Lab9 Pre-Report
Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Jiho Park, 2019142056
School of Electrical and Electronic Engineering, Yonsei University

Abstract

By sacrificing interpretability, deep neural networks have achieved great performance in accuracy. This paper proposes the method to avoid this trade-off. Grad-CAM is a visual explanation technique for CNN (Convolutional Neural Network)-based models. The linear combination of the last convolution filters with the weight 'neuron importance', gives us the activation map. The neuron importances are computed with the gradient of the each neuron of the filter. This Grad-CAM technique does not need any modification in architecture or re-training even with state-of-the art deep models, thus avoids the interpretability vs. accuracy trade-off. Also it can be applied to various models(e.g. CNN with fully-conected layer, CNN with multi-modal inputs or ouputs). Finally, Grad-CAM provides a 'good' visual explanation since it's activation map is class-discriminative and high-resolution.

1 Theoretical Explanation

Grad-CAM mainly follows the idea of Class Activation Map(CAM) [1], but overcomes it's drawbacks. Since the last convolutional layer remains the spatial information and has the best class-specific information, the class activation map is obtained by linear combination of the last feature maps. The CAM and Grad-CAM are different in the method to obtain the weight of this linear combination. This weight is called 'neuron importance' in this paper.

1.1 CAM

The CAM obtains the neuron importance of each feature map from the last fully-connected layer, which is for classification. So there exists C(# of class) different weights for each feature map. The weight represents the importance of the corresponding feature map for corresponding class. This method successfully obtains the activation map of the CNN-based models. However, drawbacks still remains. Since CAM requires the weights of each feature map, 'global average pooling → fully-connected layer → softmax classification' architecture is necessary. So for models without this architecture, additional training with architecture modification is needed.

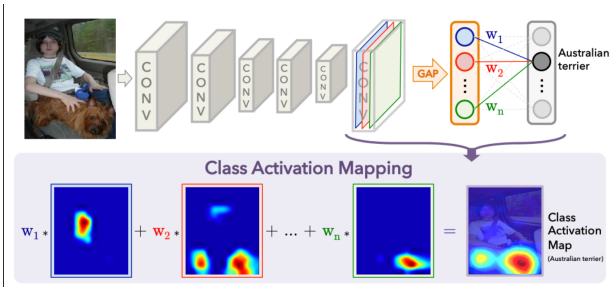

Figure 1: Class Activation Map method

1.2 Grad-CAM

Grad-CAM proposes the method using the gradient to obtain the neuron importance. The partial derivative can represent the importance of each neuron. Since the partial derivative can be obtained by back-propagation, the neuron importance can be computed regardless of the model architecture.

Grad-CAM method computes the average gradient of each feature maps for the neuron importance α_k^c(1). Computed neuron importance is identical with the neuron importance from the CAM(The mathematical proof is shown in the paper.). As a result, Grad-CAM is generalized version of CAM.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{i,j}^k}}_{\text{gradients via backprop}} \tag{1}$$

For final activation map, ReLU function is applied (2). The reason for applying ReLU, is the influence of the negative pixels. Author says that negative pixels are likely to belong to other categories in the image [2]. Also empirically without ReLU, the activation map sometimes highlighted more than just the desired class and performed worse [2]. They also found out that negative pixel leads to counterfactual explanation.

$$L_{\text{Grad-CAM}}^c = ReLU\Big(\underbrace{\sum_k \alpha_k^c A^k}_{\text{linear combination}}\Big) \tag{2}$$

1.3 Guided Grad-CAM

Guided Grad-CAM is to highlight the fine-grained details, which Grad-CAM itself can't do. This is done by fusing the Guided Back-Propagation and Grad-CAM visualizations by element-wise multiplication. Guided Backpropagation visualizes gradients of the input neuron with ReLU layer. As in Fig 2 (b) and (h), Guided Backpropagation does catch the fine-grained details but it's not class-discriminative. Therefore, the element-wise multiplication with the Grad-CAM localization map makes the class-discriminative highlight(Fig 2 (d), (j)). As a result, Gudied Grad-CAM method achieves the class-discriminative and high-resolution visual explanation.

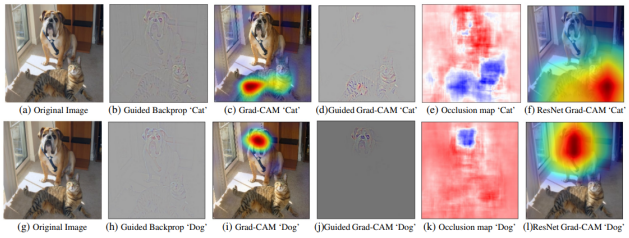

Figure 2: Localization Visualization Results

Grad-CAM methods can be applied to any differentiable CNN-based models. The final application architecture is as Fig 3.

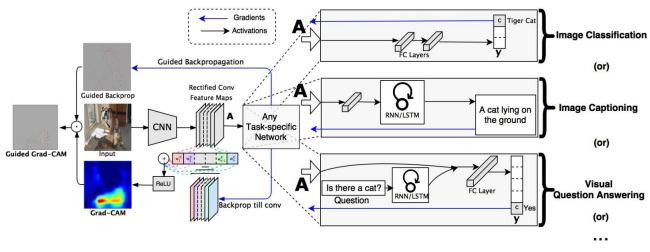

Figure 3: Grad-CAM Architecture

2 Evaluating Localization

Two weakly-supervised localization tasks were evaluated, bounding-box prediction and semantic segmentation. For both evaluation, only image annotations were shown to the model during training, without ground-truth bounding-box or segmentation.

Following ILSVRC-15, the top-1 and top-5 bounding-box errors was evaluated. Grad-CAM achieved better performance than c-MWP, Simonyan et al, and CAM.

For weakly-supervised segmentation, work of Kolesnikov et al, is used. Although, their algorithm achieved great performance, it was sensitive to the choice of the localization seed. By replacing CAM into Grad-CAM, 49.6 IoU(Intersection over Union) score was achieved, which was 44.6 previously. The segmentation example is shown in Fig 4.

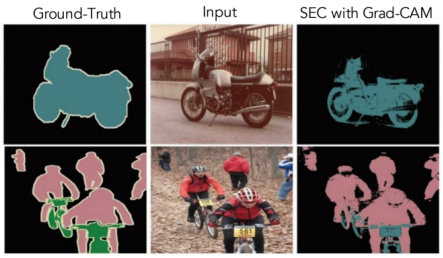

Figure 4: Segmentation result of SEC with Grad-CAM as seed

3 Evaluating Visualization

Class-Discrimination

Intuitively, a good visualization is one that produces class-discriminative explanation. In the experiment, human subjects make choice of class from the Guided Grad-CAM and Guided Back-propagation results(Fig 5 (b)). As a result, Guided Grad-CAM increased the human performance by 16.79%, which was 44.44 with Guided Back-propagation. Additionally, Grad-CAM also improved class-discriminativity of the models with Deconvolution.

Trust

The trust to the deep learning models were evaluated by AMT interface(Fig 5 (c)). The Guided Grad-CAM result and Guided-backprop with Deconvolution were compared. Even with same classification result, Guided Grad-CAM achieved higher score. Thus, their visualization actually help human to trust deep learning models.

Faithfulness and Interpretability

In general, the faithfulness and interpretability have relation of trade-off. Faithfulness is its ability to accurately explain the trainned mapping. Through localization experiment and occlusion map correlation, author claim that Grad-CAM visualizations are more interpretable, and more faithful to the model.

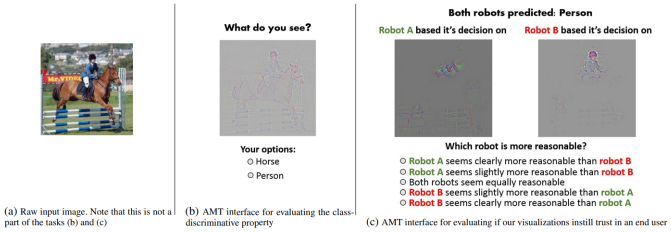

Figure 5: Class-discriminaivity, Trust Evaluation

4 Applications

4.1 Diagnosing CNNs

The failure modes of classification are occasionally seen in deep convolution models. The Guided Grad-CAM helps to explore the reason of the model's failure. As in Fig 6, we can see that seemingly unreasonable predictions have reasonable explanations [2].

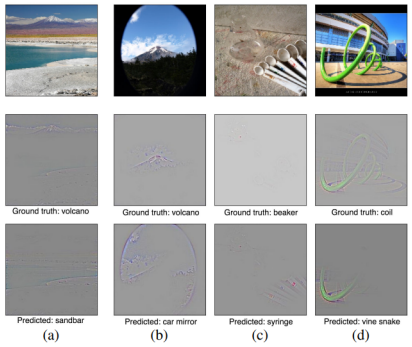

Figure 6: Failure Mode Visualization Explanation

Adversarial Attack: The Grad-CAM method is robust to adversarial attack. Even the classification gone wrong, the localization of class in interest are accurate still.

Bias in datatset: The Grad-CAM method also helps to find the bias in dataset. In experiment of the nurse and doctor classification, they found out that model looks person's face and hairstyle to distinguish nurses. Since this make the model struggle to distinguish male nurse, they can diagnose to add male nurse images to the dataset.

4.2 Image Captioning and VQA

Grad-CAM method also works well in multi-modal input/outputs. The neuraltalk2 was used for image captioning experiment. By the method of Bau et al, neurons of the last convolution layer were automatically named after training. As in Fig 7, we can confirm that important concept is well activated in visualization result.

For VQA(Visual Question Answering), they found out that the activation map is relevant to the answer of the question. Even with the same question but different answers, the Grad-CAM visualization captured the property of the corresponding answer. These results prove the interpretability of Grad-CAM in multi-modal tasks.

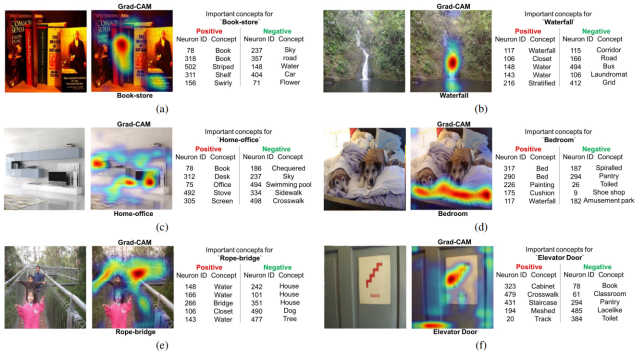

Figure 7: Important Concept Visualization of Image Captioning

Reference

[1] L. A. A. Oliva B. Zhou, A. Khosla and A. Torralba. Deep features for discriminative localization. *CVPR*.

[2] Abhishek Das Ramakrishna Vedantam Devi Parikh Dhruv Batra Ramprasaath R. Selvaraju, Michael Cogswell. Grad-cam: Visual explanations from deep networks via gradient-based localization.