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

Shrey N Pandit

August 19, 2020

Abstract

With the current improvement in Machine learning techniques especially in the field of Convolutional Neural Network, the focus goes back to what the model actually looked at and what part of image lead to classify its classification, this is where Visual Explanations come into play.
In this project I tried implementing the Research paper based on Ablation Cam, it's comparison with the current State of the art method in visual explanation (Grad-Cam) and it's performance on various types of models.

1 Introduction to the Paper

The authors of the paper - **Saurabh Desai and Harish G. Ramaswamy** saw various shortcomings in Grad-Cam which involved gradients and the common problem of gradients were associated with it such as gradient saturation and loss of spatial information in a fully connected layer so they wanted to find a method which was independent of gradients.

The main idea of Gradient-free Localization is that the whole model should be independent of calculation and application of gradients calculated in Grad-cam by this way we can get rid of the various problems that Gradcam had. The main idea of this paper is that if a part of the image is important for its classification then removing it from the image or ablating it, would result in a dip in prediction score. The part of the image is ablated by replacing the activation value of each feature map to 0. This dip in prediction value is then calculated and is used to assign importance to the different activation layers.

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

Figure 1: Here the y^c is the prediction value of class c and y_k^c is the predicted value of class c when the activation layer k is ablated . Also $||A||$ is the norm of the activation layer k

So the paper proposes that this effective slope is better than the ”instantaneous slope” proposed in Grad-Cam. But as the norm term in the denominator is much bigger than the numerator there is a slight change in the formula. This value can be simply interpreted as the fraction of drop-in activation score of class c when feature map A_k is removed. So the new proposed formula changes to -

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

Figure 2: Here the y^c is the prediction value of class c and y_k^c is the predicted value of class c when the activation layer k is ablated .

It is then this new weights of each activation map that gets multiplied to the map and signifies its importance. After this ReLU function is applied on the map ensuring only those map are kept, absence of which would result in a drop in class score.

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

Figure 3: Here w_k^c is the weight of the k'th activation map. A_k is the activation map .

Now a bit of explanation of how Grad-cam works , It similar to Ablation-Cam just the difference is in the method of calculating the weights (for the Activation-map). Here we calculate the Gradient of each value in the activation map with respect to the change in value of the class score. The summation of all such points in the activation map is considered as the final score/weight for the whole activation map A_k.

$$\sum_i \sum_j \frac{\partial y^c}{\partial A^k_{i,j}}$$

Figure 4: This is the formula used for calculating the weight that correspond to the importance of each activation map .

2 Implementation of the Paper

The model used in the paper and the data set it is trained on are -:

1)Xception model pre-trained on imagenet weights.
2)Custom model (7 layered) trained on CIFAR-100 (Not pre-trained).

After the models were trained on the data set, Two models are created 1) Activation Model and 2) Classifier model

The function of Activation model was to output the values of activation layer of the last convolutional layer , whereas the classification model is used to generate class prediction value from the activation layer.

Ablation cam is implemented as follows :-
The image needed to classify is passed through the activation model and the output of this model (activation layer) is stored as A . This output is then passed into the classifier model and the predicted value is noted as y_c. Then Each activation map is ablated and is passed with other non-ablated maps into the classifier model and the value is noted as y_c^k .

After all the values of y_c and y_k are calculated our next step is to calculate the weight ratio W_k. It is calculated using the formula Figure.2.

Our final step in the process is to multiply the weight ratio calculated in the previous step and multiply it to the corresponding activation maps A_k.

The results can be visualized by creating the heat map of the final activation maps and superimposing it to the original image .

The method was followed and results were generated for different classes to see the adaptability and results over various different custom images downloaded from the web (not in the training set).

Further Grad-Cam was implemented using pre-written code of Grad-CAM and results were generated for the same images for comparison.

3 Results

In this section I have presented the results of the Ablation cam on various images of different classes.

3.1 Pre-trained Model

Following results are of Xception model on pre-trained on Imagenet Data set. We can observe the various part of images that the model looks as to classify it the way it did.

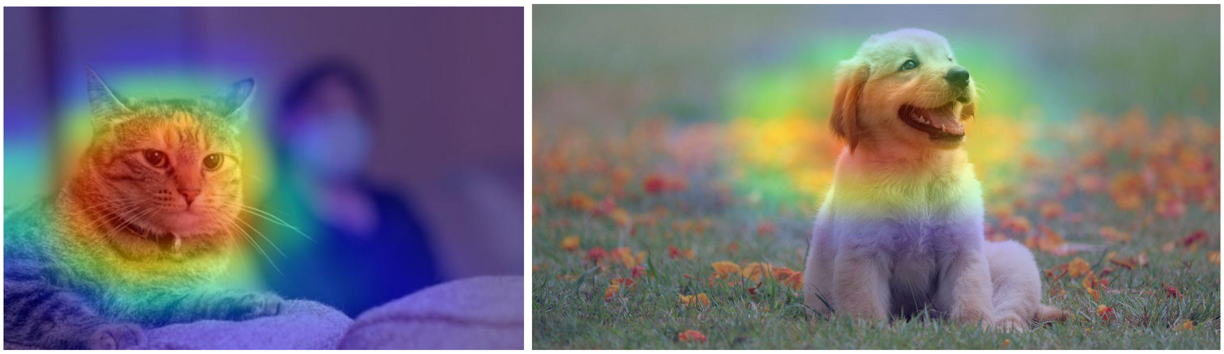

Figure 5: The figure shows result of Ablation cam on Cat and Dog Image

Further I tried the model on Object class like aeroplane car and ship.

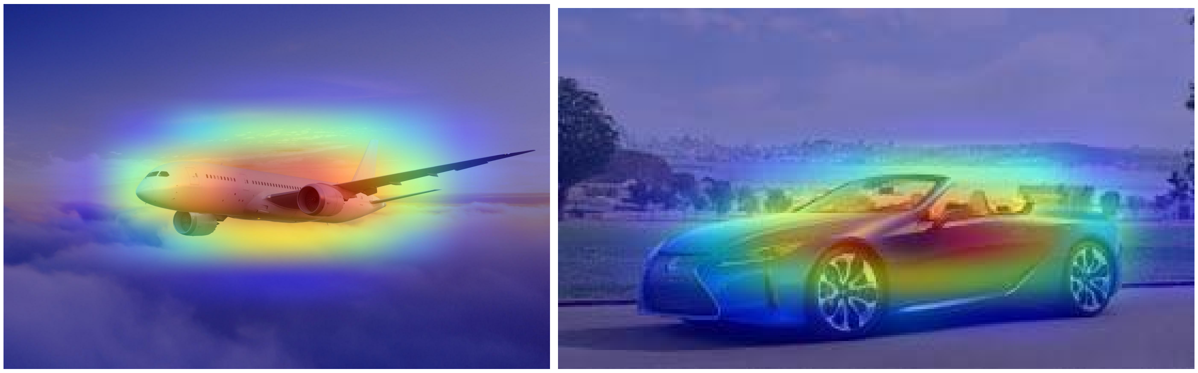

Figure 6: The figure shows result of Ablation cam on Aeroplane and Car Image

As we can see the architecture worked great on identifying the exact position in image upon which it classified that particular image into the class.

3.2 Cifar - 100 Model

Below are the results obtained by cifar -100 model trained to 60% Validation Accuracy.

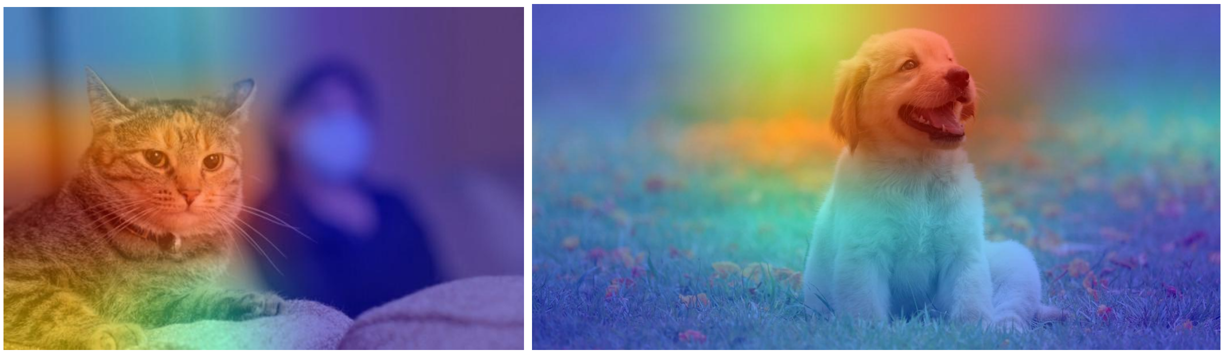

Figure 7: The figure shows result of Ablation cam on Cat and Dog Image

As we can see that the part highlighted in the image now contains a bit of background and not exactly the face . This shows us that Ablation -Cam is very much depended on the type of Model and the performance of Model .

4 Ablation-Cam vs Grad-Cam

Here we will compare the results of the current SOTA model (Grad-Cam) and the newly implemented Ablation cam .

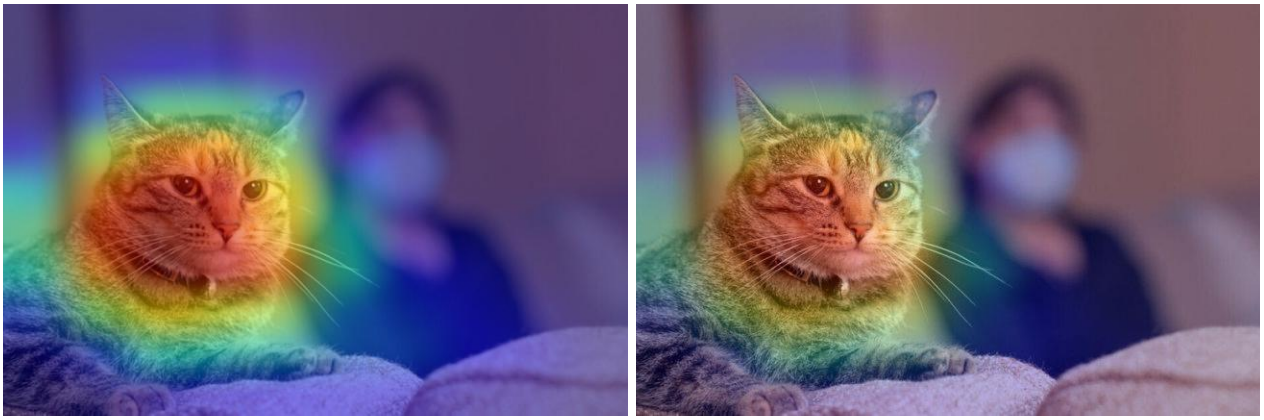

Figure 8: Figure on the left is result of Ablation-Cam and figure on the right is of Grad-Cam both images are of the same model(pre-trained Xception Model)

Here we can see that the Ablation-Cam model has performed much better than the Grad-Cam model . Grad -Cam model focused more on the body of the person/animal whereas the Ablation cam focused more on the face which is more relevant for classification.

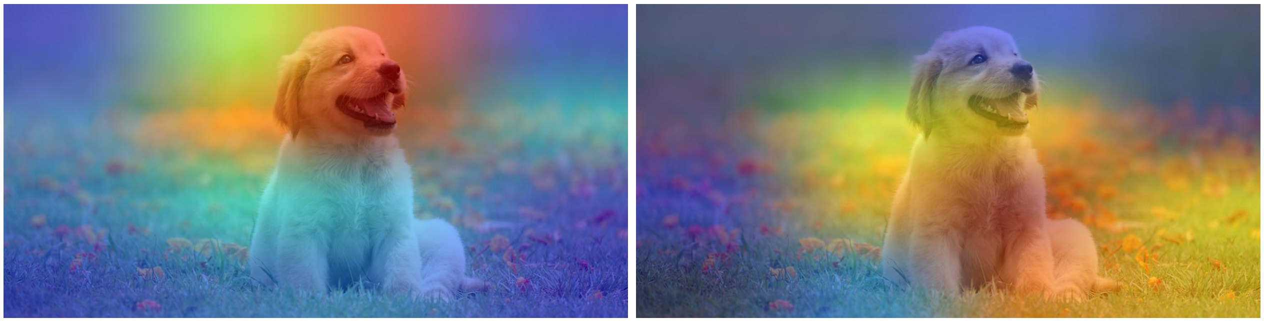

Figure 9: Figure on the left is the result of Ablation-Cam and figure on the right is of Grad-Cam both images are of the same model (Cifar-100)

5 Drawback of Ablation-Cam

In this section, we will discuss few of the drawbacks of the Ablation-Cam model.

The computational time required for Ablation-Cam is much more than that of Grad-cam because in Ablation-Cam we need processing for each activation-map whereas in Grad-Cam it only requires one backward pass.

Running time of Ablation-Cam - 96 seconds on CPU
Running time of Grad-cam - 26 seconds on CPU

As we can see from the comparison results when trained on a pre-trained Xception model the difference between both the cam is very small. This is because Grad-cam performs equivalent when there are no fully connected layers and instead have Global Average Pooling layer.

Also in kind of images where the where there are objects obstructing a bit of image the ablation-cam approach did not perform well.

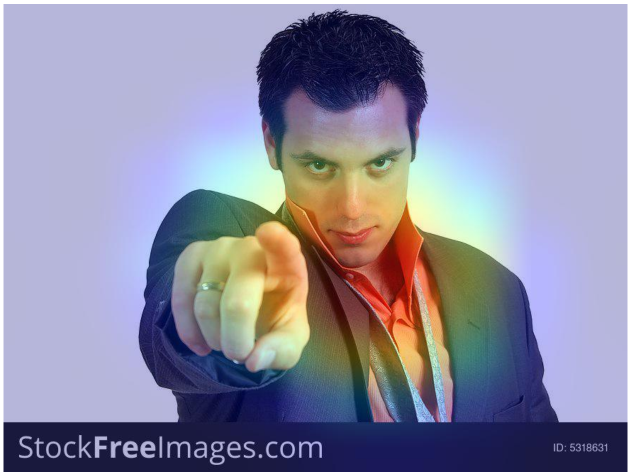

Figure 10: We can see that the model did not perform so well on this kind of images

6 A Different Approach

While Implementing the paper, I had an idea for a different approach with the same ideology.
Instead of making each map ablated and then passing it through the classifier, what if I passed an ablated image (An image with all values 0) and then see which activation maps were still activated. If a feature map was activated that meant, it gets activated without actually interacting with a shape/object.
Further used the same concept of assigning weights to the maps but now on the basis of , if they got activated even if an ablated image is passed through, and then the weight ratio was calculated.
The heat map was finally generated from the activation layer and was superimposed with the original image to see the results.
The results were not as good as Ablation cam be still they can be thought to be used as they are less computationally expensive .
The example of such approach on an image is given below

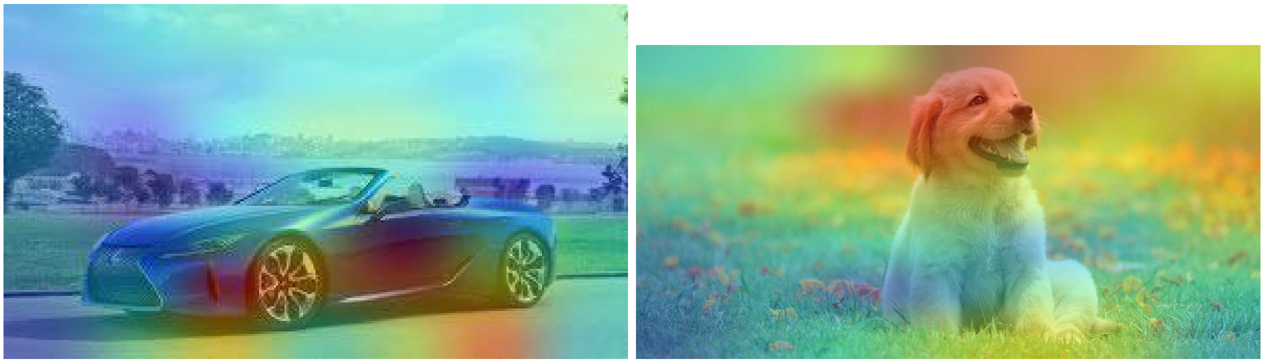

Figure 11: Result of image applied on model with a different approach

7 Future Work

1) Further Research could be done for the results and performance of Ablation-Cam for a model trained adversarial and whether it overcomes the Gabon-Panda adversary attack.

2) Same method can be used in Natural Language Processing to find out the importance of particular words in text / Sentiment classification.

8 Citation

@articleablation-cam,
title=Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization,
author=Saurabh Desai and Harish G. Ramaswamy,
year=2020

@articlegrad-cam,
title=Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization,
author=Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, Dhruv Batra,
journal=arXiv preprint arXiv:1610.02391v4,
year=2019

@articlexception,
title=Xception: Deep Learning with Depthwise Separable Convolutions,
author=François Chollet,
journal=arXiv preprint arXiv:1610.02357v3,
year=2017

9 Refrences

1) https://www.youtube.com/watch?v=e0s80kZHGF0 - Video where this paper has been explained by the author himself

2)https://www.youtube.com/watch?v=3JMKX51dots

3)https://towardsdatascience.com/demystifying-convolutional-neural-networks-using-gradcam-554a85dd4e48 - Article on Grad-Cam

4)https://colab.research.google.com/github/keras-team/keras-io/blob/master/examples/vision/ipynb/gra
$IeQV8J - LWgmc - GradCamCodeUsed.$