Let's check Grad-CAM

Jae-uk Shin
Earth
Korea Earth
jack@earth.com

Dolphin
Earth
Ocean of Earth
dd@earth.com

Abstract

In classification problems using CNN architectures, even when results are good, it isn't easy to understand where the object was located and why such results were produced. CAM(Class Activation Mapping) solved this problem. CAM calculated the weights that influence the final feature map to generate a heat map. However, it has the limitation that it must satisfy the CNN-GAP-FC layer structure. This paper examines Grad-CAM, which overcomes this structural issue. Grad-CAM utilizes the derivative value obtained through back-propagation as the weight. Consequently, it can be applied without modifying the model architecture. Furthermore, it enables one to view results from intermediate layers. The results were validated using the "ResNet 50 model" and the "Stanford dogs dataset".

1. Introduction

In classification problems using CNN architectures, even when the results are good, it is difficult to understand where the object was located and why such results were produced. CAM (Class Activation Mapping) solved this problem[3]. CAM solved this problem by generating a heatmap from the values obtained by multiplying the feature maps of the last convolution layer by weights. However, it has the limitation that it must satisfy the CNN-GAP-FC layer structure. This limitation may necessitate partial structural changes and retraining of the model. It also has the drawback of only showing results for the final feature. This paper examines Grad-CAM[2], which was proposed to overcome these shortcomings. Grad-CAM modifies the structure for obtaining weights. Unlike CAM, it uses the derivative values obtained through back-propagation from the result back to the feature map as weights. Consequently, it produces results similar to CAM without requiring model structural changes. Furthermore, it allows viewing intermediate layer results, enabling observation of step-by-step changes from the input stage to the output stage. This paper conducted experiments using the "ResNet50 model" and the "Stanford dogs" dataset[1]. The experimental results confirmed that Grad-CAM can be used to produce results similar to CAM without modifying the model structure. It also confirmed that changes in results due to variations in layer depth can be observed.

2. Related Work

2.1. CAM

CAM calculates the heatmap by multiplying the feature maps from the last convolution layer by weights. After obtaining the heatmap for each feature map, they are all summed to produce the final heatmap. To determine the weights to multiply each feature map, the last layer needed to be modified. In the classification model combining CNN and FC layers, the FC layer was removed. Instead, we applied GAP (Global Average Pooling) to the final feature map and connected it to an FC layer. We then multiplied the weights of this new FC layer by the feature maps from the last convolution layer to generate the heatmap. The result visually represented the location of the classified object and how much the model focused on different areas to produce that result.

However, it has three structural drawbacks. First, a FC layer must always follow GAP. Therefore, modifying part of the model structure may be necessary depending on requirements. This carries the potential issue of reflecting distorted results. Second, additional training is required to obtain the weights for this newly configured FC layer. Third, heatmaps can only be generated for the feature map produced by the final convolution layer.

3. Grad-CAM

To address these issues, the Grad-CAM architecture was proposed. The most significant structural limitation of conventional CNNs is the mandatory requirement for a GAP-FC layer structure to obtain the weights for constructing heatmaps. Grad-CAM obtained these weights by calculating the derivative through backpropagation from the output to each feature map. The derivative values computed for

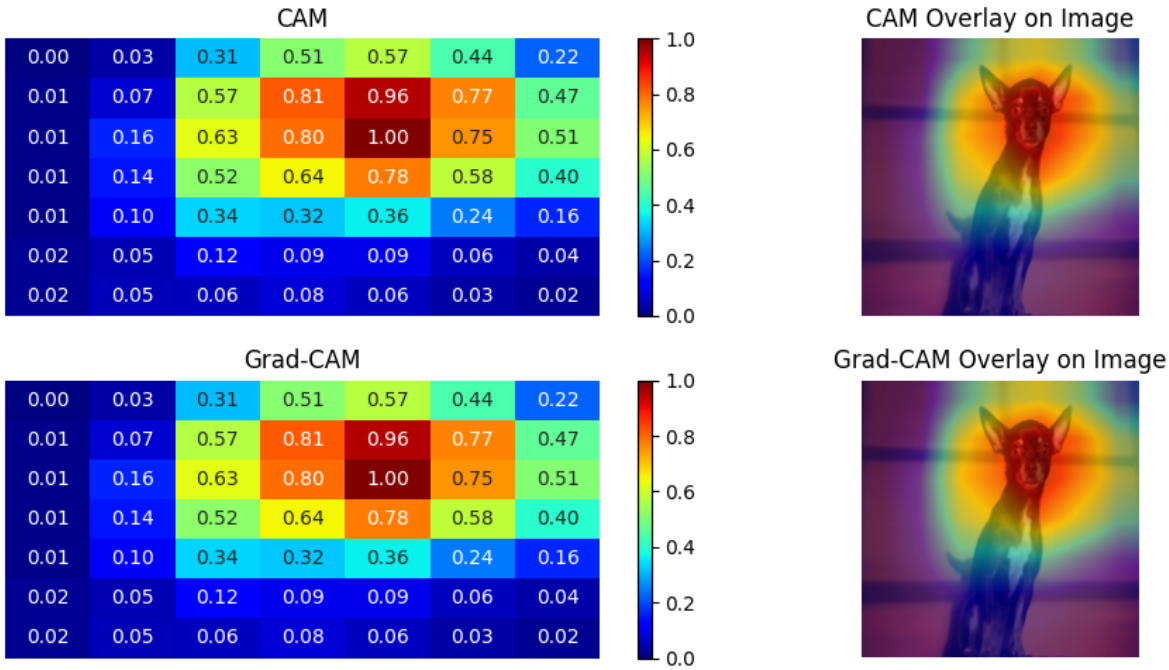

Figure 1. CAM vs Grad-CAM : The CAM and Grad-CAM outputs for the last layer in the same layer structure are identical.

each pixel in the feature map are aggregated into a single weight via GAP. Subsequently, similar to CAM, each feature map is multiplied by its weight to generate heatmaps, and all heatmaps are summed to produce a final heatmap. This structural change yields the same results as CAM while resolving CAM's inherent structural issues. In CNN-GAP-FC models, CAM and Grad-CAM produce identical results. FC is a linear combination of the form y = ax. CAM uses the coefficient a that produces the model's output, while Grad-CAM uses the differentiated gradient (a), because they ultimately use the same value. Strictly speaking, a linear scaling difference exists in the final heatmap result, but it disappears after min-max normalization at the end.

However, differences arise when the final layer is not GAP-FC. CAM requires retraining with the final layer converted to GAP-FC form, whereas Grad-CAM can use the currently trained model without additional training. Furthermore, Grad-CAM can be applied without layer modification, enabling its use in intermediate layers, unlike CAM which is only applicable to the final layer.

4. Experiments

4.1. Grad-CAM, Stanford dogs Classification

The model used in the experiment was a pretrained ResNet50 model modified to have only 120 outputs in the final FC layer. The modified model was fine-tuned using the "Stanford dogs" dataset, which has 120 classes. To compare CAM and Grad-CAM under identical conditions, a CNN-GAP-FC architecture model was employed. Therefore, no additional model modifications or retraining were required for CAM.

ResNet50 is broadly divided into four layers. Since CAM can only be applied to the final layer, only one result was displayed. Grad-CAM can also obtain heatmaps from intermediate layers, so a total of four results were displayed. The intermediate layer heatmaps are expected to show features considered important at the input stage spread out broadly, while higher-dimensional features extracted as the model progresses toward the output stage become concentrated in specific regions.

4.2. Result

First, examining the results of the final layer confirms that the CAM and Grad-CAM results match. As explained earlier, although a linear scaling difference exists in the heatmap results, applying min-max normalization at the end demonstrates that they become completely identical(Figure.1).

Furthermore, Grad-CAM allows us to examine results from intermediate layers. The findings align with expectations. At lower-level layers, simple features like points and lines are extracted, causing areas of interest to be scattered across the entire image. As we move to higher-level layers, higher-order features are extracted, and we observe that areas of interest gradually become more concen-

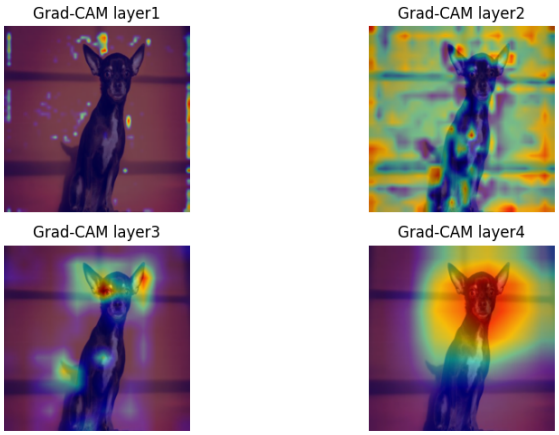

Figure 2. Grad-CAM results by layer

trated(Figure.2).

5. Conclusion

Grad-CAM was proposed to overcome the structural limitations of CAM. CAM enabled visual confirmation of where objects were located and why they produced certain results in classification problems using CNN structures. However, it had structural limitations that restricted its applicability to specific architectures. Grad-CAM overcomes these structural limitations while producing the same results as CAM. It requires no additional training since it can be applied without structural modifications, eliminating the distortion of model outputs. Furthermore, it allows observation not only of results for the final layer but also for intermediate layers, enabling the tracking of feature changes and shifts in focus areas as the model deepens.

5.1. Future works

However, it has limitations in locating the object's position. Grad-CAM highlights regions within the object that significantly influence the classification result. This does not reflect the entire area where the object is located. Therefore, further research is needed to use Grad-CAM results to display the object's location in the form of a bounding box.

References

[1] http://vision.stanford.edu/aditya86/imagenetdogs/main.html. 1

[2] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 1

[3] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 1