

Fig. 3. Examples of saliency maps obtained by artificially sparsifying the saliency scores. The original saliency map is generated using Score-CAM applied on a ResNet50 model tested on the CUB-200-2011 dataset. Despite having different visual appearances, the four maps have the same DAUC and IAUC metric values because these metrics ignore the score values and only take into account the ranking of the scores. On the other hand, the Sparsity metric depends on the score distribution and reflects the amount of focus of the map. In this figure, only the saliency map is modified, the decision process is left unchanged.

3 Score Aware Metrics

As mentioned in the previous paragraph, DAUC and IAUC ignore the actual score values and only take into account the saliency score S_{ij} ranking. To complement these metrics, we propose three new metrics, namely Sparsity, Deletion Correlation (DC), and Insertion Correlation (IC).

3.1 The Sparsity Metric

An important visual aspect of saliency maps that has not been studied until now by the community is what we call sparsity. As shown by Fig. 3, Saliency maps can be more or less focused on a specific point depending on the score distribution, without changing the score ranking. This aspect could impact the interpretability of the method, as it significantly changes the visual aspect of the map and therefore could also affect the perception of the user. For example, one could argue that a high sparsity value implies a map with a precise focus that highlights only a few elements of the input image, making it easier to understand for humans. The Sparsity metric is defined as follows:

$$\text{Sparsity} = \frac{S_{max}}{S_{mean}} \tag{3}$$

where S_{max} and S_{mean} are respectively the maximum and mean score of the saliency map S. Note that the saliency methods available in the literature generate saliency maps with scores that are comprised in a large number of ranges. Therefore, the map should be first normalized as follows:

$$S' = \frac{S - S_{min}}{S_{max} - S_{min}} \tag{4}$$

This means that, after normalization, $S'_{max} = 1$ and Eq. (3) can be simplified to

$$\text{Sparsity} = \frac{1}{S'_{mean}} \tag{5}$$

A high sparsity value means a high S_{max}/S_{mean} ratio, i.e., a low mean score S_{mean} which indicates that the map's activated areas are narrow and focused. As shown by Fig. 3b, this metric is indeed sensitive to the actual saliency scores values and reflects the various amount of focus observed in the saliency maps.

3.2 The DC and IC Metrics

As previously mentioned, the DAUC and IAUC metrics ignore the score values of the saliency maps and only take into account the ranking of the scores. This means that these metrics ignore the sparsity of the map, which is why we proposed to quantify this aspect. Another potentially interesting property of saliency maps is the calibration. The concept of calibration has seen a recent surge of interest in the deep learning community [10,19,28], but previous work focused exclusively on calibrating prediction scores. A pixel S_{ij} from a well-calibrated saliency map S would reflect through its luminosity the importance it has on the class score. More precisely, we say that an explanatory map S is perfectly calibrated if for any two elements S_{ij} and $S_{i'j'}$, we have $S_{ij}/S_{i'j'} = v/v'$, where

v and v' are respectively the impact of S_{ij} and $S_{i'j'}$ on the class score. To evaluate this, we propose to quantify how correlated the saliency scores and their corresponding impact on the class score are. To the best of our knowledge, this is the first time that an objective metric is proposed to measure the calibration of explanation methods. In practice, such a metric could be used in a user study to evaluate to what extent the calibration property is useful.

We take inspiration from the DAUC and IAUC metrics and propose to gradually mask/reveal the input image by following the order suggested by the saliency map, but instead of computing the area under the class score vs. pixel rank curve, we compute the linear correlation of the class score variations and the saliency scores. The correlation measured when masking the image is called Deletion Correlation (DC) and the one measured when revealing the image is called Insertion Correlation (IC). The following paragraph details the computation of these two metrics.

DC is computed using the same progressive masking and inference method as DAUC. Once the scores c_k have been computed, we compute the variation of the scores $v_k = c_k - c_{k+1}$. Finally, we compute the linear correlation between the v_k and the s_k where s_k is the saliency score of the area masked at step k. For the IC metric, we take inspiration from IAUC, and instead of masking the image, we start from a blurred image, and gradually reveal the image according to the saliency map. Once the image is totally revealed, the score variations are computed $v_k = c_{k+1} - c_k$ and we compute the linear correlation of the v_k with the s_k. Note that the order of the subtraction is reversed compared to DC because when revealing the image, the class score is expected to increase.

When computing DC/IC on a well-calibrated saliency method, we expect that when the class score variation is high, the saliency score should also be high, and conversely, when the class score variation is low, the saliency score should be also low.

The DC and IC metrics measure the calibration, which is an aspect that is ignored by the DAUC and IAUC metrics but also by the Sparsity metric. To illustrate, the DC and IC metrics are computed for the examples visible in Fig. 3.

3.3 Limitations

The Sparsity Metric Does Not Take into Account the Prediction Scores. Indeed, this metric only considers the saliency score dynamic and ignores the class score produced by the model. However, this is not necessarily a problem as this metric was designed to be used as a complement to other metrics like DAUC, IAUC, DC, or IC, which takes the class score into account.

The DC and IC Metrics also Generate OOD Images. As we took inspiration from DAUC and IAUC and also passed masked/blurred examples to the model, one can make the same argument as for DAUC/IAUC to show that the reliability of DC and IC could probably be improved by preventing OOD samples.

4 Benchmark

We compute the five metrics studied in the work (DAUC, IAUC, DC, IC, and Sparsity) on post-hoc generic explanation methods and attention architectures that integrate the computation of the saliency map in their forward pass. The post-hoc methods are Grad-CAM [21], Grad-CAM++ [3], RISE [20], Score-CAM [27], Ablation CAM [6]. The architectures with native attention are B-CNN [12], BR-NPA [9], the model from [13] which we call IBP (short for Interpretability By Parts), ProtoPNet [4], and ProtoTree [18]. These attention models generate several saliency maps (or *attention* maps) per input image but the metrics are designed for a single saliency map per image. To compute the metrics on these models, we selected the first attention map among all the ones produced, as, in these architectures, the first is the most important one.

Table 1 shows the performances obtained. The most important thing to notice is the overall low values of correlation, especially for IC, where most values are very close to 0, meaning the saliency scores reflect the impact on the class score as much as random values. This highlights the fact that attention models and explanation methods are currently not designed for this objective, although it could be an interesting property.

Table 1. Evaluation of the interpretability on the CUB-200-2011 dataset.

Model	Viz. method	Accuracy	DAUC	IAUC	DC	IC	Sparsity
ResNet50	Ablation CAM	0.842	0.0215	0.26	0.36	−0.04	8.54
	Grad-CAM		0.0286	0.16	0.35	−0.12	5.28
	Grad-CAM++		0.0161	0.21	0.35	−0.07	6.73
	RISE		0.0279	0.18	**0.57**	−0.11	6.63
	Score-CAM		0.0207	0.27	0.32	−0.05	5.96
	AM		0.0362	0.22	0.31	−0.09	4.04
B-CNN	–	0.848	0.0208	0.3	0.27	−0.02	12.74
BR-NPA		**0.855**	0.0155	**0.49**	0.41	−0.02	**16.02**
IBP		0.819	0.0811	0.48	0.23	−0.04	6.56
ProtoPNet		0.848	0.2964	0.37	0.1	−0.06	2.18
ProtoTree		0.821	0.2122	0.43	0.17	**0.04**	13.75

5 Discussion

One common limit of all the metrics discussed here is that they are designed for methods and architectures producing single saliency maps, making their use for multi-part attention architectures like B-CNN, BR-NPA, IBP, ProtoPNet, and ProtoTree less straightforward. In Table 1, we chose to only select the most important attention map but the other ones should also be taken into account to fully reflect the model's behavior. We also could have computed the mean attention maps from all the ones produced by the model but this would also not be faithful towards the model. Indeed, that would amount to considering that all attention maps have the same weight in the decision, which is not true, as the first attention map has more importance in the decision than the second, which is more important than the third, etc. One possibility would be to estimate the weight of each map and to compute a pondered mean but there remains the issue of computing the weight, which may be difficult in the general case, due to the variety of architectures.

Note that the low values of DC and IC in Table 1 do not imply that the models and methods provide unsatisfying performance, but simply show that the calibration property has not been studied until now.

We propose to quantify the sparsity and the calibration of the saliency maps as these properties have not been studied until now and may be relevant for interpretability. However, to what extent this is true remains to be tested in a subjective experiment. More generally, all the metrics discussed in this paper should be tested against a user experiment. As shown by [5,7], there is a great variety of possible experimental setups depending on what should be explained, in which context, and for who the explanation is targeted.

Notably, how the explanation is presented to the user is still an open question. For example, [22] applies a mask on the input image whereas [2] proposed to superimpose the explanation over the image. Also, various tasks can be given to the user to evaluate the explanation. Slack et al. proposed to mask the input image at the most salient areas and ask users to try to recognize the object with the mask [22]. Instead of guessing the label, Alqaraawi et al. asked users to predict the network's prediction to evaluate if the access to a saliency map helps to improve their prediction [2]. We could use a similar setup with various saliency map methods and evaluate the users' accuracy. Then we could rank the methods according to the impact they have on the users' accuracy and see how this ranking relates to the ranking provided by the objective metrics. Like this, we could deduce which metrics best reflect the impact of an explanation method on the user's understanding of the model.

The main issue is that current user studies seem to show that providing saliency maps to the user affects little their understanding of the model and also does not affect the trust in the model [15]. One study found that the presence of a saliency map helps users to better predict the model's output [2]. However, the effect size measured was small and it could be argued that the effect of changing the explanation method would be smaller or even difficult to observe.

6 Conclusion

In this work, we first studied two aspects of the DAUC and IAUC metrics. We showed that they may generate OOD samples which might negatively impact their reliability. Also, we show that they only take into account the ranking of the saliency scores and show that the visual appearance of a saliency map can significantly change without the DAUC and IAUC metrics being affected. Then, we propose to quantify two aspects that were previously unstudied on saliency maps, the sparsity and the calibration (DC and IC). Finally, we conclude with general remarks on the studied metrics and discuss the issues of a user study that could be used to evaluate the usefulness of these metrics.

References

1. Adebayo, J., Gilmer, J., Goodfellow, I., Kim, B.: Local explanation methods for deep neural networks lack sensitivity to parameter values (2018)
2. Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., Berthouze, N.: Evaluating saliency map explanations for convolutional neural networks: a user study. In: IUI 2020, pp. 275–285. Association for Computing Machinery, New York (2020). https://doi.org/10.1145/3377325.3377519
3. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-CAM++: generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 839–847 (2018). https://doi.org/10.1109/WACV.2018.00097
4. Chen, C., Li, O., Barnett, A., Su, J., Rudin, C.: This looks like that: deep learning for interpretable image recognition. In: NeurIPS (2019)
5. Chromik, M., Butz, A.: Human-XAI interaction: a review and design principles for explanation user interfaces. In: Ardito, C., et al. (eds.) INTERACT 2021. LNCS, vol. 12933, pp. 619–640. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85616-8_36
6. Desai, S., Ramaswamy, H.G.: Ablation-CAM: visual explanations for deep convolutional network via gradient-free localization. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 972–980 (2020). https://doi.org/10.1109/WACV45572.2020.9093360
7. Ferreira, J.J., Monteiro, M.S.: What are people doing about XAI user experience? A survey on AI explainability research and practice. In: Marcus, A., Rosenzweig, E. (eds.) HCII 2020. LNCS, vol. 12201, pp. 56–73. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-49760-6_4
8. Ghosh, S., Shet, R., Amon, P., Hutter, A., Kaup, A.: Robustness of deep convolutional neural networks for image degradations. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2916–2920. IEEE (2018)
9. Gomez, T., Ling, S., Fréour, T., Mouchère, H.: Improve the interpretability of attention: a fast, accurate, and interpretable high-resolution attention model (2021)
10. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, 06–11 August 2017, vol. 70, pp. 1321–1330. PMLR (2017)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)
12. Hu, T., Qi, H.: See better before looking closer: weakly supervised data augmentation network for fine-grained visual classification. CoRR abs/1901.09891 (2019)

13. Huang, Z., Li, Y.: Interpretable and accurate fine-grained recognition via region grouping (2020)
14. Jung, H., Oh, Y.: LIFT-CAM: towards better explanations for class activation mapping. arXiv arXiv:2102.05228 (2021)
15. Kenny, E.M., Ford, C., Quinn, M., Keane, M.T.: Explaining black-box classifiers using post-hoc explanations-by-example: the effect of explanations and error-rates in XAI user studies. Artif. Intell. **294**, 103459 (2021). https://doi.org/10.1016/j.artint.2021.103459
16. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777 (2017)
17. McInnes, L., Healy, J., Melville, J.: UMAP: uniform manifold approximation and projection for dimension reduction (2020)
18. Nauta, M., van Bree, R., Seifert, C.: Neural prototype trees for interpretable fine-grained image recognition (2021)
19. Nixon, J., Dusenberry, M.W., Zhang, L., Jerfel, G., Tran, D.: Measuring calibration in deep learning. In: CVPR Workshops, vol. 2 (2019)
20. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models (2018)
21. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
22. Slack, D., Hilgard, A., Singh, S., Lakkaraju, H.: Reliable post hoc explanations: modeling uncertainty in explainability. In: Advances in Neural Information Processing Systems, vol. 34 (2021)
23. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: SmoothGrad: removing noise by adding noise (2017)
24. Tsai, C.H., Brusilovsky, P.: Evaluating visual explanations for similarity-based recommendations: user perception and performance, pp. 22–30. Association for Computing Machinery, New York (2019)
25. van der Waa, J., Nieuwburg, E., Cremers, A., Neerincx, M.: Evaluating XAI: a comparison of rule-based and example-based explanations. Artif. Intell. **291**, 103404 (2021). https://doi.org/10.1016/j.artint.2020.103404
26. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD Birds-200-2011 Dataset. Technical report, CNS-TR-2011-001, California Institute of Technology (2011)
27. Wang, H., et al.: Score-CAM: score-weighted visual explanations for convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 24–25 (2020)
28. Zhang, J., Kailkhura, B., Han, T.Y.J.: Mix-n-Match: ensemble and compositional methods for uncertainty calibration in deep learning. In: Singh, A., et al. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, 13–18 July 2020, vol. 119, pp. 11117–11128. PMLR (2020)