# Learning Unsupervised Localized Activation Maps through Classification Model

# Abstract

Most deep learning models, although powerful, acts like a black box where it is hard to explain a model’s decision-making process. We review the Class Activation Mapping (CAM) technique that modifies the last layer of a classification model on some image dataset, so that in addition to the usual logits’ prediction, one can view a map that tells us which region from the image activates such prediction. When training an image classifier, researchers found out that the model can learn a certain map indicating the region of interest on the image without explicit supervision. Such activation map can be used to explain the reason of classification. Our report will be focus on building a simple baseline model on creating the activation maps of our datasets, and then improve the architecture hoping to generate more precise activation maps. We use the Flower-17 dataset that is available on Kaggle to experiment the CAM technique, which has not been done before. To speed up our experiment, we use the feature extractor of a pretrained VGG16 model and freeze its parameter during training, which allow efficient memory usage when storing our trained models.

# Introduction and Background

We have seen competitive advancements on deep learning in the past few years, and it is only going faster. Among all of the deep learning categories, image classification remains the most researched and there are plenty of state-of-the-art models available. Some examples include Residual Neural Network (ResNet) that alleviates gradient vanishing issue in a very deep neural network (Kaiming He et al., 2016; Ross Wightman et al., 2021) and the YOLO series that keep improves the result of classification (Chien-Yao Wang et al., 2022).

With such improvements on classification accuracies, it strikes curiosity in our heart to wonder how exactly does a classification model make a decision that an image belong to one of the many classes? This process is kind of like mathematical theories, where people first assume something heuristically work and then take plenty of times to justify the rationale of its mechanics, in contrasts to common beliefs that mathematical theorems are first rigorously proved before widely used. However, current deep learning model, especially the ones involving convolutional neural network (CNN) layers are too complex to be able to clearly justified, thus we might take a step back and try to explain better than what we can for now. Class Activation Mapping (CAM) is one of such technique that improves the interpretability of a classifier (Bolei Zhou et al., 2015), as shown in Figure 1. The CAM technique clearly explain the decision of classifying the data as “Australian terrier” by showing that the last convolutional layer is activated by the “dog” part of the image, rather than the “human” part. In the discussion below we will use CAM to refer to the class activation mapping technique and the class activation map itself interchangeably.

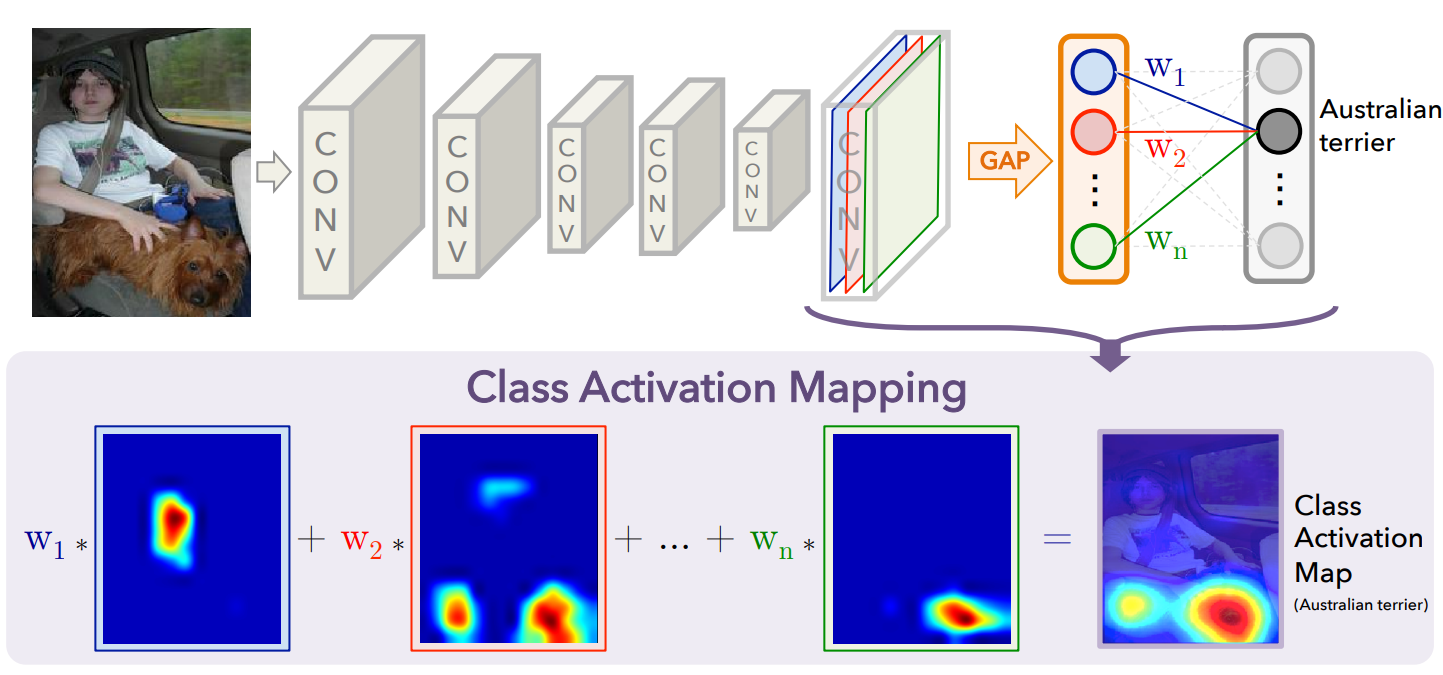


Figure : CAM visualization. (Image taken from Bolei Zhou et al., 2015)

The importance of CAM can sometimes help explain some ambiguity in a classifier, especially when the dataset is biased. For example, one classifier could be handling a task of classifying animals, and one of the classes is “polar bear”. It is highly possible that all of the “polar bear” images­ are taken in a snow-white scenery because it is unlikely for them to live in a more tropical environment, classifier could falsely rely on the white scenery for identifying “polar bear” class, which might lead to false classification when there is a white fox in a snow scenery. Therefore, we can analyze the CAM of some “white bear” images to know the model’s CAM is relying on the snow scenery or the bear itself, informing us potential issue with the model in advanced.

Our main objective is to investigate various implementations of CAM technique by using a pretrained VGG16 network and analyze its result critically. Moreover, different implementations will be carried out to generate fine grained CAM to better visualize the region of interest by the classifier. This project is mostly interesting in a particular way, that CAM generation can be done in a weakly supervised manner. This is because it only relies on the image level label where no mask is given. This provides another practical entry where instance segmentation is needed without ground truth masks at hand (Yang Liu et al., 2022; Zhaozheng Chen et al., 2022).

# References