Visualizing Deep Features using Class Activation Mapping (CAM)

Contents

List of Figures

Explainable AI - Introduction

The application of AI systems in healthcare is a challenging task mainly because the factors involved in arriving at a decision by the machines are not explainable. Questions like, how did the machine arrive at this decision? or what did the machine see to predict the particular class of a condition? will always be asked to understand a machine's way of taking the decisions in healthcare.

Interpretability matters when machines take decisions on doctor's behalf. For machines to arrive at a particular medical decision, health diagnostic or the treatment course, they have to be trustable. If machine based intelligent systems have to be integrated into the healthcare systems, their decisions have to be meaningful and transparent.

CAM Technique Overview

(Zhou *et al.,* 2015) in their paper titled *"Learning Deep Features for Discriminative Localization"* utilize the Global Average Pooling layer to demonstrate its ability to support localization of objects in am image. Though, it has been majorly used for its regularizing capabilities to prevent overfitting, the authors of the paper say that GAP layer can also be used to retain the spatial structure of the feature maps and identify the discriminative regions of the image.

The output of the final convolutional layers are fed to the fully connected dense layers which results in a loss of the spatial structure. However, performing a global average pooling operation on the convolutional feature maps just before the final softmax layer, *not only retains the spatial structure but also help identify the important regions in the image by projecting back the weights of the output layer onto the convolutional feature maps.*

Class Activation Mappings

A Class Activation Map indicates the discriminative image regions used by the convolutional network to classify that image into a particular class. Since the authors utilize the global average pooling layer, CAM technique is architectural bound such that just before the final output layer

(softmax layer), a *global average pooling is performed on the convolutional feature maps* and the resultant features are fed to a fully-connected (softmax) layer that produces the desired output. This modification in the network connectivity towards the final layers, helps to *identify the important regions of the image by projecting back the weights of the output layer on to the convolutional feature maps*. This technique is known as the class activation mapping.
Figure 1 shows the general overview of CAM implementation utilizing the global average pooling layer. (Image Source - (Zhou *et al.*, 2015))

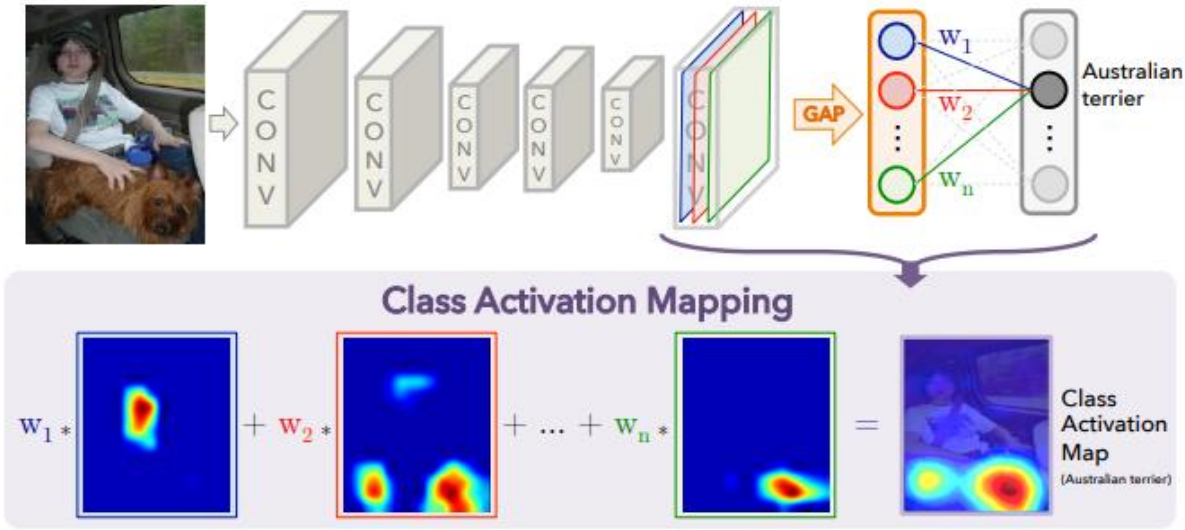

Figure 2. Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class activation maps (CAMs). The CAM highlights the class-specific discriminative regions.

Figure 1- CAM Overview

GAP computes the spatial average of each unit's feature map in the final convolutional layer. A weighted sum of these values gives the final output. Similarly, *a weighted sum of the feature maps from the final convolutional layer results in the class activations.*

CAM Implementation

As discussed above, CAM technique is bound to having the GAP layer just prior to the final softmax classification layer. The full technique is a 2 step process to get the final activation maps.

1. Compute the class activations.
2. Up sample the heatmap to the size of the input image.

Step 1 - Compute Class Activations

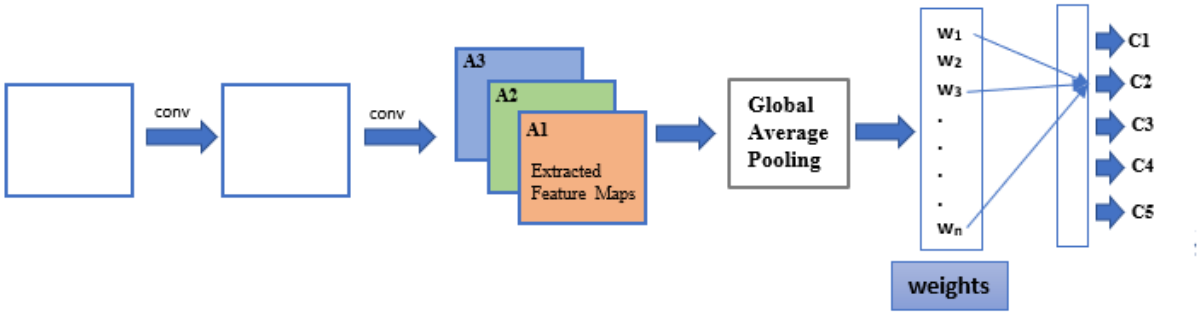

Figure 2- CAM Building Blocks

Let's say, in Figure 2 the input image is classified as class C2. So, the final score for the class C2 is computed as per the formula -

$$y^{c2} = \sum_{k=1}^{K} w_k^{c2} * \frac{1}{z} \sum_{i=1}^{u} \sum_{j=1}^{v} A_{ij}^k$$

where,

$\frac{1}{z} \sum_{i=1}^{u} \sum_{j=1}^{v} A_{ij}^k$ is the result of the Global Average Pooling operation for activation A^k.

$A^k \in \mathbb{R}^{u \times v}$ width - 'u' and height - 'v'.

Now, let M_c be the class activation map for thee class C2. where each spatial element is given by

$$M_c(i,j) = \sum_{k} w_k^{c2} * A^k$$

$M_c(i,j)$ indicates the importance of the activation at the spatial location (i, j) which eventually leads to classifying the image as belonging to the class c2. Each unit in the convolutional layers is activated by some visual pattern within its receptive field. A^k represents the activations of this visual pattern. *The class activation map is simply a weighted linear sum of the presence of these visual patterns at different spatial locations*.

Step 2 - Up Sample Heatmap

The spatial dimension of the final convolutional layer would be too small compared the dimensions of the input image. So would be the dimensions of the computed class activation maps from this layer.

So, for better visualization purpose, the class activation map should be resized to match the dimensions of the input image.

Global average pooling (GAP) vs global max pooling (GMP)

GAP encourages the network to identify the extent of the object as compared to GMP which basically identifies just one discriminative part of it. When computing the average of an activation map, the value can be maximized by finding all discriminative parts of an image as all low activations reduce the output of the particular map. Whereas, for GMP, since the maximum value is considered, the low scores for all image regions except the most discriminative one do not impact. The authors show that while GMP could achieve similar classification performance as GAP, *GAP outperforms GMP for localization*. So, global average pooling is considered a better choice when activation maps are computed for achieving localization.

Localization

To achieve localization, the authors segment out those regions in the image where the heatmap has a value 20% of the maximum value of the class activation.

CAM Visualizations

A few CAM viz from the experiment for visualizing the classification of skin cancer tumors as ***benign*** or ***malignant*** are shown in Figure 3, Figure 4 and Figure 5. Images for training have been referenced from (Tschandl, Rosendahl and Kittler, 2018; Codella *et al.*, 2019) respectively.

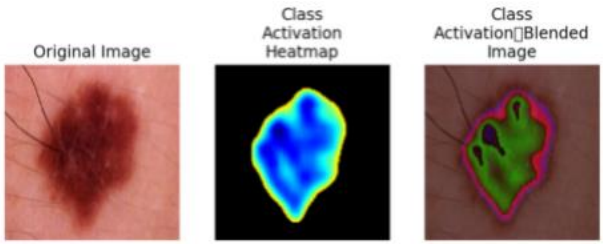

Figure 3- CAM Viz 1

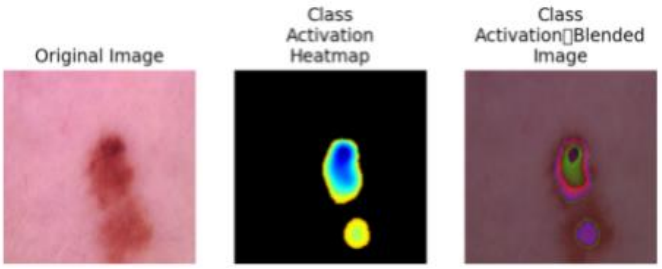

Figure 4- CAM Viz 2

Figure 5- CAM Viz 3

The visualizations seen in the images above, show the original image and the up-sampled heatmap. The blended image shows that the heatmap coincides with the location of the tumor while having good amount of intensity at the tumor location.

Conclusion

The visualizations seen in the images above, show the original image and the up-sampled heatmap. The blended image shows that the heatmap coincides with the location of the tumor while having good amount of intensity at the tumor location and hence provide a good amount of explainability showing where the model looks at in an image.

A quick summary -

1. CAM can be utilized for a weakly supervised object localization.
2. CAM is bound by architectural constraints, i.e., only those architectures performing Global Average Pooling over convolutional maps immediately before final softmax layer could take advantage of the CAM visualizations.
3. The modified model needs to be retrained, which could computationally expensive when trained over the SOTA CNN architectures.
4. Since the fully connected Dense layers are replaced the performance of the model can suffer and the prediction score may the actual picture of the model's ability to classify images.

Further Reading

1. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. Available at - https://arxiv.org/abs/1610.02391
2. Grad-CAM++: Improved Visual Explanations for Deep Convolutional Networks. Available at - https://arxiv.org/abs/1710.11063
3. Tell Me Where to Look: Guided Attention Inference Network. Available at - https://arxiv.org/abs/1802.10171

References

Codella, N. *et al.* (2019) 'Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)'. Available at: http://arxiv.org/abs/1902.03368 (Accessed: 27 May 2021).

Tschandl, P., Rosendahl, C. and Kittler, H. (2018) 'The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions', *Scientific Data*, 5(1), p. 180161. doi: 10.1038/sdata.2018.161.

Zhou, B. *et al.* (2015) 'Learning Deep Features for Discriminative Localization'. Available at: https://arxiv.org/abs/1512.04150 (Accessed: 7 June 2021).

Interpretability in Deep Learning - https://towardsdatascience.com/interpretability-in-deep-learning-with-w-b-cam-and-gradcam-45ba5296a58a