Reviewing and Analyzing the Visualization of Network Activations for Explainability

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

Abstract

As the request for deep learning solutions increases, the need for explainability becomes 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 research activity, 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 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 [Hendricks *et al.*, 2016; Goyal *et al.*, 2019], 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 [Simonyan *et al.*, 2013; Zeiler e Fergus, 2014], gradient-based [Springenberg *et al.*, 2014; Sundararajan *et al.*, 2017], and perturbation-based approaches [Fong e Vedaldi, 2017; Petsiuk *et al.*, 2018]. Among them, Class Activation Mapping (CAM) [Zhou *et al.*, 2016; Selvaraju *et al.*, 2017; Chattopadhay *et al.*, 2018; Fu *et al.*, 2020; Wang *et al.*, 2020] 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 [Selvaraju *et al.*, 2017; Wang *et al.*, 2020].

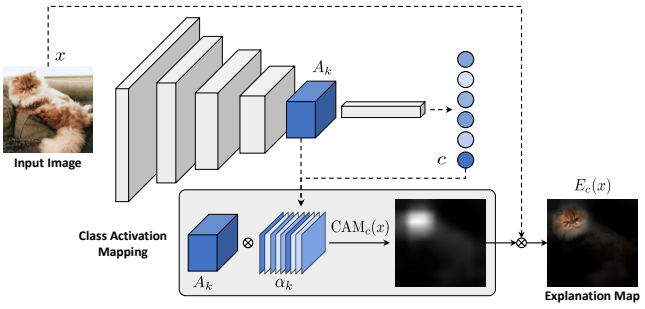

Figure 1: Overview of CAM approaches for explaining predictions: explanation maps are produced via a linear combination of the activations of a convolutional layer.

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 research activity 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 Class Activation Mapping (CAM)

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* [Zhou *et al.*, 2016] 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 [Zhou *et al.*, 2016] 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 [Selvaraju *et al.*, 2017] generalizes [Zhou *et al.*, 2016] 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 [Fu *et al.*, 2020] has been introduced to improve Grad-CAM's sensitivity [Sundararajan *et al.*, 2017] and conservation [Montavon *et al.*, 2018].

Grad-CAM++ [Chattopadhay *et al.*, 2018], 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 [Omeiza *et al.*, 2019] by adding a smoothening technique in the gradient computation. Score-CAM [Wang *et al.*, 2020], 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 [Chattopadhay *et al.*, 2018; Petsiuk *et al.*, 2018; Fu *et al.*, 2020], which although has not been unified throughout the community. 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.

3.1 Limitations

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 image, 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 [Russakovsky *et al.*, 2015] surpasses 45% – 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,

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

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

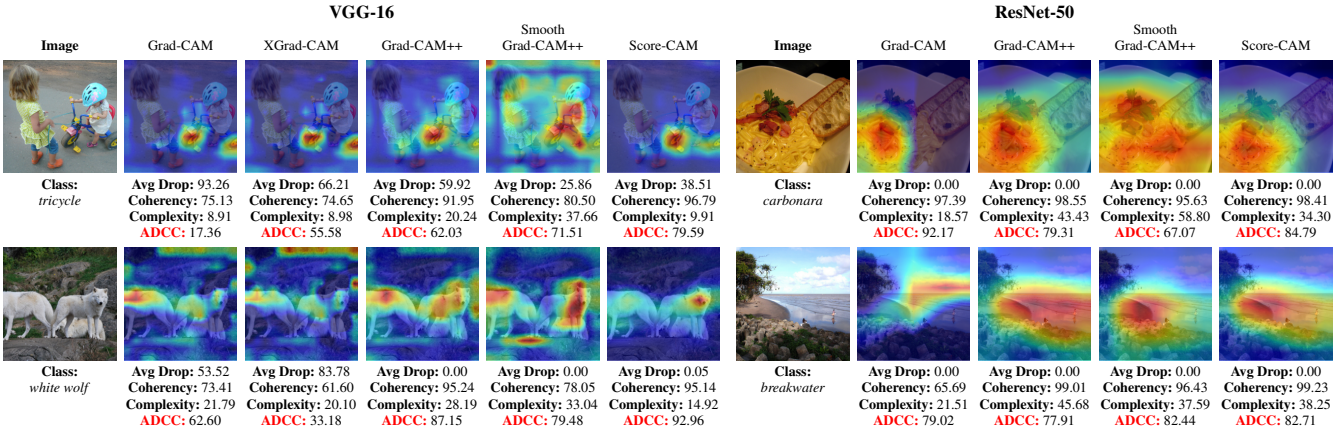

Figure 2: Explanation maps and evaluation scores of different approaches on sample images from ImageNet validation set.

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 [Riche *et al.*, 2013; Bylinskii *et al.*, 2018], 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. 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. 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 complexity of the explanation map as well as the coherency of the CAM approach to be evaluated.

3.3 Experiments

Experimental Setup. Differently from previous works which conducted the evaluation on randomly selected images, we conduct experiments on the full ILSVRC 2012 [Russakovsky *et al.*, 2015] validation set, consisting of 50 000 images, each representing one of the 1 000 possible object classes. We use two different CNNs for object classification – *i.e.*, VGG-16 [Simonyan e Zisserman, 2015] and ResNet-50 [He *et al.*, 2016], applying each CAM approach on the last convolutional layer. According to the original paper [Fu *et al.*, 2020], 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.

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 [Wang *et al.*, 2020] 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++ [Omeiza *et al.*, 2019] has a lower drop in confidence, but it is less coherent and more complex.

In Table 1 we report the values obtained with existing and proposed metrics on all backbones. As it can be seen, the results of Fake-CAM are better than true CAM approaches on average drop and average increase on both considered backbones. On the contrary, the proposed ADCC score correctly penalizes the complexity of explanation maps generated by Fake-CAM. Comparing true CAM methods, generally Score-CAM [Wang *et al.*, 2020] achieves the best results on almost all metrics, except for complexity. Grad-CAM [Selvaraju *et al.*, 2017] produces less complex but less coherent maps, and

Method	VGG-16					ResNet-50				
	Avg Drop ↓	Avg Inc ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**	Avg Drop ↓	Avg Inc ↑	**Coherency ↑**	**Complexity ↓**	**ADCC ↑**
Fake-CAM	***0.15***	***45.51***	*100.00*	*100.00*	*0.01*	***0.38***	***47.54***	*100.00*	*100.00*	*0.01*
Grad-CAM	66.42	5.92	69.20	15.65	53.52	32.99	24.27	82.80	**22.24**	75.27
XGrad-CAM	73.84	4.09	66.69	**13.68**	46.29	-	-	-	-	-
Grad-CAM++	32.88	20.10	89.34	26.33	75.65	12.82	40.63	97.84	43.99	75.86
SmoothGrad-CAM++	36.72	16.11	82.68	28.09	71.72	15.21	35.62	97.47	42.25	76.19
Score-CAM	26.13	24.75	**93.83**	20.27	**81.66**	8.61	46.00	**98.12**	42.05	**78.14**

Table 1: Evaluation of different CAM-based approaches with existing and proposed metrics.

with higher confidence drops. The ADCC score jointly accounts for all three properties, providing a single metric from which approaches can be easily compared.

4 Conclusion and Future Works

Motivated by the request of explainability solutions, we presented a novel evaluation protocol for CAM-based approaches which provides an effective mean of comparison. Further works in this research line will entail the development of visualization techniques for Vision Transformer-based networks, and for networks trained on other vision tasks like object detection, segmentation, image and video classification and captioning.

Acknowledgments

This work has been funded by the InSecTT (www.insectt.eu) project. InSecTT has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 876038. The JU receives support from the European Union's Horizon 2020 research and innovation programme and Austria, Sweden, Spain, Italy, France, Portugal, Ireland, Finland, Slovenia, Poland, Netherlands, Turkey. The document reflects only the author's view and the Commission is not responsible for any use that may be made of the information it contains.

References

[Bylinskii *et al.*, 2018] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, e Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Trans. PAMI*, 41(3):740–757, 2018.

[Chattopadhay *et al.*, 2018] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, e Vineeth N Balasubramanian. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. In *WACV*, 2018.

[Fong e Vedaldi, 2017] Ruth C Fong e Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*, 2017.

[Fu *et al.*, 2020] Ruigang Fu, Qingyong Hu, Xiaohu Dong, Yulan Guo, Yinghui Gao, e Biao Li. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In *BMVC*, 2020.

[Goyal *et al.*, 2019] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, e Stefan Lee. Counterfactual visual explanations. In *ICML*, 2019.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, e Jian Sun. Deep Residual Learning for Image Recognition. In *CVPR*, 2016.

[Hendricks *et al.*, 2016] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, e Trevor Darrell. Generating visual explanations. In *ECCV*, 2016.

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

[Omeiza *et al.*, 2019] Daniel Omeiza, Skyler Speakman, Celia Cintas, e Komminist Weldermariam. Smooth Grad-CAM++: An Enhanced Inference Level Visualization Technique for Deep Convolutional Neural Network Models. In *IntelliSys*, 2019.

[Petsiuk *et al.*, 2018] Vitali Petsiuk, Abir Das, e Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *BMVC*, 2018.

[Riche *et al.*, 2013] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, e Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *ICCV*, 2013.

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, e Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *IJCV*, 115(3):211–252, 2015.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, e Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. In *ICCV*, 2017.

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

[Simonyan e Zisserman, 2015] Karen Simonyan e Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*, 2015.

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

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, e Qiqi Yan. Axiomatic attribution for deep networks. In *ICML*, 2017.

[Wang *et al.*, 2020] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, e Xia Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *CVPR Workshops*, 2020.

[Zeiler e Fergus, 2014] Matthew D Zeiler e Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, 2014.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, e Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016.