GRAD-CAM

Visual Explanations from Deep Networks via Gradient-based Localization

Introduction:

The paper proposes a technique for producing visual explanations for decisions from a large class of Convolutional Neural Network (CNN)-based models, making them more transparent. Gradient-weighted Class Activation Mapping (Grad-CAM), uses the gradients of any target concept (say logits for 'dog' or even a caption), flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

Contribution:

Unlike previous approaches, Grad-CAM is applicable to a wide variety of CNN model-families:

- CNNs with fully-connected layers (e.g. VGG)
- CNNs used for structured outputs (e.g. captioning)
- CNNs used in tasks with multi-modal inputs (e.g. VQA) or reinforcement learning

without architectural changes or re-training. Grad-CAM helps untrained users successfully discern a stronger deep network from a weaker one.

Approach:

The last convolutional layers have the best compromise between high-level semantics and detailed spatial information (which is lost when we move to the fully connected layers). The neurons in these layers look for semantic class-specific information in the image (say object parts). Grad-CAM uses the gradient information flowing into the last convolutional layer of the CNN to understand the importance of each neuron for a decision of interest.

In order to obtain the class-discriminative localization map Grad-CAM $L^c_{Grad-CAM} \in \mathbb{R}^{u \times v}$ of width u and height v for any class c , we first compute the gradient of the score for class c, y^c (before the softmax), with respect to feature maps A^k of a convolutional layer, i.e. $\frac{\partial y^c}{\partial A^k}$. These gradients flowing back are global average-pooled to obtain the neuron importance weights α^c_k:

$$\alpha^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}}$$

This weight α^c_k represents a partial linearization of the deep network downstream from A, and captures the importance of feature map k for a target class c. Then perform a weighted combination of forward activation maps, and follow it by a ReLU to obtain,

$$L^c_{Grad-CAM} = ReLU(\sum_k \alpha^c_k A^k)$$