Grad-CAM pre-report

Taehun Kim[1]
[1] Department of Computer Science and Engineering, Pusan National University.

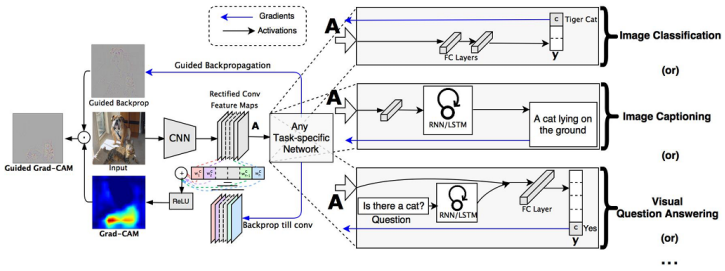

Figure 1: Grad-CAM overview

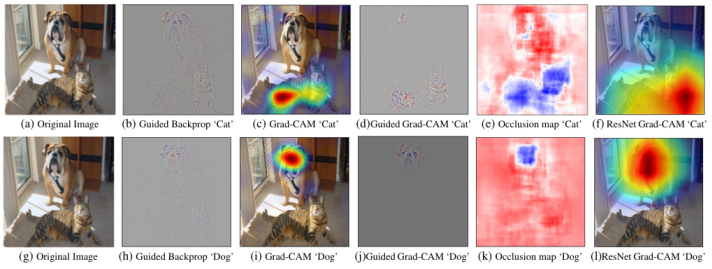

Figure 2: Guided Backpropagation, Grad-CAM, and Guided Grad-CAM

1 Introduction

Class Activation Mapping(CAM)[3] modifies image classification CNN architectures replacing fully connected layers with global average pooling and a 1×1 convolutional layer for visual explanation. this approach cannot be applied to any other tasks except the image classification, and need to re-train.

In contrast, Gradient-weighted Class Activation Mapping(Grad-CAM)[1] is the technique for producing visual explanations, without modifying the CNN architecture or re-training the model. Therefore this approach can be applied to various CNN model like image classification, segmentation, object detection, and image captioning.

This report summarize the Grad-CAM paper[1] with emphasis on its approach and architectural design.

2 Grad-CAM

The last convolutional layer in a CNN has the best compromise between high-level semantics and spatial information. Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN. This method identifies the regions that contributed to the CNN's decision-making process.

Let activation maps of last convolutional layer is $A \in \mathbb{R}^{H \times W \times K}$, which $H \times W$ is resolution of the activation map and K is the number of the activation maps. Let y^cdenote the class score for class c before the softmax activation. First, The paper computes gradient of class score y^c with respect to the kth map, ith column, jth row of A. Second, They apply global average pooling to the gradient to get the gradient of class score y^c with respect kth A, α_k^c:

$$\alpha_k^c = \frac{1}{HW} \sum_{i=1}^{W} \sum_{j=1}^{H} \frac{\partial y^c}{\partial A_{ij}^k}$$

, where the A_{ij}^k is the kth map, ith column, jth row of activation maps of the last convolutional layer. Since A is the last convolutional layer, the gradient calculation involves successive matrix multiplications of the weight matrices and the gradient with respect to activation from class score to last convolutional layer through Backpropagation(chain rule). α_k^c captures the importance of kth feature map for a target class c.

Additionally, They calculate localization map for class c using α_k^c,A^k, and ReLU:

$$L_{Grad-CAM}^c = ReLU(\sum_k \alpha_k^c A^k)$$

The product of α_k^c and A^k highlights the significant regions within the activation map A^k. ReLU is applied to retain only features with positive contributions, as negative values likely correspond to other classes.

$$\text{ReLU}(x) = \begin{cases} 0 & \text{if } x \leq 0, \\ x & \text{if } x > 0. \end{cases}$$

2.1 Guided Grad-CAM

As shown in the figure 2, Grad-CAM can localize region relevant to the class, but lacks to highlight fine-grained details. However, Guided Backpropagation[2], which is visualize gradient that suppress the negative gradient(and activation) during the backpropagation, capture the details of the objects, but not class-discriminative.

The authors combine Guided Backpropagation and Grad-CAM through element-wise multiplication to integrate both information. Grad-CAM output is upsampled to the input image size for multiplication. This fusion can visualize the interesting region with high-resolution and class-discriminative.

3 Conclusion

We summarize the Gradient-weighted Class Activation Mapping (Grad-CAM) paper [1]. Grad-CAM visualizes which parts of an image are important for a specific class without modifying the pre-trained model architecture or requiring re-training.

Additionally, Guided Grad-CAM provides more precise visualizations by combining Grad-CAM with Guided Backpropagation [2].

[1] 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. *International Journal of Computer Vision*, 128(2):336–359, October 2019. ISSN 1573-1405. doi: 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-01228-7.

[2] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net, 2015. URL https://arxiv.org/abs/1412.6806.

[3] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization, 2015. URL https://arxiv.org/abs/1512.04150.