Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization

Saurabh Desai [2] Harish G. Ramaswamy[1,2]

[1]Department of Computer Science and Engineering, Indian Institute of Technology Madras
[2]Robert Bosch Centre for Data Science and Artificial Intelligence (RBC-DSAI),
Indian Institute of Technology, Madras
saketd403@gmail.com, hariguru@cse.iitm.ac.in

Abstract

In response to recent criticism of gradient-based visualization techniques, we propose a new methodology to generate visual explanations for deep Convolutional Neural Networks (CNN) - based models. Our approach - Ablation-based Class Activation Mapping (Ablation CAM) uses ablation analysis to determine the importance (weights) of individual feature map units w.r.t. class. Further, this is used to produce a coarse localization map highlighting the important regions in the image for predicting the concept. Our objective and subjective evaluations show that this gradient-free approach works better than state-of-the-art Grad-CAM technique. Moreover, further experiments are carried out to show that Ablation-CAM is class discriminative as well as can be used to evaluate trust in a model.

1. Introduction

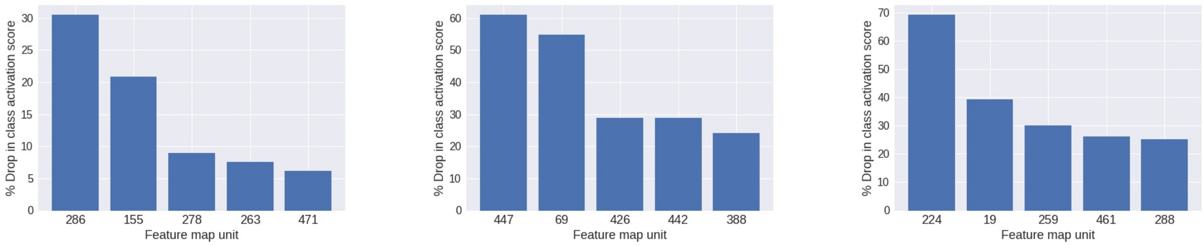
Convolutional Neural Networks (CNNs) are known to show near human-level performance on various computer vision tasks such as image classification [8], object detection [5], semantic segmentation [10] and have performed well on tasks such as image captioning [19] and visual question answering [2]. This is due to the improved architectures of CNNs [4][6] and availability of greater computational power. Despite their superior performance, these deep networks act as black box and are hard to interpret. They are prone to failing without providing any plausible explanation and consequently, users could not place trust in network's decisions [13]. This lack of human trust has limited the meaningful integration of deep learning systems in everyday applications. This issue becomes even more critical for sectors such as healthcare, finance, security etc. where stakes are high for every single decision made. In order to make CNN models trustworthy, it is important to explain their decisions. This transparency will help in understanding failure modes and debugging models as well as identifying and eliminating potential bias in training data [14].
For interpreting convolutional network, it will be useful to locate the regions of input image the model looked at in order to assign a class label to it. Grad-CAM [14] is the state-of-the-art visualization technique to generate such localization maps. This technique relies on the gradients flowing from the decision nodes to final convolutional layer to produce explanations. But each of these output nodes is a non-linear function of the input image as well as previous layers. Hence, Grad-CAM suffers from the problem of gradient saturation which causes the backpropagating gradients to diminish and therefore, adversely affect the quality of visualizations.

We propose a novel “gradient-free” visualization approach - Ablation-CAM to produce visual explanations for interpreting CNNs. This technique avoids use of gradients and at the same time, produces high quality class-discriminative localization maps. Further, we show that, as in case of Grad-CAM, it is possible to fuse pixel-space gradient visualizations such as Guided Backpropagation [18] with Ablation-CAM to produce high resolution localization maps.
The key contributions of this paper are as follows-

- We propose Ablation-CAM, a class-discriminative localization technique that can generate gradient-free visual explanations for any CNN based architecture.

- We demonstrate situations where Ablation-CAM produces more reliable visualizations than Grad-CAM. We show that Ablation-CAM overcomes the limitations inherent with Grad-CAM visualizations.

- We show by subjective and objective evaluation that overall performance of Ablation-CAM is better than



(a) Leopard

(b) Grey Whale

(c) Digital watch

Figure 1: Activation score drop in decision nodes due to ablation of feature map units in final convolutional layer of VGG-16 network trained on Imagenet data for categories leopard(a), grey whale(b) and digital watch(c).

the state-of-the-art Grad-CAM technique. We repeat experiments from Grad-CAM to evaluate class-discrimination of Ablation-CAM. We further show that Ablation-CAM can help users place trust in a model and assist in model selection.

2. Related Work

Our work draws on recent work in ablation analysis, visualizing CNNs, evaluating trust in a model and unreliability of saliency methods.

Visualizing CNNs : One of the earliest efforts to interpret CNNs was made by Zeiler and Fergus [20] by highlighting the pixels in image responsible for activation of a neuron in a higher layer. They achieved this by using deconvolution approach which allows data to flow from a neuron activation in higher layer back to input image. Further, Simonyan *et al.* [15] obtained the partial derivatives of predicted class scores w.r.t. input pixels to produce class-specific saliency maps. Springenberg *et al.* [18] extended this work to Guided Backpropagation which modifies the backpropagating gradients to improve quality of saliency maps. These works are compared in [11]. The visualizations produced by Guided Backpropagation and Deconvolution, though high resolution, are not class-discriminative i.e. for a given image, visualizations w.r.t. different class nodes will be almost identical [14]. Sundarajan *et al.* [17] used integrated gradients to attribute the prediction of CNN to input pixels. Chattopadhyay *et al.* [3] attempted to objectively evaluate efficacy of saliency visualizations.

Above methods provide explanations for individual image instances. Simonyan *et al.* [15] uses gradient ascent to synthesize images that maximally activates a neuron to understand overall notion of concept it represents. Zhou *et al.* [22] show that activation maps in higher convolutional layers act as object detectors and trigger for specific concepts.

Trust evaluation : Lipton *et al.* [9] emphasized the need for interpretable and trustworthy networks. Ribeiro *et al.* [13] conducted human studies to assess if humans can place trust in a classifier.

Unreliability of saliency methods : Adebayo *et al.* [1] and Kindermans *et al.* [7] exposed the unreliability of gradient-based methods citing gradient saturation to be one of the main reasons.

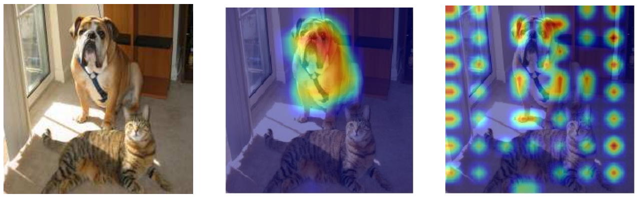
Ablation studies : Morcos *et al.* [12] used ablation analysis to quantify the reliance of network output on single neurons. According to this work, class selectivity is a poor predictor of neuron's importance towards overall performance of network. Zhou *et al.* [23] extends this work to show that ablation of highly selective units, though having negligible effect on overall accuracy, has severe impact on accuracy of specific classes.

Our approach is highly inspired from two visualization techniques i.e. CAM and Grad-CAM. For CNNs with Global Average Pooling (GAP) layer as penultimate layer, Class Activation Mapping (CAM) [21] produces class-discriminative visualization maps. This map is weighted linear combination of feature maps of final convolutional layer where weights are obtained from trained linear classifier of target class node. Since CAM is limited to CNNs with GAP layer, it cannot generate explanations for CNN architectures with fully-connected layers or CNNs trained for tasks such as image captioning and visual question answering.

Gradient-CAM (Grad-CAM) [14] provides a generalization of CAM to be able to explain CNNs irrespective of their architectures. This method utilizes the gradients backpropagating from output node to compute the weights for feature maps as follows -

$$\alpha_k^c = \frac{1}{M} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k} \tag{1}$$

where M is the total number of cells in a feature map, y^c is activation class score for target class c, A_{ij}^k represents activation of cell at spatial location i, j for feature map A^k.



(a) Original

(b) Unit 437 Visualization

(c) Gradient matrix visualization for unit 437

Figure 2: Visualizing feature map 437 (b) and corresponding gradient matrix (c) of VGG model's final convolutional layer for input image (a).

This weight α_k^c is the 'importance' of feature map k for target class c.

As shown in Figure 4, Grad-CAM has some limitations. Grad-CAM fails to provide *faithful* explanations for highly confident decisions due to gradient saturation. Many times, Grad-CAM highlights relatively smaller, incomplete regions of an object in image which might not be enough for the users to place their trust in the system. Also, it fails to detect multiple occurrences of same object in an image. We show that our proposed approach - Ablation-CAM addresses all these shortcomings and proves to be a better visualization technique.

3. Motivation

Current techniques for visualizations depend on gradients backpropagating from output class nodes. These nodes are complex non-linear function of the input image as well as preceding convolutional layers. Gradient-based methods suffer from problem of gradient saturation (discussed in section 5) wherein the backpropagating gradients diminish and hence visualization methods fail to localise relevant regions in image.

Grad-CAM uses gradients of decision node w.r.t. individual cells in feature maps to find the weight (importance) of feature map units for a decision node. Feature maps with lesser spatial footprints fade away in the final saliency map in Grad-CAM. Chattopadhyay *et al.* [3] tries to fix this by taking a weighted average of the gradients of individual cells. These gradients denote the contribution of individual cells for a decision and not of individual feature map. Hence, finding the importance of entire feature map representation by aggregating these seems inappropriate.

Moreover, the backpropagated gradients fail to retain the spatial information. As shown in Figure 2 , the feature map 437 of VGG's final convolutional layer activates for body portion of dog. But backpropagating gradients do not have any spatial correlation to this. We find this for networks with fully connected layers whereas this behaviour is absent for networks without fully connected layers such as Inception and Resnet. As per our knowledge, we are the the first to report this behaviour.

CNN models depend on activations flowing through the network to arrive at a decision whereas the visualization techniques look at the gradients (slope of the learnt function) to understand them. We found this to be counterintuitive. Morcos *et al.* [12] conducted ablation analysis to understand the importance of individual neurons for trained networks. Their findings show that a well-generalized network is less reliant on single neurons and ablation (setting activation value to zero) of individual units will have negligible effect on overall network performance. This paper does not take into account the effect of ablation of units on performance of network for individual classes. Zhou *et al.* [23] showed that removing single feature map units had a severe impact on accuracy of specific classes. Figure 1 shows the effect of ablation of certain units on activation scores of output class nodes. We consider this drop to be an indicator of *how important is an unit for a particular class*. Hence, this ablation drop can be used, instead of global average pooled gradients, to act as weights for feature maps in final convolutional layer.

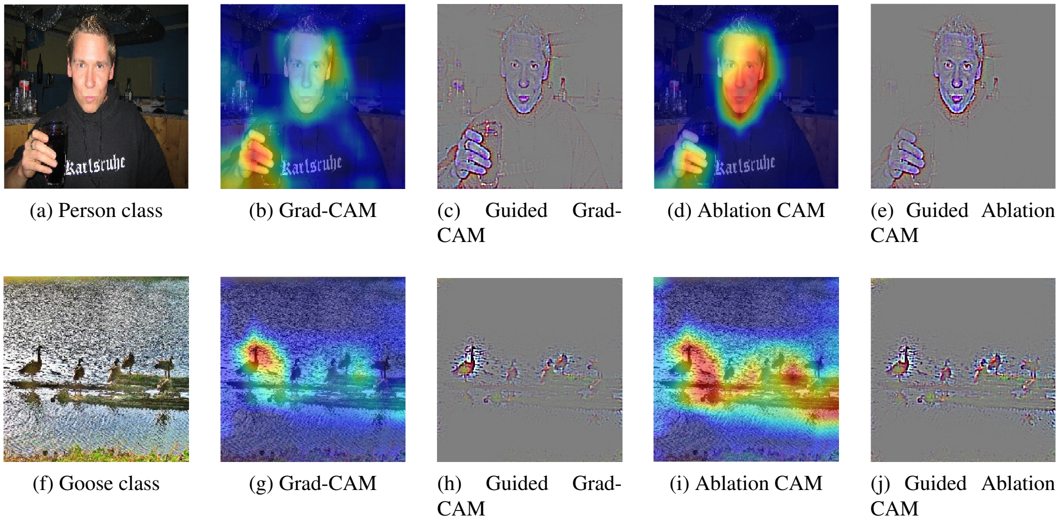
4. Approach

Consider a case where we are required to generate a localization map for an image I using a CNN trained for image classification task. A forward pass through the model is made to obtain the class activation score y^c of class c. Lets assume this class score to be a non-linear function of feature map A_k of final convolutional layer, then y^c will be the value of this function when activations of A_k are present. Set all the individual activation cell values of feature map A_k to zero and repeat the forward pass of same image I. The ablation of unit k leads to a (possibly) reduced activation score y_k^c. Now, y_k^c is value of the function for absence of unit k and acts as a baseline for A_k. Hence, the slope describing the effect of ablation of unit k is given by

$$\text{slope} = \frac{y^c - y_k^c}{||A_k||}. \tag{2}$$

We argue that this effective slope, is better than the "instantaneous slope" arrived via the gradient in Grad-CAM. This approach is immune to both saturation which marks an important filter as not important, and explosion which marks a filter that has very little value as having high importance.

In our approach, we use a slight variant of the slope as a measure of importance of the filter k to class c. This is because norm $||A_k||$ is very large compared to the numerator


(a) Person class (b) Grad-CAM (c) Guided Grad-CAM (d) Ablation CAM (e) Guided Ablation CAM

(f) Goose class (g) Grad-CAM (h) Guided Grad-CAM (i) Ablation CAM (j) Guided Ablation CAM

Figure 3: Ablation CAM and Grad CAM visualizations are shown for two images. Clearly, Ablation-CAM tend to produce better localization for person class (first row) than Grad-CAM. Ablation-CAM is even better at detecting multiple occurrences of goose class (second row).

and hence the slope assumes a very small value.

$$w_k^c = \frac{y^c - y_k^c}{y^c} \tag{3}$$

This importance value, can be simply interpreted as the fraction of drop in activation score of class c when feature map A_k is removed.

Ablation-CAM can then be obtained as weighted linear combination of activation maps and corresponding weights from Equation 3, in a fashion similar to that of Grad-CAM.

$$L^c_{Ablation-CAM} = \text{ReLU}\left(\sum_k w_k^c A_k\right) \tag{4}$$

The ReLU ensures that we only retain units with positive drop values i.e. those units whose absence cause a drop in class score y^c. Similar to Grad-CAM, we resize the map $L^c_{Ablation-CAM}$ to size of the original image to localize important regions in the image. To the best of our knowledge, we are the first to employ this technique.

It should be noted that we have chosen drop in (unnormalized) class activation scores and not drop in confidence scores returned by softmax layer. This is because drop in confidence scores can be achieved via an increase in activation scores of other classes while drop in class scores only focuses on class in question [15]. We experimented by considering drop in confidence scores but we found the visualizations to be less trustworthy.

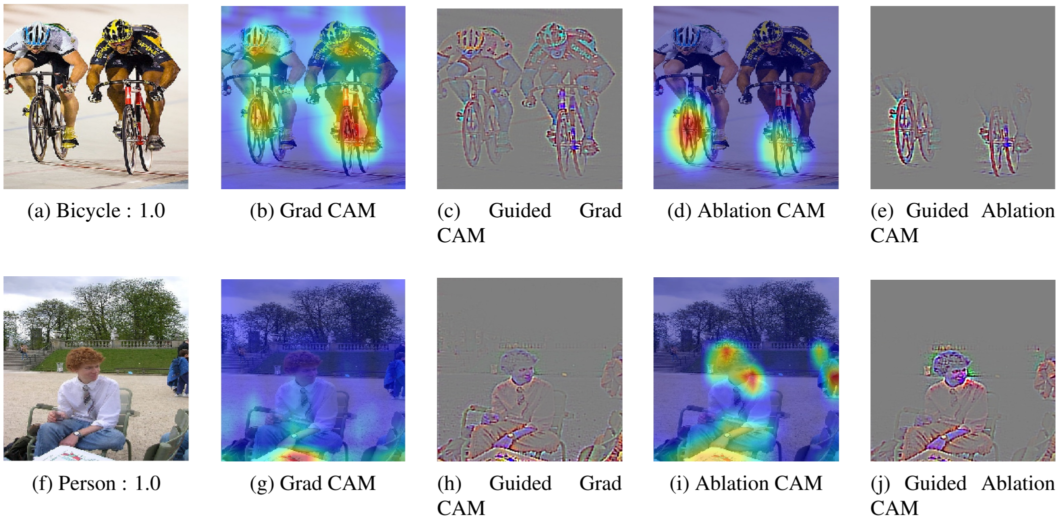
Our approach is similar to the Integrated gradients approach [17], which also attacks the gradient saturation problem. However, instead of choosing a common baseline of a black image for all inputs, classes and filters, we use a natural baseline of zeroing out the corresponding filter activation which varies based on the image and the filter. Also, it should be noted that unlike integrated gradients, our method is not a pixel-space visualization technique and hence, noise-free and class-discriminative.

Guided Ablation-CAM - The heatmaps generated by Ablation-CAM highlight relevant image regions but these do not depict fine-grained details like guided backpropagation visualizations do. Similar to Guided Grad-CAM, Guided Ablation-CAM is obtained by pointwise multiplication of Ablation-CAM and guided backpropagation visualizations.

Figure 3(d)(i) & 3(e)(j) show the visualizations generated by Ablation-CAM and Guided Ablation-CAM respectively.

5. Case for Ablation-CAM

Many times, Grad-CAM visualizations highlight only bits and parts of region of interest and hence, fail to generate considerable amount of trust in human users. Ablation-CAM overcomes this limitation to some extent. Figure 3(b) & 3(d) shows the localization maps for person class generated by Grad-CAM and Ablation-CAM respectively. Clearly, Ablation-CAM visualization provides more complete and trustworthy explanation for person class as compared to Grad-CAM and hence proves to be a better tool



(a) Bicycle : 1.0
(b) Grad CAM
(c) Guided Grad CAM
(d) Ablation CAM
(e) Guided Ablation CAM

(f) Person : 1.0
(g) Grad CAM
(h) Guided Grad CAM
(i) Ablation CAM
(j) Guided Ablation CAM

Figure 4: The first and second rows show decrease in class-discrimination of Grad-CAM maps for high confidence decisions (Person: 1.0 and Bicycle : 1.0). Ablation-CAM seems to produce class-discriminative and trustworthy localization maps.

for trust evaluation. Another shortcoming that Grad-CAM faces is its inability to highlight multiple occurrences of same object in an image. We see that in Figure 3(i), Ablation-CAM clearly does a better job at detecting multiple instances of goose class in the image than Grad-CAM (Figure 3(g)). The backpropagated gradients seem to have high variance due to its dependence on slope of non-linear function learnt by a class node. This causes some pixels of image to be overly emphasized which suppresses the rest of the pixels in a heatmap visualization. On the other hand, our methodology is gradient-free and hence, produces better visualizations due to more uniform emphasis on pixels.

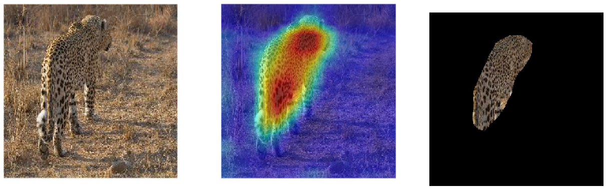
In this part, we demonstrate situations where Grad-CAM technique fails while Ablation-CAM operates robustly providing a strong case for our approach. For purpose of our experiment, we fine-tuned an off-the-shelf VGG-16 network, pretrained on Imagenet data, on Pascal VOC 2007 dataset. As this is multi-label classification problem, we will be using sigmoid activation function for final layer. To test class-discrimination ability of visualizations, we only chose the images containing at least one instance of two different classes. Consider a case where for a given input, the model is highly confident for a class. Then, this class node operates in a region of sigmoid curve where the slope is saturated to almost zero. This causes the backpropagating gradients, required for producing Grad-CAM, to diminish. As these gradients vanish, the localization maps lose their property of discriminaton and fail to localize relevant regions as shown in Grad-CAM visualizations of Figure 4(b)(c)(g)(h). Though Grad-CAM visualizations are generally generated w.r.t. unnormalized nodes (before applying sigmoid or softmax function), these nodes are nevertheless complex non-linear functions of previous layers and are prone to fail due to saturation of gradients. We choose normalized nodes as we know the region where sigmoid saturates (values close to 1). We cannot visualize or anticipate the regions along which the unnormalized nodes would have saturated. The above analysis serve as an illustration of how Grad-CAM could fail. Ablation-CAM, being gradient-free, is robust to such cases and generates high quality visualizations (Figure 4 (d)(i)).

6. Experiments and Results

Our experiments involve both the objective and subjective evaluation of Ablation-CAM and its comparison with Grad-CAM. We have used VGG16 [16] and Inception-v3 [4] models pretrained on Imagenet for experiments. We have repeated experiments from Grad-CAM in sections 6.4 and 6.5 to show that Ablation-CAM have similar capabilities as Grad-CAM.

6.1. Empirical evaluation of Ablation-CAM

We leveraged the study used in [3] for objective evaluation of Ablation-CAM. For every image I, a class-conditional localization map (heatmap) is generated using a visualization technique such as Grad-CAM or Ablation-CAM. This heatmap will highlight the most important discriminative regions as red. The basic idea behind explanation map is to generate an image which contains only the sub-regions of the original image which is emphasized by a visualization technique. We experimented by retaining



(a) Original image

(b) Ablation-CAM

(c) Explanation map for Ablation-CAM

Figure 5: Original image of leopard followed by Ablation-CAM visualization and corresponding explanation map.

only top 20 percent of pixels of localization map. We have also experimented with top 50 and 30 percent values and found similar results. In order to generate explanation map, the localization heatmap is modified so that top 20 percent pixels are 1 and the rest 0. An explanation map is generated by point wise multiplication of original image with modified localization map. Figure 5 shows the explanation map generated for Ablation CAM visualization containing only certain amount of pixels of original image. Unlike the explanation maps in GradCAM++ [3], we choose to use top x% pixels to compare the effectiveness of heatmaps produced by two techniques. This ensures that one technique does not outperforms other simply by highlighting more number of pixels but rather it captures more relevant information for given number of pixels.

We evaluate the performance of explanation maps produced by Ablation-CAM and Grad-CAM using six metrics: (i) Average drop in confidence. (ii) Average drop in activation score. (iii) Percent increase in confidence. (iv) Percent increase in activation score. (v) % Win in confidence. (vi) % Win in activation score. All the results are computed on Imagenet validation set for VGG-16 and Inception-v3 models.

(i & ii) Average drop in confidence and activation score : A good class-conditional explanation map will cover most of relevant parts of the object in the image necessary for arriving at a decision. Hence, a better explanation map, when provided as input instead of full image, is expected to result in lower drop in model's output scores. We used this to compare the visualizations produced by Ablation CAM with that of the Grad CAM. The metric is given as the percentage drop in model's scores when only explanation map is provided as input :

Metric	Ablation CAM	Grad CAM
Average % drop in confidence (lower is better)	**41.52**	45.15
Average % drop in activation (lower is better)	**29.23**	34.07
Percent increase in confidence (higher is better)	**24**	23.06
Percent increase in activation (higher is better)	**14**	12.57
Win % in confidence (higher is better)	**47.97**	34.68
Win % in activation (higher is better)	**56.61**	34.30

Table 1: Results for VGG-16 on Imagenet 2012 validation data.

$$Average\,drop\,\% = \frac{1}{N}\sum_{i=1}^{N}\frac{max(0, Y_i^c - O_i^c)}{Y_i^c} * 100 \quad (5)$$

where Y_i^c is the output score (confidence score or activation score) when original image i is provided as input and O_i^c is the output score when explanation map is provided as input. N is total number of images in dataset. We use max to eliminate cases where $O_i^c > Y_i^c$. Table 1 shows that Ablation-CAM beats Grad-CAM on this metric by causing a lower drop in output scores. On the other hand, we see that both the techniques perform equally well for models without fully connected layers such as Inception-v3 (Table 2).

(iii & iv) Percent increase in confidence and activation score : It is observed that there are instances where providing the explanation map instead of full image increases the output confidence and activation scores (especially when the context is acting as noise for the class). A good explanation map is expected to do this often. This metric is defined as rate at which model's output scores increases when only explanation map is provided as input for an entire dataset. Formally, this can be expressed as :

$$Rate\,of\,increase\,in\,scores = \sum_{i=1}^{N}\left(\frac{1_{Y_i^c < O_i^c}}{N}\right)100 \quad (6)$$

where 1_x is an indicator function that returns 1 when argument is true. As seen from Table 1 Ablation-CAM has more cases where it increased the confidence and activation scores than Grad-CAM.

(v) Win % : To further add to the above metrics, we also computed the number of times in an entire dataset the drop

Metric	Ablation CAM	Grad CAM
Average % drop in confidence (lower is better)	**23.91**	24.24
Average % drop in activation (lower is better)	**10.27**	10.81
Percent increase in confidence (higher is better)	41.16	**41.95**
Percent increase in activation (higher is better)	41.91	**42.73**
Win % in confidence (higher is better)	**33.28**	30.96
Win % in activation (higher is better)	**32.83**	30.72

Table 2: Results for Inception-v3 on Imagenet 2012 validation data.

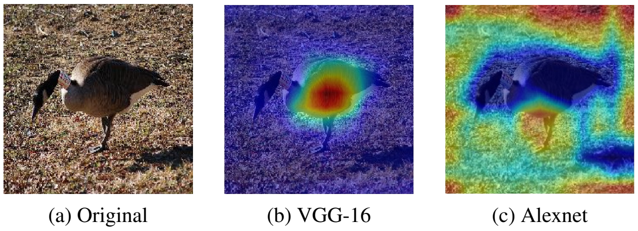

(a) Original (b) VGG-16 (c) Alexnet

Figure 6: Ablation-CAM visualizations for models VGG-16 and Alexnet.The subjects were asked to choose which of the two explanation maps is more trustworthy. For this instance, clearly VGG-16 produces more reliabe explanation than Alexnet.

in model's output scores for an explanation map generated by one technique is less than that for another. This is expressed as percentage. Ablation-CAM outperforms Grad-CAM for over 50 percent of cases (Table 1). Here we have only considered positive drop values and zeroed all the negative values (which cause an increase in scores).

We see that overall Ablation-CAM performs better than Grad-CAM for all metrics if the model has fully-connected layers (Table-1). On the other hand, for Inception-v3, there is not much difference in performance (Table 2) as Grad-CAM effectively becomes CAM and we use the weights of final linear classifier.

6.2. Evaluation using axiomatic approach

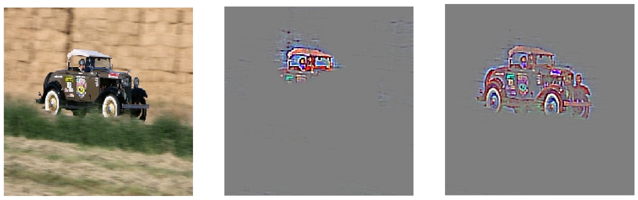
Sundarajan *et al*. [17] used an axiomatic aprroach for evaluating attribution methods. An attribution method requires to satisfy two properties (a) *Sensitivity* and (b) *Implementation invariance*. Ablation-CAM satisfies Sensitivity by employing a baseline for each feature map input, and in some sense try to compute "discrete gradients" instead of instantaneous gradients. Like any CAM, Ablation-CAM may seem to violate the Implementation Invariance property as the visualization depends on size of feature map of intermediate convolutional layer. This size may vary for two networks even though they might be functionally equivalent. Inspite of this, we argue that Ablation-CAM provides better localization and noise-free discriminative visualizations as compared to pixel-space techniques like deconvolution, integrated gradients.

6.3. Subjective evaluation of Ablation-CAM

In 6.1, we carried out an objective evaluation of *faithfulness* of the explanations generated by Ablation-CAM and compared it to those generated by Grad-CAM. Here, we perform comparative assessment of human *trust* in the localization maps generated by the two methods. For this experiment, we chose the classes with highest F1-score from validation dataset to ensure that underlying model VGG-16 performs well on these categories. This led to a total of 250 images as each class had 50 images in validation set. For each image, localization maps were generated using Grad-CAM and Ablation-CAM. These localization maps were shown to 10 human subjects (who had no knowledge of the deep learning field). These subjects were provided with a class label for each image and were asked to select the map which best highlighted the object(s) in the class. The subjects also had the option to select "same" if the two localization maps were very similar. Suppose for a given image, 3 subjects chose Grad-CAM map as better one, 5 subjects chose Ablation-CAM map and the rest 2 chose the option"same". The respective normalized scores of Grad-CAM, Ablation-CAM and the option "same" are 0.3, 0.5 and 0.2. Hence, the maximum achievable score for any option is 250. Ablation-CAM achieved a score of 130.8 as compared to 62.8 of Grad-CAM. The remaining 56.4 was labeled as "same" by the subjects. This empirical study provides strong evidence for the case that Ablation-CAM visualizations are more trustworthy as compared to those generated by Grad-CAM.

6.4. Ablation-CAM for model selection

In the previous section, we compared the Ablation-CAM and Grad-CAM maps generated for the same model. Here we compare the visualizations generated by Ablation-CAM maps for two different models. We leverage the study in [14] to test the hypothesis that a model with better generalization performance will produce better Ablation-CAM visualizations. We compared Ablation-CAM visualizations of VGG-16 with that of Alexnet, where VGG-16 is known to perform better than Alexnet with top 1 percent test er-



(a) A person driving a car (b) Is this person or car? (c) Is this person or car?

Figure 7: Original image of a person driving a car. This image contains two categories - (i) Person (ii) Car. Based on visualization, subject need to answer which class is being depicted.

ror of 28.41 (vs 43.45) on Imagenet classification. We expect that for many instances the visualizations for Alexnet will highlight less relevant parts of the object in the image as compared to those produced by VGG-16. Again for the purpose of these experiments, we chose the same 5 classes of Imagenet. In order to seperate the efficacy of visualizations from accuracy of models, we only considered the images for which both the models made the same predictions as ground truth. Given Ablation-CAM visualizations for Alexnet and VGG-16 along with the object category, 10 human subjects were asked to rate the reliability (*which visualization describes the object best*) on a scale of (+2/-2) if first visualization is clearly more/less reliable than second , (+1/-1) if it is slightly more/less reliable and equally reliable (0) (Figure 6) . The subjects had no idea of which one of the two visualizations belonged to VGG-16 or Alexnet. Moreover, we randomly switched VGG-16 and Alexnet visualizations to be the first option. The subjects assigned VGG model a score of 1.31 meaning it is clearly more reliable than Alexnet. This means the subjects were able to identify the more accurate model on account of better visualizations it produces. This confirms our hypothesis. Thus, Ablation-CAM can help users to place trust and assist in model selection.

6.5. Evaluating class discrimination

An important property of any good visualization is class-discrimination i.e. *how selective is a visualization for a class when more than one category objects are present in the image.* It is important that visualization for a particular object category do not highlight parts of another category. Class-discrimination is another way to ensure faithfulness of visualization to the model. For this experiment, we will use VGG-16 finetuned on Pascal VOC 2007 train set and use validation set to generate visualizations. We select images from VOC 2007 val set that contain exactly two categories. For each image, we produce two category-specific visualizations using Guided Ablation-CAM. These visualizations were shown to 10 human subjects who were asked to answer a simple question: Which of the two object categories are highlighted by the visualizations ? (Figure 7). They also had the option to choose "both". We evaluated 50 image-category pairs with 10 responses for each pair.Note that the option "both" was considered as incorrect answer for any image-category pair. We repeated the experiment for Guided Grad-CAM visualization. For Guided Ablation-CAM, the subjects were able to correctly identify the category being visualized in 73.6 % of the cases while, for Guided Grad-CAM the score was 65.4 %.

7. Limitations of Ablation-CAM

The computational time required to generate a single Ablation-CAM is greater than that required for Grad-CAM. The reason being we have to iterate over each feature map to ablate it and check the corresponding drop in class activation score. On the other hand, Grad-CAM requires single back-propagation to generate a Grad-CAM. The difference in generation time can be considerably reduced by using proper multiprocessing as we did for our experiments.

Secondly, table (2) shows that Ablation-CAM performs only slightly better than Grad-CAM when we use models such as Resnet-50 and Inception-v3 which do not have any fully-connected layers. Essentially, for these models Grad-CAM boils down to CAM and the output nodes are linear combination of global-average pooled feature maps of last convolutional layer. In such situations, Ablation-CAM works as good as any other CAM. Ablation-CAM will perform better for tasks such as image captioning where last convolutional layer is not followed immediately by decision nodes.

8. Conclusion

In this work, we proposed a novel technique - Ablation-CAM to produce class-discriminative localization maps for explaining individual decisions of CNN-based models. Unlike previous techniques, this technique is "gradient-free". We showed through objective and subjective evaluations (sections 6.1 and 6.3) that Ablation-CAM out performs the existing state-of-the-art Grad-CAM. We also show through further experiments that Ablation-CAM is class-discriminative, can be used to place trust in models and assist in model selection. In future work, we plan to apply Ablation-CAM to produce explanations for non-vision tasks such as reinforcement-learning, natural language processing, etc.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, M. H. Ian Goodfellow, and B. Kim. Sanity checks for saliency maps. *NIPS*, 2018.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. *ICCV*, 2015.

[3] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-cam++: Improved visual explanations for deep convolutional networks. *arXiv preprint arXiv:1710.11063*, 2018.

[4] S. Christian, L. Wei, J. Yangqing, S. Pierre, R. Scott, A. Dragomir, E. Dumitru, V. Vincent, and A. Rabinovich. Going deeper with convolutions. *arXiv preprint arXiv:1512.03385*, 2015.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *CVPR*, 2014.

[6] K. He, X. Zhangand, S. Ren, and J. Sun. Deep residual learning for image recognition. *CVPR*, 2016.

[7] P.-J. Kindermans, S. Hooker, J. Adebayo, M. Alber, K. T. Schütt, S. Dähne, D. Erhan, and B. Kim. The (un)reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[9] Z. C. Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

[10] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. *CVPR*, 2015.

[11] A. Mahendran and A. Vedaldi. Salient deconvolutional networks. *European Conference on Computer Vision.*, 2016.

[12] A. S. Morcos, D. G. Barrett, N. C. Rabinowitz, and M. Botvinick. On the importance of single directions for generalization. *ICLR*, 2018.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. Why should i trust you? : Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM,2016.

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam:visual explanations from deep networks via gradient-based localization. *arXiv preprint arXiv:1610.02391*, 2017.

[15] K. Simonyan, A. Veldadi, and A. Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[17] M. Sundarajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. *ICML*, 2017.

[18] J. Tobias, Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. Striving for simplicity: The all convolutional net. *ICLR Workshop*, 2015.

[19] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR*, 2015.

[20] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. *ECCV*, 2014.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. *arXiv preprint arXiv:1512.04150*, 2015.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Object detectors emerge in deep scene cnns. *International Conference on Learning Representations*, 2015.

[23] B. Zhou, Y. Sun, D. Bau, and A. Torralba. Revisiting the importance of individual units in cnns via ablation. *arXiv preprint arXiv:1806.02891*, 2018.