CAM

October 22, 2020

1 Introduction

Paper [1] presents CAM. The aim is to exploit the global average pooling layer and demonstrate how it enables the CNN to have a higher and remarkable localization ability, despite not being trained for them. Moreover, the paper explains how to build CAMs in the ego RNN project [2].

2 Main theoretical concepts

Convolutional units of various layers of CNNs actually behave as object detectors despite no supervision on the location of the object was provided. Despite they have this ability to localize objects in the convolutional layers, this power is lost when fully connected layers are used for classification. So, there is the proposal of avoiding the use of the fully connected layers in view to minimize the number of parameters while maintaining high performances.
To achieve this goal, a global average pooling (GAP) [3] results to be better than the max pooling (most used until this project). This layer acts as a structural regularizer, preventing overfitting during training. Moreover, with a little tweaking, the network can retain its remarkable localization ability until the final layer. So, it is easier to identify the discriminative image regions in a single forward-pass for a variety of tasks. More technically, GAP layers perform a stronger type of dimensionality reduction, where a tensor with dimensions hwd is reduced in size to have dimensions $11d$. GAP layers reduce each hw feature map to a single number by simply taking the average of all hw values, as you can see in figure 1. Max pooling: it limits the localization to a point lying in the boundary of the object rather than determining the full extent of the object;
Average pooling: it encourages the network to identify the complete extent of the object. The loss for average pooling benefits when the networks identifies ALL discriminative regions in an object.

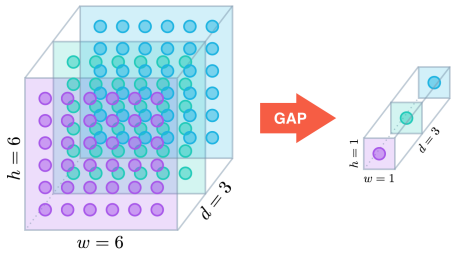

Figure 1: Global average pooling schema

3 The process

CAM: it indicates the discriminative image regions used by the CNN to identify a particular category. The architecture: it largely consists of convolutional layers and, just before the final output layer (softmax in classification), a global average pooling is applied on the convolutional feature maps. These features are used as features for a fully connected layer which produces the desired output. Given this structure, it will be possible to identify the importance of the image regions by projecting back the weights of the output layer on the convolutional feature map (CLASS ACTIVATION MAPPING).

4 Pseudocode

- Global average pooling outputs the spatial average of the feature map of each unit at the last convolutional layer;
- A weighted sum of these values is used to generate the final output;
- Compute a weighted sum of the feature maps of the last convolutional layer to obtain our class activation maps.

References

[1] *CAM's reference paper at.* URL: `https://arxiv.org/pdf/1512.04150.pdf`.

[2] *ego RNN's reference paper at.* URL: `https://arxiv.org/pdf/1807.11794.pdf`.

[3] *Global average pooling's reference article at.* URL: `https://alexisbcook.github.io/2017/global-average-pooling-layers-for-object-localization/`.