Revisiting The Evaluation of Class Activation Mapping for Explainability: A Novel Metric and Experimental Analysis

Samuele Poppi Marcella Cornia Lorenzo Baraldi Rita Cucchiara
University of Modena and Reggio Emilia
186923@studenti.unimore.it, {marcella.cornia,lorenzo.baraldi,rita.cucchiara}@unimore.it

Abstract

As the request for deep learning solutions increases, the need for explainability is even more fundamental. In this setting, particular attention has been given to visualization techniques, that try to attribute the right relevance to each input pixel with respect to the output of the network. In this paper, we focus on Class Activation Mapping (CAM) approaches, which provide an effective visualization by taking weighted averages of the activation maps. To enhance the evaluation and the reproducibility of such approaches, we propose a novel set of metrics to quantify explanation maps, which show better effectiveness and simplify comparisons between approaches. To evaluate the appropriateness of the proposal, we compare different CAM-based visualization methods on the entire ImageNet validation set, fostering proper comparisons and reproducibility.

1. Introduction

Explaining neural network predictions has been recently gaining a lot of attention in the research community, as it can increase the transparency of learned models and help to justify incorrect outputs in a human-friendly way. While there have been diverse attempts to provide explanations about the inference process in different forms [14, 15, 12], the graphical visualization of a quantity of interest (*e.g.* regions of the input) remains the most straightforward and effective explanation approach.

Because of the effectiveness of visualizations, there has been a surge of methods to solve the task, including gradient visualization tools [28, 35], gradient-based [30, 31, 16, 27, 1], and perturbation-based approaches [23, 9, 10, 21, 3, 32]. Among them, Class Activation Mapping (CAM) [36, 26, 4, 11, 33, 22, 19, 18] provides effective visual explanations by taking a weighted combination of activation maps from a convolutional layer. The motivation behind the approach is that each activation map contains different spatial information about the input, and when the selected convolutional layer is close to the classification stage of the network, its activations are sufficiently high-level to provide a visual localization that explains the final prediction. Identifying a proper way of calculating the importance (*i.e.*, the weight) of each channel is the main issue that has been tackled by all recent CAM approaches [26, 11, 33].

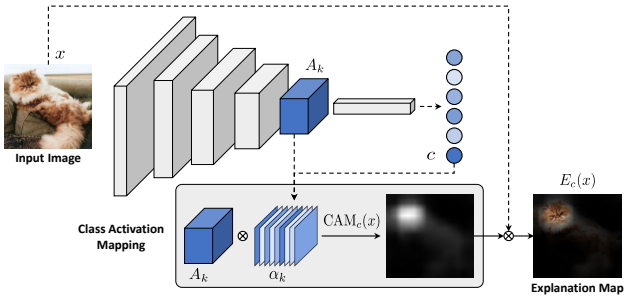

Figure 1. Overview of CAM approaches for explaining predictions: explanation maps are produced via a linear combination of the activations of a convolutional layer.

As it often happens when a new field emerges, the comparison of different CAM approaches has been done mainly in a qualitative way, through the visual comparison of explanation maps, or via quantitative metrics which, however, are not completely effective and sometimes fail to numerically convey the quality of the explanation. At the same time, the evaluation has mostly been limited to few backbones and using protocols that imply a random selection of data and are not completely replicable. With the aim of improving the evaluation of CAM-based approaches, in this paper, we propose a novel set of metrics for CAM analysis, which provides a better ground for evaluation and simplifies comparisons. The effectiveness of the proposed metrics is assessed by comparing a variety of CAM-based approaches and by running experiments in a fully replicable setting.

2. Preliminaries

Let f be a CNN-based classification model and c a target class of interest. Given an input image x and a convolutional layer of f, the *Class Activation Mapping* [36] with

respect to c can be defined as a linear combination of the activation maps of the convolutional layer (Fig. 1), as follows:

$$\text{CAM}_c(x) = \text{ReLU}\left(\sum_{k=1}^{N_l} \alpha_k A_k\right), \quad (1)$$

where N_l denotes the number of channels of the convolutional layer, A_k is the k-th channel of the activation, and α_k are weight coefficients indicating the importance of the activation maps with respect to the target class. Depending on the specific CAM approach, these weights can be in scalar or matrix form, so that it is possible to apply a pixel-level weighting of the activation map. A ReLU activation is employed to consider only the features that have a positive influence on the class of interest, *i.e.*, pixels whose intensity should be increased to increase the score for class c.

Regardless of the particular CAM approach at hand, $\text{CAM}_c(x)$ is usually upsampled to the size of the input image to obtain fine-grained pixel-scale representations. From this, an explanation map $\text{E}_c(x)$ can be generated by taking the element-wise multiplication between $\text{CAM}_c(x)$ and the input image itself (see Fig. 1).

The concept of CAM has been firstly defined in [36] for CNNs with a global average pooling layer after the last convolutional layer. In this case, weights α_k were defined as the weights of the final classification layer. Subsequently, several more sophisticated approaches for computing α_k have been proposed. Grad-CAM [26] generalizes [36] to be applied to any network architecture. It computes the gradient of the score for the target class with respect to the activation map and then applies a global average pooling. Recently, an axiom-based version [11] has been introduced to improve Grad-CAM's sensitivity [31] and conservation [17].

Grad-CAM++ [4], instead, takes a true weighted average of the gradients. Each weight of the average is in turn obtained as a weighted average of the partial derivatives along the spatial axes, so to capture the importance of each location of activation maps. The approach has been further extended in [20] by adding a smoothening technique in the gradient computation. Score-CAM [33], finally, avoids the usage of gradients and instead computes the weights α_k using a channel-wise increase of confidence, computed as the difference in confidence when feeding the network with the input x multiplied by A_k and that of a baseline input.

3. Evaluating CAMs

Ideally, the explanation map produced by a CAM approach should contain the minimum set of pixels that are relevant to explain the network output. While this has been mainly qualitatively evaluated, a quantitative evaluation of explanation capabilities is still in an early stage, with the appearance of different evaluation metrics [4, 21, 11], which although has not been unified throughout the community.

	VGG-16		ResNet-50	
Method	Avg Drop ↓	Avg Increase ↑	Avg Drop ↓	Avg Increase ↑
Grad-CAM [26]	66.42	5.92	32.99	24.27
XGrad-CAM [11]	73.84	4.09	-	-
Grad-CAM++ [4]	32.88	20.10	12.82	40.63
Smooth Grad-CAM++ [20]	36.72	16.11	15.21	35.62
Score-CAM [33]	26.13	24.75	8.61	46.00
Fake-CAM	**0.15**	**45.51**	**0.38**	**47.54**

Table 1. Average Drop and Average Increase values of different CAM approaches, in comparison with Fake-CAM.

In particular, we focus on the following metrics which have been recently proposed, and which focus on the change of model confidence induced by the explanation map.

Average Drop. It measures the average percentage drop in confidence for the target class c when the model sees only the explanation map, instead of the full image. For one image, the metric is defined as $(\max(0, y_c - o_c)/y_c) \cdot 100$, where y_c is the output score for class c when using the full image, and o_c the output score when using the explanation map. The value is then averaged over a set of images.

Average Increase. It computes, instead, the number of times the confidence of the model is higher when using the explanation map compared to when using the entire image. Formally, for a single image it is defined as $\mathbb{1}_{y_c < o_c} \cdot 100$, where $\mathbb{1}$ is the indicator function. The value is then again averaged over different images.

Insertion and Deletion. Deletion measures the drop in the probability of the target class as important pixels (given by the CAM) are gradually removed from the image, while insertion computes the rise in the target class probability as pixels are added according to the CAM. Both metrics are expressed in terms of the total Area Under the Curve.

3.1. Limitations

While the rise of quantitative evaluation approaches is valuable for the field, many of the proposed metrics lack in providing a proper and affordable evaluation for explainability. From a numerical point of view, having a set of different metrics rather than a single-valued score makes comparison between different approaches cumbersome. Secondly, while average increase is too discrete for evaluating the rise of model confidence, average drop alone can easily bring to a misleading evaluation.

To further showcase the limitations of average drop and average increase, we build a "fake" CAM approach in which weights α_k do not depend on confidence scores. Specifically, for each activation map Fake-CAM produces a weight α_k in matrix form, in which all pixels are set to $1/N_l$, where N_l is the number of activation maps, except for the top-left pixel, which is set to zero. The result is a class activation map which is 1 almost everywhere, except for the top-left pixel which is set to 0. Because the resulting explanation map is almost equivalent to the original im-

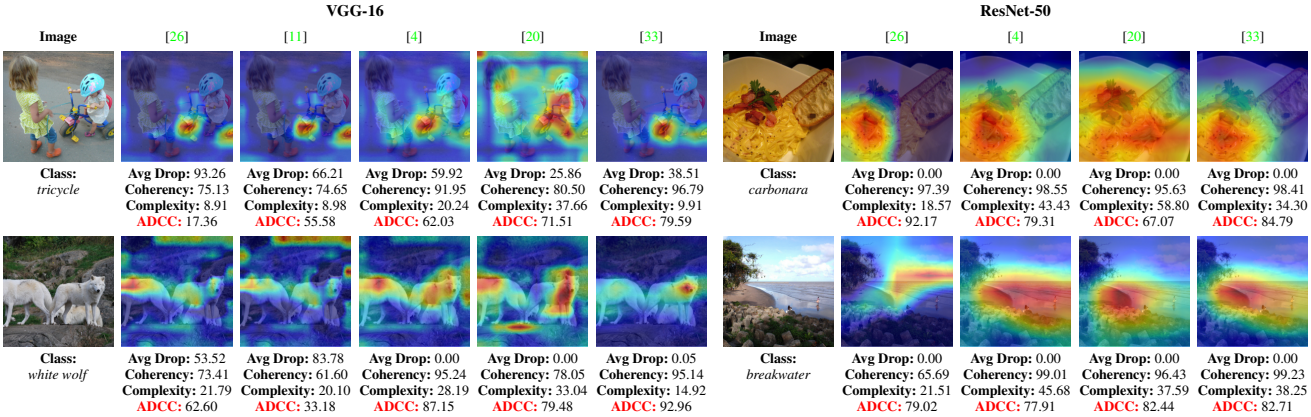

Figure 2. Explanation maps and evaluation scores of different approaches on sample images from ImageNet validation set. We compare the results of Grad-CAM [26], XGrad-CAM [11] (for VGG-16 only), Grad-CAM++ [4], SmoothGrad-CAM++ [20], and ScoreCAM [33].

age, except for one pixel in the corner which is unlikely to contain the target class, the average increase of Fake-CAM is usually very high. When computed on the entire ImageNet validation set [25] surpasses 45% when employing most backbones – a value that is superior to that of any true CAM approach (see Table 1). Similarly, the average drop of Fake-CAM is almost zero, because of the similarity between the input image and the explanation map. While Fake-CAM clearly does not help to explain the model predictions, it achieves almost ideal scores in terms of both Average Drop and Average Increase.

3.2. Proposed Metrics

In order to define a better evaluation protocol, we start by defining which properties an ideal attribution method should verify. The final proposed metric is a combination of three scores, each tackling one of the ideal properties.

Maximum Coherency. The CAM should contain all the relevant features that explain a prediction and should remove useless features in a coherent way. As a consequence, given an input image x and a class of interest c, the CAM of x should not change when conditioning x on the CAM itself. Formally,

$$\text{CAM}_c(x \odot \text{CAM}_c(x)) = \text{CAM}_c(x). \tag{2}$$

Notice that this is equivalent to requiring that the CAM of one image should be equal to that of the explanation map obtained with the same CAM approach. To measure the extent to which an approach satisfies the coherency property, we define a metric that measures how much the CAM changes when smoothing pixels with a low attribution score. Following previous works in the comparison of saliency maps [24, 5, 6, 2, 7, 8], we use the Pearson Correlation Coefficient between the two CAMs considered in Eq. 2:

$$\text{Coherency}(x) = \frac{\text{Cov}(\text{CAM}_c(x \odot \text{CAM}_c(x)), \text{CAM}_c(x))}{\sigma_{\text{CAM}_c(x \odot \text{CAM}_c(x))}\sigma_{\text{CAM}_c(x)}}, \tag{3}$$

where Cov indicates the covariance between two maps, and σ the standard deviation. Since the Pearson Correlation Coefficient ranges between -1 and 1, we normalize the Coherency score between 0 and 1 and, following existing metrics, we also define it as a percentage. Clearly, Coherency is maximized when the attribution method is invariant to change in the input image.

Minimum Complexity. Beyond requiring that the CAM should be coherent in removing features from the input image, we must also require it to be as less complex as possible, *i.e.*, it must contain the minimum set of pixels that explains the prediction. Employing the L_1 norm as a proxy of the complexity of a CAM, we define the Complexity measure as:

$$\text{Complexity}(x) = \|\text{CAM}_c(x)\|_1. \tag{4}$$

Complexity is minimized when the number of pixels highlighted by the attribution method is low.

Minimum Confidence Drop. An ideal explanation map should produce the smallest drop in confidence with respect to using the original input image. To express this third property, we directly employ the Average Drop metric, which linearly computes the drop in confidence.

Average DCC. Finally, we combine the three scores in a single metric, which we name Average DCC, by taking their harmonic mean, as follows:

$$\text{ADCC}(x) = 3\left(\frac{1}{\text{Coherency}(x)} + \right.$$
$$\left. + \frac{1}{1 - \text{Complexity}(x)} + \frac{1}{1 - \text{AverageDrop}(x)}\right)^{-1} \tag{5}$$

Compared to the usage of separate metrics as done in the past, Average DCC has the additional merit of being a single-valued metric with which direct comparisons between approaches are feasible. From a methodological point of view, instead, it takes into account the complex-

Method	VGG-16							ResNet-18						
	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**
Fake-CAM	*0.15*	*45.51*	*32.87*	*35.70*	*100.00*	*100.00*	*0.01*	*0.24*	*45.37*	*31.12*	*33.44*	*100.00*	*100.00*	*0.01*
Grad-CAM [26]	66.42	5.92	11.12	19.56	69.20	15.65	53.52	42.90	16.63	13.43	41.47	81.03	**23.04**	69.98
XGrad-CAM [11]	73.84	4.09	11.59	14.95	66.69	**13.68**	46.29	-	-	-	-	-	-	-
Grad-CAM++ [4]	32.88	20.10	**8.82**	36.60	89.34	26.33	75.65	17.85	34.46	12.30	44.80	98.18	44.63	74.24
Smooth Grad-CAM++ [20]	36.72	16.11	10.57	31.36	82.68	28.09	71.72	20.67	29.99	12.83	43.13	97.53	43.11	74.20
Score-CAM [33]	**26.13**	**24.75**	9.52	**47.00**	**93.83**	20.27	**81.66**	**12.81**	**40.41**	**10.76**	**46.01**	**98.35**	41.78	**77.30**

Method	ResNet-50							ResNet-101						
	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**
Fake-CAM	*0.38*	*47.54*	*38.06*	*38.72*	*100.00*	*100.00*	*0.01*	*0.36*	*43.98*	*43.66*	*41.64*	*100.00*	*100.00*	*0.01*
Grad-CAM [26]	32.99	24.27	17.49	48.48	82.80	**22.24**	75.27	29.38	29.35	18.66	47.47	81.97	**22.51**	76.40
Grad-CAM++ [4]	12.82	40.63	14.10	53.51	97.84	43.99	75.86	11.38	42.07	14.99	56.65	98.28	43.94	76.34
Smooth Grad-CAM++ [20]	15.21	35.62	15.21	52.43	97.47	42.25	76.19	13.37	37.76	**14.32**	58.23	97.76	42.61	76.54
Score-CAM [33]	**8.61**	**46.00**	**13.33**	**54.16**	**98.12**	42.05	**78.14**	**7.20**	**47.93**	14.63	**59.57**	**98.37**	42.04	**78.55**

Method	ResNeXt-50							ResNeXt-101						
	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**	Avg Drop ↓	Avg Inc ↑	Deletion ↓	Insertion ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**
Fake-CAM	*0.34*	*46.70*	*41.67*	*43.31*	*100.00*	*100.00*	*0.01*	*0.26*	*42.43*	*48.90*	*46.79*	*100.00*	*100.00*	*0.01*
Grad-CAM [26]	28.06	29.42	20.73	50.30	82.72	**25.57**	**76.09**	24.12	36.37	20.47	61.04	82.94	**25.45**	**77.62**
Grad-CAM++ [4]	11.12	41.38	17.07	56.05	97.30	48.66	73.16	9.74	42.63	17.63	62.90	95.05	46.27	74.61
Smooth Grad-CAM++ [20]	12.70	36.58	16.90	56.76	97.32	47.48	73.58	9.49	40.43	17.67	**64.16**	96.81	49.24	73.03
Score-CAM [33]	**7.20**	**45.70**	**15.59**	**57.92**	**98.00**	46.86	75.38	**5.37**	**47.70**	**17.30**	63.61	**97.03**	46.83	75.60

Table 2. Evaluation of different CAM-based approaches with existing and proposed metrics, on six different backbones.

ity of the explanation map as well as the coherency of the CAM approach to be evaluated.

4. Experiments

Experimental Setup. Differently from previous works which conducted the evaluation on randomly selected images, we conduct experiments on the entire ImageNet validation set (ILSVRC2012) [25] consisting of 50 000 images, each representing one of the 1 000 possible object classes. To increase the generality of the evaluation, we use six different CNNs for object classification – *i.e.*, VGG-16 [29], ResNet-18, ResNet-50, ResNet-101 [13], ResNeXt-50, and ResNeXt-101 [34], applying each CAM approach on the last convolutional layer. According to the original paper [11], XGrad-CAM is equivalent to Grad-CAM when applied to ResNet models and for this reason, we report the results of this method only on VGG-16. All images are resized and center cropped to 224×224, using mean and standard deviation values computed over the ImageNet training set to normalize the results.

Experimental Results. Fig. 2 shows the scores obtained by the proposed metrics on some sample images, using both VGG-16 and ResNet-50. As it can be seen, the three metrics are complementary in evaluating explanation maps and are equally weighted in the final ADCC score. For instance, in the top-left example, the map produced by Score-CAM [33] achieves the best ADCC score, as it performs favorably in terms of reduced confidence drop while maintaining high levels of coherency and being less complex than other maps. On the contrary, the map produced by SmoothGrad-CAM++ [20] has a lower drop in confidence, but it is less coherent and more complex. Turning to the *carbonara* example, the maps reported by all approaches have the same drop in confidence (*i.e.*, 0), and similar coherency values. The map produced by Grad-CAM [26] has the lowest complexity, thus obtaining the best final score.

In Table 2 we report the values obtained with existing and proposed metrics on all six backbones, on the entire ImageNet validation set. As it can be seen, the results of Fake-CAM are better than true CAM approaches on average drop and average increase for all considered backbones. On the contrary, the proposed ADCC score correctly penalizes the complexity of explanation maps generated by Fake-CAM, thus confirming the appropriateness of the proposed evaluation metric. Comparing true CAM methods, generally Score-CAM [33] achieves the best results on almost all metrics, except for complexity. It shall be noted, also, that Score-CAM [33] performs favorably on VGG and ResNet backbones, while Grad-CAM [26] achieves the best ADCC score when employing ResNeXt models – something which was never tested in literature before. In terms of complexity, Grad-CAM [26] produces less complex but less coherent maps, and with higher confidence drops. The ADCC score jointly accounts for all three properties, providing a single-valued metric from which different approaches can be easily compared.

5. Conclusion

We presented a novel evaluation protocol for CAM-based explanation approaches. The proposed ADCC score takes into account the variation of model confidence, the coherency, and the complexity of explanation maps in a single score, providing an effective mean of comparison. Experiments have been conducted on the entire ImageNet validation set, with six different CNN backbones, testifying the appropriateness of the proposed score and its generality across different settings.

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*, 2018. 1

[2] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, 2018. 3

[3] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *Proceedings of the International Conference on Learning Representations*, 2019. 1

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2018. 1, 2, 3, 4

[5] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Multi-Level Net: A Visual Saliency Prediction Model. In *Proceedings of the European Conference on Computer Vision Workshops*, 2016. 3

[6] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Visual Saliency for Image Captioning in New Multimedia Services. In *Proceedings of the IEEE International Conference on Multimedia & Expo Workshops*, 2017. 3

[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 3

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. SAM: pushing the limits of saliency prediction models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 3

[9] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, 2017. 1

[10] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 1

[11] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, and Biao Li. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In *Proceedings of the British Machine Vision Conference*, 2020. 1, 2, 3, 4

[12] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the International Conference on Machine Learning*, 2019. 1

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 4

[14] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision*, 2016. 1

[15] Lisa Anne Hendricks, Ronghang Hu, Trevor Darrell, and Zeynep Akata. Grounding visual explanations. In *Proceedings of the European Conference on Computer Vision*, 2018. 1

[16] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, 2017. 1

[17] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018. 2

[18] Rakshit Naidu, Ankita Ghosh, Yash Maurya, and Soumya Snigdha Kundu. IS-CAM: Integrated Score-CAM for axiomatic-based explanations. *arXiv preprint arXiv:2010.03023*, 2020. 1

[19] Rakshit Naidu and Joy Michael. SS-CAM: Smoothed Score-CAM for sharper visual feature localization. *arXiv preprint arXiv:2006.14255*, 2020. 1

[20] Daniel Omeiza, Skyler Speakman, Celia Cintas, and Komminist Weldermariam. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. In *Proceedings of the Intelligent Systems Conference*, 2019. 2, 3, 4

[21] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *Proceedings of the British Machine Vision Conference*, 2018. 1, 2

[22] Harish Guruprasad Ramaswamy et al. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, 2020. 1

[23] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?" Explaining the Predictions of Any Classifier. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, 2016. 1

[24] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE International Conference on Computer Vision*, 2013. 3

[25] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3, 4

[26] 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*, 2017. 1, 2, 3, 4

[27] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the International Conference on Machine Learning*, 2017. 1

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 1

[29] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *Proceedings of the International Conference on Learning Representations*, 2015. 4

[30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 1

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the International Conference on Machine Learning*, 2017. 1, 2

[32] Jorg Wagner, Jan Mathias Kohler, Tobias Gindele, Leon Hetzel, Jakob Thaddaus Wiedemer, and Sven Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 1

[33] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1, 2, 3, 4

[34] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 4

[35] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proceedings of the European Conference on Computer Vision*, 2014. 1

[36] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 1, 2