Student ID : 20194293

Name : Go, Kyeong Ryeol

[AI 502] Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

1. Paper Summary

While deep neural networks show remarkable performance in various tasks, they lack interpretability where decomposing into intuitive and understandable components is difficult. However, the transparency must be equipped to be reliably deployed so that identifying failure modes, establishing trust and confidence in users, and machine teaching can be done.

The author proposed a class-discriminative localization technique which is named as Grad-CAM. Comparing to Class Activation Mapping (CAM), Grad-CAM does not alter the structure of the model. Thus, this made existing state-of-the-art deep models interpretable avoiding the trade-off between accuracy and interpretability. Moreover, Guided Grad-Cam is also devised which combines Grad-CAM with the Guided back-propagation so that high-resolution was further achieved by capturing fine-grained details in the image.

The last convolutional layers are expected to have the best compromise between high-level semantics and detailed spatial information. Therefore, the Grad-Cam used the gradient flowing into this layer to understand the importance of each neuron for the decision of interest. The gradient of the score for class (y^c) with respect to feature map (A^k) are global-average-pooled to obtain the neuron importance weights (α_k^c). Then, a weighted combination of forward activation maps is activated by ReLU to focus on the pixels whose intensity should be increased in order to increase the class score. As a result, the class discriminative localization map $L_{Grad-CAM}^c$ can be computed as follows. Here, the Guided back-propagation can be fused via point-wise multiplication on the up-sampled Grad-CAM to the input image resolution using bi-linear interpolation.

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k} \qquad L_{Grad-CAM}^c = \mathrm{ReLU}\left(\sum_k \alpha_k^c A^k\right)$$

The experiments show that the proposed approaches outperform all existing approaches on both interpretability and faithfulness to original image. Furthermore, extensive human studies reveal the truth-worthiness of a classifier where discriminating between classes and identifying the bias in data were more accurate. Finally, the broad applicability to various off-the-shelf available architectures was verified for tasks like image classification, image captioning and visual question answering by providing faithful visual explanations for possible model decisions.

2. Discussion

Here I would like to offer 2 discussion points. To begin with, how can we quantitatively measure the quality of visualization? This paper mainly utilized the human studies, but it is preferable to measure its quality so that the model parameters can be further adapted to maximize the metric. Since it's hard to objectively label the original data, it must be processed in unsupervised setting. I suggest the metric learning through knowledge distillation where the teacher network is the well-pretrained model such as ResNet and the student network is the shallow network with several convolutional layers. Next, instead of the last convolutional layers, what can we do to interpret the intermediate convolutional layers? I suggest drawing a histogram of hidden neurons activities and the filters.