Ante-Hoc Generation of Task-Agnostic Interpretation Maps

Akash Guna R T[1] Raul Benitez[2,*] Sikha O K[1,2,*]
[1]Department of Computer Science and Engineering,
Amrita School of Engineering, Coimbatore, Amrita Vishwa Vidyapeetham-India
[2]Departament of Automatic Control, Universitat Politècnica de Catalunya-BarcelonaTech, Spain
* Corresponding Authors
cb.en.u4cse19302@cb.students.amrita.edu,raul.benitez@upc.edu,ok_sikha@cb.amrita.edu

Abstract

Existing explainability approaches for convolutional neural networks (CNNs) are mainly applied after training (post-hoc) which is generally unreliable. Ante-hoc explainers trained simultaneously with the CNN are more reliable. However, current ante-hoc explanation methods mainly generate inexplicit concept-based explanations tailored to specific tasks. To address these limitations, we propose a task-agnostic ante-hoc framework that can generate interpretation maps to visually explain any CNN. Our framework simultaneously trains the CNN and a weighting network - an explanation generation module. The generated maps are self-explanatory, eliminating the need for manual identification of concepts. We demonstrate that our method can interpret classification, facial landmark detection, and image captioning tasks. We show that our framework is explicit, faithful, and stable through experiments. To the best of our knowledge, this is the first ante-hoc CNN explanation strategy that produces visual explanations generic to CNNs for different tasks.

1. Introduction

Convolutional Neural Networks (CNNs) are widely used for image-related tasks in fields such as healthcare [1] and security [12], and are known to learn complex patterns that are difficult for humans to comprehend. However, in critical applications, it is essential to explain how CNNs make their predictions. Existing research that explains CNNs post-training (post-hoc) has been successful, but such methods are oblivious to how the CNN learns its features. Moreover, when a post-hoc method fails, it is difficult to find if the method or the CNN is at fault [13]. These issues make the post-hoc methods unreliable. Ante-hoc explanation methods [3, 8, 13] counter these drawbacks by training the CNN to provide explanations in addition to predictions.

Figure 1. Dominant features identified by our framework on classifying Tiny-Imagenet using ResNet-50

However, the additional explanations deter the performance of the CNN. Ante-hoc explainers usually produce concepts as explanations which cannot provide satisfactory explanations [11] and labelling concepts learned without supervision [13, 17] is prone to human error. In this paper, we propose an ante-hoc framework that produces visual interpretations and can be embedded into CNNs learned for any task to overcome these limitations. Our method was able to capture dominant features used by the CNN for prediction as shown in Fig. 1.

2. Proposed method

Our ante-hoc framework produces implicit visual interpretations by the addition of a weight network. A generic CNN can be considered as a combination of a feature extractor $F(\cdot)$ and a task-specific prediction network $P(\cdot)$. The feature extractor $F(\cdot)$ obtains a meaningful latent representation from the input, which is then fed into the prediction network $P(\cdot)$ which produces predictions from the features. Therefore, a CNN can be defined as $\text{CNN}(\cdot) = P(F(\cdot))$, and is normally trained using a task-specific loss function.

In addition to the previously mentioned elements of the traditional CNN pipeline, we introduce a weight network ($W(\cdot)$) to incorporate concurrent learning of interpretable concepts. The weight network consists of an autoencoder followed by a pixel-wise softmax function. Soft-attention

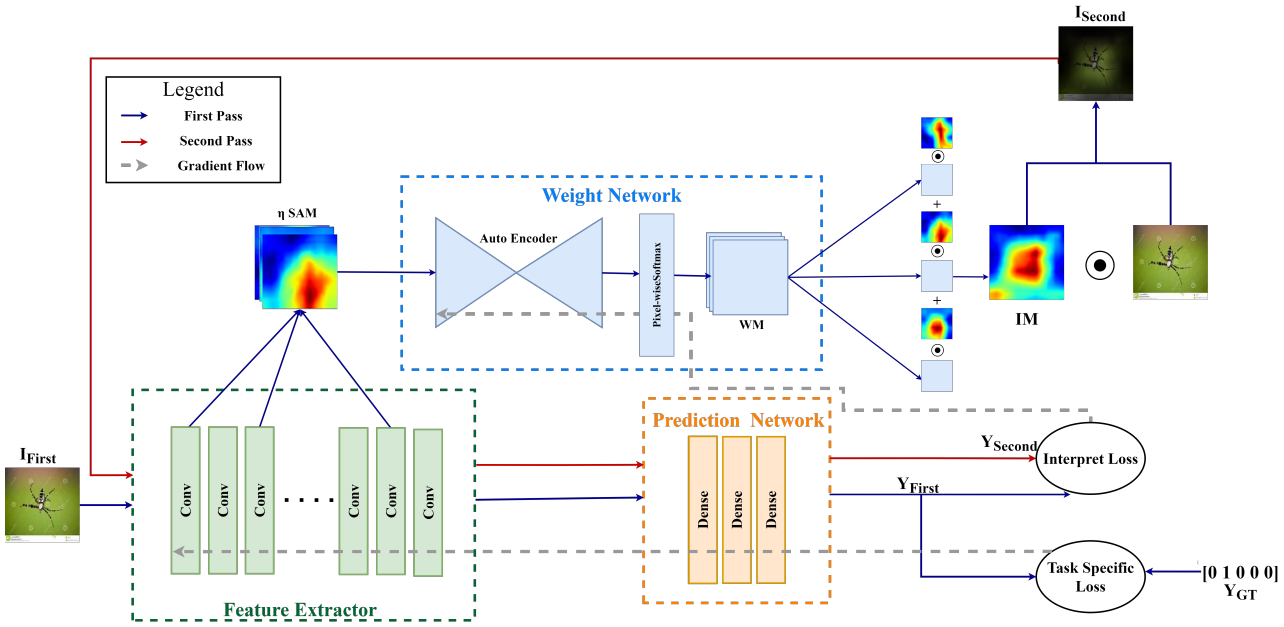

Figure 2. Proposed approach for training the weight network concurrent with a CNN classifier. The First Pass of the training process generates the interpretation map and trains the classifier. In the Second Pass of the training, the weight network is trained.

maps ($\text{SAM}_i = \text{Softmax}(F_{0..i}(\cdot))$ are generated for a selected number of convolutional layers (η) of the backbone CNN, which are passed into the autoencoder to output η feature maps. These feature maps on applying pixel-wise softmax are denoted as weight maps WM. Pixel-wise softmax is the softmax of pixels at a specific two-dimensional position in η SM. The interpretation map I_{map} is then generated. I_{map} is defined as:

$$I_{\text{map}} = \sum_{n=1}^{\eta} \text{WM}_n \odot \text{SAM}_n. \tag{1}$$

where $\odot$ is the hadamard product.

2.1. Training procedure

We propose a two-pass training strategy in which the CNN and the weight network are trained separately. Fig. 2 displays our training strategy. In the first pass, the input to the back-bone CNN is the image I_{first}, which produces predictions Y_{first} and η SAM. The SAM are passed into the weight network W to produce the interpretation map (I_{map}) using Eq. (1). Only the CNN is trained during the first pass using a *task-specific loss* L_{task} defined as

$$L_{\text{task}} = L(Y_{\text{first}}, Y_{\text{GT}}) \tag{2}$$

where Y_{first} is the prediction made by the CNN during the first pass, Y_{GT} is the ground truth and $L(\cdot)$ is a loss function suited for the specific task(for instance, a categorical cross-entropy for a classification problem).

In the second pass, the input to the CNN is the Hadamard product of the image and the generated interpretation map from the previous pass ($I_{\text{second}} = I_{\text{first}} \odot I_{\text{map}}$) which we denote as *Energy Map* in the following sections. The weight network is trained in the second pass using an *interpretation loss* $L_{\text{interpret}}$. The interpretation loss is defined as:

$$L_{\text{interpret}} = \text{MSE}(Y_{\text{first}}, Y_{\text{second}}). \tag{3}$$

The interpretation loss $L_{\text{interpret}}$ penalizes the CNN when the learned concepts are unable to produce the same prediction produced by the CNN with the original image. In other words, the interpretation loss forces the weight network to emphasize significant regions used by CNN for prediction.

3. Experiments

To conduct the experiments, we resize all images to 128×128 and use ResNet-50 [5] as the backbone CNN. We select SAM at fixed intervals of the CNN, including the maps from the first and last convolutional layers. We set $\eta = 8$ for the selection of SAM. We resize the SAM to match the input dimensions. Our results show that the proposed ante-hoc framework generates meaningful explanations while maintaining competitive prediction accuracy compared to state-of-the-art classification pipelines.

3.1. Identification of concepts learned for each class

Concept Identification in existing ante-hoc frameworks is an error-prone manual process. Our framework produces visual interpretations, eliminating the need for the

identification of concepts (dominant regions). We identify concepts by binarizing the produced energy map using a high threshold of 0.8. Fig. 3 shows different concepts that are learned by the proposed explanation method for the Animals-10 dataset [2]. For some images multiple concepts were important for decision, such as the third image in the cat class where both eyes and nose of the cat are used by CNN. The results show that the concepts learned by the proposed explanation framework are human-understandable, providing valuable insights into the CNN's learning process.

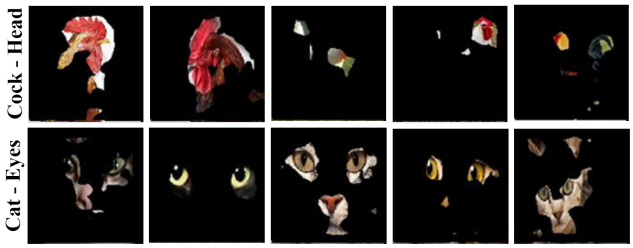

Figure 3. Sample Concepts learned by the CNN for Animal-10 classification. In few instances, the CNN uses multiple concepts to classify images.

3.2. Explaining Landmark Detection

A simple CNN [1] with a convolutional feature extractor followed by dense layers was learned on AFLW2000 [18] dataset to detect distinct facial features. The AFLW2000 dataset includes 2000 faces and their facial landmarks. We used the nose, eyes and mouth landmarks to train three different CNNs, along with the proposed ante-hoc framework. We also tested the performance when all three facial features were learned together. Sample explanations generated by the proposed framework for detecting different facial features are shown in Fig. 4. It is evident from the figure that the proposed method successfully localizes the respective facial features for the different landmark detection tasks.

3.3. Quantitative performance evaluation

In this section, we evaluate the faithfulness and stability of the proposed ante-hoc framework. A method is considered faithful if the trained CNN can predict ground truth labels solely using the key features (concepts) identified by the explanation framework [3, 4]. Our metrics to evaluate the faithfulness of interpretations are inspired by Adithya *et al*. [4]. We use three metrics namely Drop(%), Inc(%) and Win(%). When the input image is replaced with only concept regions, the classification probability either increases due to concentrated representation or decreases due to elimination of important regions. Drop(%)

[1] Landmark Detection Repository

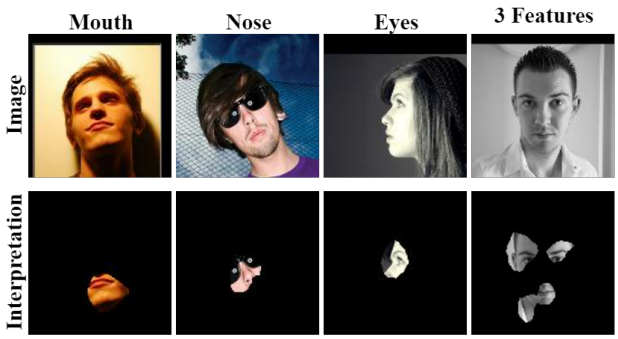

Figure 4. Explanations of proposed framework for landmark detection on AFLW2000 Dataset [18]. Each column displays the binarized energy map when identifying landmark for selected facial features.

(Drop(%) $= \frac{\sum_{k=0}^{D}(E_k^p - I_k^p)/E_k^p}{D} * 100$) measures the percentage drop in classification probability when using the energy map (E_k^p) instead of using the image (I_k^p). Here, D is the total number of cases in the test set where the classification probability drops and k is the ground truth class. Inc(%) (Inc(%) $= \frac{\sum_{k=0}^{C}(I_k^p - E_k^p)/I_k^p}{C} * 100$) measures the increase in classification probability when using energy maps. Here, C is the total number of cases in the test set where the classification probability increases. Drop(%) and Inc(%) quantify faithfulness on a variable subset of the testing data. In an ideal scenario, Drop (%) would have no samples reducing the value of the metric. To ensure that we have a valid Drop(%) or Inc(%) we use Win(%). Win(%) is defined as the percentage of occurrences for which the conditional probability when using the energy map is greater than the conditional probability when using the image. Low Drop(%), high Inc(%) and high Win(%) are desired.

Comparing classification with post-hoc methods: We assess the faithfulness of classification by providing (i) Energy Maps (ii) Binarized Energy Maps as inputs to CNN for evaluation. Energy maps highlight input regions the CNN utilizes for prediction. Binarized energy maps segment the most important regions for the CNN, therefore were considered along with energy maps. The binarized energy maps were obtained using a high threshold of 0.8 since it retains only highly important regions. We use visual post-hoc methods Grad-CAM [14], Grad-CAM++ [4] and Score-CAM [16] as baselines for the experiments. Tabs. 1 and 2 shows the results of the experiment on Tiny-Imagenet [10], CIFAR-10 [9], Animals-10 [2] and Food-11 [15] datasets. We find that our framework produced more faithful explanations compared to post-hoc methods for both input methods. The results using binarized energy maps shows that our framework is better at capturing dominant regions.

Comparing classification with Ante-Hoc Methods:

Table 1. Comparison of the proposed framework with post-hoc methods

Dataset	Grad-CAM			Grad-CAM++			Score-CAM			Ours		
	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)
CIFAR-10	83.86	38.5	13	57.28	47.4	26	33.68	62.76	29	**23.25**	**72.79**	**37**
Animals-10	32.47	15.55	20	27.34	16.74	22	30.85	13.23	26	**19.24**	**16.81**	**34**
Food-11	61.18	65.74	59	62.56	**66.42**	64	60.90	63.87	58	**59.2**	65.98	**67**
Tiny-Imagenet	82.67	57.08	19	80.65	62.18	24	82.87	59.12	18	**78.14**	**67.21**	**33**

Table 2. Comparison of the proposed framework with post-hoc methods when binarizing heatmaps

Dataset	Grad-CAM			Grad-CAM++			Score-CAM			Ours		
	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)	Drop(%)	Inc(%)	Win(%)
CIFAR-10	85.03	21.48	16	67.71	20.48	27	65.43	32.74	38	**22.61**	**63.08**	**45**
Animals-10	84.53	28.76	32	78.19	37.64	39	79.27	42.45	41	**70.50**	**50.54**	**48**
Food-11	88.91	63.4	33	78.56	65.0	35	84.75	44.56	27	**76.23**	**74.67**	**33**
Tiny-Imagenet	87.47	39.60	13	84.58	43.98	21	78.54	29.71	7	**80.56**	**48.33**	**28**

Table 3. Comparison of Acc(%) of ResNet-18 on classification with existing ante-hoc methods

Method	Animals-10	CIFAR-10	Food-11	Tiny Imagenet
SENN	82.64	84.50	75.84	33.85
CBAH	87.35	82.14	73.54	35.64
Ours	**90.56**	**90.86**	**84.32**	**38.82**

We compared our method with ante-hoc methods such as Self Explaining Neural Network (SENN) [3] and Concept based Ante-Hoc framework (CBAH) [13] as shown in Tab. 3. We compared these methods by measuring the classification accuracy (Acc(%)). We chose ResNet-18 [5] as the base CNN to classify. Our method outperformed both ante-hoc approaches across all datasets. The main reason for improved performance is that our framework doesn't affect the task being solved which is highlighted especially in Food-11 where obtaining class generic concepts is complicated therefore other methods deter the classification accuracy. This experiment shows that our visual ante-hoc framework is superior to existing ante-hoc methods.

Assessing performance on Image Captioning: We compare the performance of the proposed ante-hoc framework for an image captioning task. We used a simple CNN-based image captioning network[2] trained on Flickr-8k dataset [7]. The feature extractor of the image captioning network had a ResNet-50 [5] followed by LSTM [6] layers. Tab. 4 shows the quantitative comparison of our method with post-hoc methods for explaining an image captioning task. Existing ante-hoc frameworks focus on explaining only classification hence, post-hoc methods were considered. We used the Drop(%) and Inc(%) for this comparison. Usage of Win(%) for a task that *generates* captions is debatable and therefore was omitted. Our method had a lower Drop(%), higher Inc(%) compared to post-hoc methods [4, 14]. This illustrates that our method can yield faithful ante-hoc explanations irrespective of task.

[2]Image Captioning Repository

Table 4. Comparing our framework with post-hoc methods for interpreting an image captioning task

Metrics	Grad-CAM	Grad-CAM++	Score-CAM	Ours
Drop(%)	45.31	43.68	32.40	**27.93**
Inc(%)	12.25	13.88	10.28	**14.12**

Table 5. Stability of the framework compared with post-hoc methods. Noise is the fraction of pixels affected by salt and pepper noise.

Noise	Grad-CAM		Grad-CAM++		Ours	
	CD(%)	CI(%)	CD(%)	CI(%)	CD(%)	CI(%)
0.1	5.67	3.85	4.85	4.62	4.12	3.28
0.2	8.60	4.82	5.22	5.94	5.38	3.72
0.3	7.54	2.78	8.42	4.84	7.48	4.36
0.4	9.38	7.84	7.34	6.16	12.78	8.15
0.5	14.88	9.58	10.68	8.56	15.90	12.24

Assessing the stability of the framework: An explanation method is stable when the explanations for similar inputs are similar [3]. We assess the stability of the proposed method on perturbing the input images with different amounts of salt and pepper noise before predictions. We perform the same perturbations for visual post-hoc methods. Comparison of stability between concepts and images is unfeasible hence comparison with ante-hoc methods was avoided. The noise level was defined as the fraction of pixels affected by noise and increased from 0.1 to 0.5. We measure the change in Drop(%) (CD) and change in Inc(%) (CI) to assess stability. The average results obtained on all classification datasets are displayed in Tab. 5. When the fraction of noise in an image reaches 0.4 or 0.5, the image is noticeably distinct from the original, unperturbed image. This is because of the high level of noise present in the image,

which can alter its appearance and obscure crucial details, making them harder to discern. Explanations are expected to have a jump in change for these fractions, which was reflected in the results. Our framework had a higher change for these fractions. We also noticed that post-hoc methods had random increase-decrease over increasing fractions. Ideally, change should increase on increase in fractions as displayed by our framework. Hence, our framework is more stable compared to visual post-hoc methods.

3.4. Choice of Weight Network

Our weight network uses an auto-encoder to generate pixel-wise weights from SAM. Usage of auto-encoder ensured different weights for different inputs. Pixel-wise weights represent importance of each pixel better, producing better regions of importance (ROIs). In this experiment, we compared our weight network with a weighing layer that with a layer that learns a single weight for each SAM which on weighted addition produces an interpretation map. Fig. 5 shows the interpretation maps produced using both methods for sample images from the Animals-10 dataset [2]. We find the interpretation maps produced by utilizing weighing network localizes ROIs better compared to learning a single weight for each soft-attention maps.

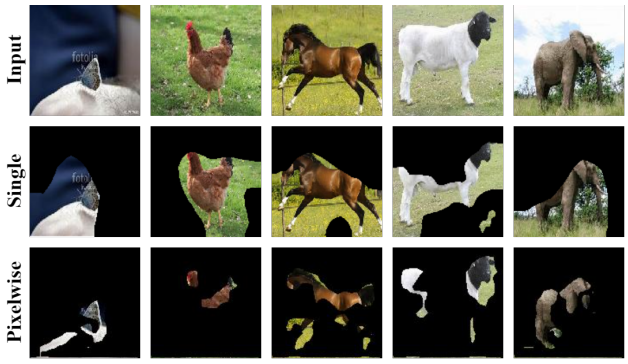

Figure 5. ROIs identified by the method using a single weight for each soft attention map and pixelwise weights (ours). The ROIs are localized better when using pixelwise weights.

4. Conclusion

In this paper, we introduced a task-agnostic ante-hoc explanation framework which produces visual explanations. Visual interpretations produced more explicit explanations than concept based ante-hoc methods. Through experiments, we found that our framework is more faithful and stable compared to existing explanation methods for different tasks such as classification and captioning over various datasets.

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation, grant number PID2020-116927RB-C22 (R.B.).

References

[1] RT Akash Guna, K Rahul, and OK Sikha. U-net xception: A two-stage segmentation-classification model for covid detection from lung ct scan images. In *International Conference on Innovative Computing and Communications*, pages 335–343. Springer, 2023. 1

[2] Corrado Alessio. Animals-10, Dec 2019. 3, 5

[3] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, 31, 2018. 1, 3, 4

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847, 2018. 3, 4

[5] 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, 4

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[7] Micah Hodosh, Peter Young, and Julia Hockenmaier. Flickr8k dataset. 4

[8] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020. 1

[9] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[10] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 3

[11] Andrei Margeloiu, Matthew Ashman, Umang Bhatt, Yanzhi Chen, Mateja Jamnik, and Adrian Weller. Do concept bottleneck models learn as intended? *arXiv preprint arXiv:2105.04289*, 2021. 1

[12] K Rahul-Vigneswaran, Prabaharan Poornachandran, and KP Soman. A compendium on network and host based intrusion detection systems. In *ICDSMLA 2019*, pages 23–30. Springer, 2020. 1

[13] Anirban Sarkar, Deepak Vijaykeerthy, Anindya Sarkar, and Vineeth N Balasubramanian. A framework for learning ante-hoc explainable models via concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10295, 2022. 1, 4

[14] 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 *ICCV*, pages 618–626, 2017. 3, 4

[15] Ashutosh Singla, Lin Yuan, and Touradj Ebrahimi. Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pages 3–11, 2016. 3

[16] 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*, pages 24–25, 2020. 3

[17] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1

[18] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face alignment across large poses: A 3d solution. *CoRR*, abs/1511.07212, 2015. 3