DEEP LEARNING LAB(EEE4423-01) Week 9 - Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Hyeon Si Eun[1], 2019142214
[1] Department of Electric and Electronic Engineering, Yonsei University

Introduction

The paper "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization"[1] proposes a technique called Gradient-weighted Class Activation Mapping (Grad-CAM) for producing visual explanations for decisions made by Convolutional Neural Network (CNN)-based models. The goal is to make these models more transparent and explainable. Grad-CAM uses the gradients of any target concept flowing into the final convolutional layer to produce a coarse localization map highlighting the important regions in the image for predicting the concept.

The authors begin by discussing the importance of transparency and interpretability in deep learning models, especially in applications such as medical diagnosis, autonomous driving, and criminal justice. They argue that visual explanations can help humans understand how these models make decisions and build trust in them. They also discuss existing techniques for visualizing CNNs such as Deconvolutional Networks, Guided Backpropagation, and CAM[3].

The authors then introduce Grad-CAM, which is a simple yet effective technique that can be applied to any CNN-based model without requiring any architectural changes or additional training. Grad-CAM produces class-discriminative localization maps that highlight the important regions in an image for predicting a particular class. The authors show that Grad-CAM outperforms existing techniques such as CAM and Guided Backpropagation on various metrics such as localization accuracy, class-discriminativeness, trustworthiness, and faithfulness.

The paper also shows how Grad-CAM can be used to diagnose image classification CNNs by identifying which parts of an image are most important for making a correct or incorrect prediction. The authors demonstrate this on several datasets such as CIFAR-10, CIFAR-100, and ImageNet. They also show how Grad-CAM can be used to identify biases in datasets by visualizing which parts of an image are most important for predicting different classes.

The paper then shows how Grad-CAM can be extended to obtain textual explanations by combining it with attention mechanisms in vision and language models such as image captioning and Visual Question Answering (VQA). The authors show that Grad-CAM can produce more accurate and interpretable explanations than existing techniques such as attention-weighted saliency maps.

What makes a good visual explanation?

Let's consider image classification. A 'good' visual explanation from a model to demonstrate the validity of a certain target category must be (a)class-discriminative (i.e. you are able to locate the category within the image) and (b)high-resolution.(i.e. you need to capture fine-grained details.)

Figure 1 shows a number of visualization outputs for the 'tiger cat' class (top, cat) and 'boxer' class (bottom, dog). Pixel-space gradient visualizations, such as guided back-propagation and deconvolution, are high-resolution and emphasize fine-grained detail within the image but are not class-discriminative. (For example, Fig. 1b and Fig. 1h are very similar.) In contrast, localization approaches such as CAM and the method proposed by the authors, Gradient-weighted Class Activation Mapping (Grad-CAM), are highly class-discriminative. (The description of 'cat' emphasizes only the 'cat' region and not the 'dog' region, as seen in Fig. 1c and Fig. 1i.) To combine the advantages of both methodologies, the authors show that it is possible to combine pixel-space gradient visualization with Grad-CAM to produce a high-resolution and class-discriminative guided Grad-CAM visualization. As a result, as shown in Figs 1d and 1j, high-resolution visualization of important regions of the image corresponding to any decision of interest can be achieved even when the image contains several possible concepts. When visualizing a 'tiger cat', Guided Grad-CAM not only highlights the cat region but also highlights the stripes on the cat, which are important for predicting a particular species of cat.

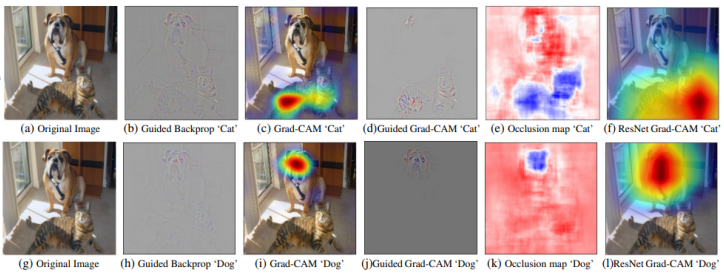

Figure 1: (a) Original image with a cat and a dog. (b-f) Support for the cat category according to various visualizations for VGG-16 and ResNet. (b)Guided Backpropagation[2]: highlights all contributing features. (c, f) Grad-CAM (Ours): localizes class-discriminative regions, (d) Combining (b) and (c) gives Guided Grad-CAM, which gives high-resolution class-discriminative visualizations. Interestingly, the localizations achieved by our Grad-CAM technique, (c) are very similar to results from occlusion sensitivity (e), while being orders of magnitude cheaper to compute. (f, l) are Grad-CAM visualizations for ResNet-18 layer. Note that in (c, f, i, l), red regions corresponds to high score for class, while in (e, k), blue corresponds to evidence for the class. Figure best viewed in color.

Grad-CAM

Numerous previous studies have argued that deeper representations in CNNs capture higher-level visual constructs. Moreover, convolutional layers inherently retain spatial information that is lost in fully-connected layers. Thus, the authors expect the last convolutional layer to strike a compromise between high-level semantics and detailed spatial information. Neurons in these layers seek semantic, class-specific information within the image. Grad-CAM is a method that assigns an importance value to each neuron for a particular decision of interest, using gradient information flowing into the last convolutional layer of the CNN.

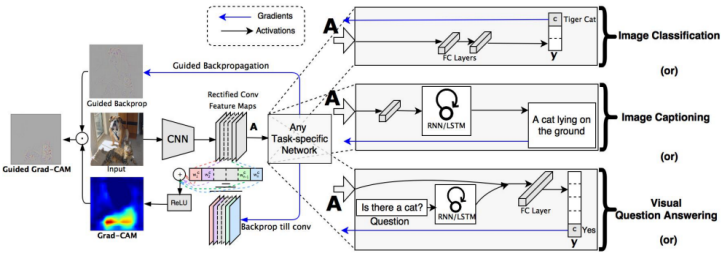

Figure 2: Grad-CAM overview

As shown in Figure 2, in order to obtain the class-discriminative localization map Grad-CAM $L^c_{Grad-CAM} \in R^{u \times v}$ for a certain class c with width u and height v, The authors first calculate the gradient for the score y^c (before the softmax) for class c with respect to the feature map activation A^k of the convolutional layer. So, it can be expressed by $\frac{\partial y^c}{\partial A^k}$.

Global average pooling is applied to gradients for width and height dimension to obtain neuron importance weights α^c_k.

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

where Z is the normalization factor, i and j are the spatial indices for the activation maps, and the summation is performed over all spatial locations. The weight α^c_k captures 'importance' of feature map k for target class c. The

This is an extended abstract. The full paper is available at the this paper

authors perform the weighted combination of forward activation maps and then ReLU the Grad-CAM localization map.

1. Compute the weighted combination of the activation maps using the weights α_k^c:

$$L^c = \sum_k \alpha_k^c A^k \tag{2}$$

2. Apply the ReLU activation function to the class activation map L^c to obtain the Grad-CAM localization map:

$$L^c_{Grad-CAM} = ReLU(L^c) \tag{3}$$

This will result in a coarse heatmap of the same size as the convolutional feature maps. The authors apply ReLU to a linear combination of maps because they are only interested in features that positively influence the class of interest. In other words, it refers to a pixel that has intensity to be increased in order to increase y^c. It is said that if ReLU is not applied, it shows worse performance in localization. y^c does not have to be a class score created by an image classification CNN, and any differentiable activations, including captions or words from answers to questions, can be applied.

Grad-CAM generalizes CAM

In this paper, the authors discuss the connection between Grad-CAM and Class Activation Mapping (CAM), and formally verify that Grad-CAM generalizes CAM to various CNN-based architectures. Assuming that the second layer from the end creates K feature maps, it is $A^k \in R^{u \times v}$, and each element is indexed by i, j. Therefore, A^k_{ij} denotes the activation at location (i, j) of feature map A^k. These feature maps are spatially pooled using Global Average Pooling (GAP) and linearly transformed to produce a score Y^c for each class c.

$$Y^c = \sum_k w_k^c \frac{1}{Z} \sum_i \sum_j A^k_{ij} \tag{4}$$

Let F^k be the global average pooled output.

$$F^k = \frac{1}{Z} \sum_i \sum_j A^k_{ij} \tag{5}$$

CAM calculates the final score as follows:

$$Y^c = \sum_k w_k^c \cdot F^k \tag{6}$$

w_c^k represents the weight of the k-th feature map associated with the c-th class. The gradient for the feature map F^k of score (Y^c) for class c can be calculated as follows:

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A^k_{ij}}}{\frac{\partial F^k}{\partial A^k_{ij}}} \tag{7}$$

If we get the partial derivatives of A^k_{ij} in (5), it is $\frac{\partial F^k}{\partial A^k_{ij}} = \frac{1}{Z}$. Substituting this in (7) gives:

$$\frac{\partial Y^c}{\partial F^k} = \frac{\partial Y^c}{\partial A^k_{ij}} \cdot Z \tag{8}$$

We get $\frac{\partial Y^c}{\partial F^k} = w_k^c$ from (6). thus,

$$w_k^c = Z \cdot \frac{\partial Y^c}{\partial A^k_{ij}} \tag{9}$$

Adding both sides of (9) for all pixels (i, j),

$$\sum_i \sum_j w_k^c = \sum_i \sum_j Z \cdot \frac{\partial Y^c}{\partial A^k_{ij}} \tag{10}$$

Since Z and w_k^c do not depend on (i, j), they can be written as follows.

$$Z w_k^c = Z \sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{ij}} \tag{11}$$

(The left side is the sum of both i and j, so the number of them, Z, remains.) Since Z represents the number of pixels in the feature map, we can divide both sides by Z as follows.

$$w_k^c = \sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{ij}} \tag{12}$$

Except for the proportionality constant ($\frac{1}{Z}$) that normalizes out, the expression w_k^c is identical to the α_k^c used by Grad-CAM.

Thus, Grad-CAM is a strict generalization of CAM.

Guided Grad-CAM

Grad-CAM is class-discriminative and locates relative image regions, but it lacks the ability to emphasize fine-grained details like pixel-space gradient visualization methods such as Guided Backpropagation and Deconvolution. Looking at Figure 1c, Grad-CAM easily locates the cat, but it is unclear why the network predicted this particular instance as a 'tiger cat' from the coarse heatmap. To combine the best of both worlds, the authors fuse Guided Backpropagation and Grad-CAM visualizations through element-wise multiplication. The resulting visualization is high-resolution and class-discriminative. A slight modification of the Grad-CAM allows us to obtain a description that highlights the regions that make the network change its predictions. As a result, removing the concepts that appear in these regions makes the model more confident about its prediction. The authors call this explanation modality counterfactual explanations. Specifically, the authors make the gradient value for feature maps A of the convolutional layer of y^c, which is the score for class c, negative. Therefore, importance weight α_k^c is replaced by the following equation.

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} - \frac{\partial y^c}{\partial A^k_{ij}} \tag{13}$$

As shown previously in (3), the authors take the weighted sum of the forward activation maps A for the weights α_k^c and pass it through ReLU to obtain counterfactual explanations shown in Figure 3.

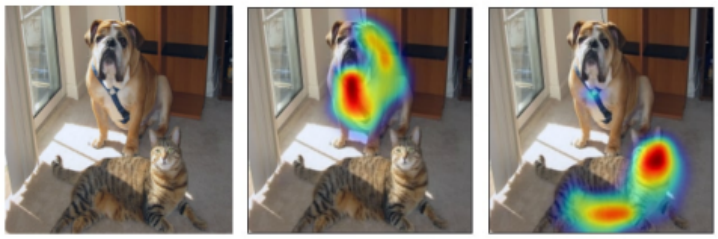

Figure 3: Counterfactual Explanations with Grad-CAM

Experiments

The authors perform extensive experiments to evaluate the effectiveness of Grad-CAM in various tasks, including image classification, image captioning, and visual question answering. In image classification, the authors evaluate Grad-CAM using the ILSVRC dataset and demonstrate that the method successfully highlights discriminative regions corresponding to the predicted class. In image captioning, the authors use the Grad-CAM method on the NeuralTalk model trained on the MSCOCO dataset. In the visual question answering task, the authors use the VQA dataset and a VQA model based on the LSTM and CNN architecture. If you want to see detailed experimental results, I recommend viewing the full paper.

Conclusion

Grad-CAM is a versatile and effective method for generating visual explanations from deep networks using gradient-based localization. It is model-agnostic, applicable to various tasks, and outperforms existing visual explanation methods in terms of localization performance, faithfulness, and robustness. Grad-CAM enables a better understanding of deep learning model behavior, which can be crucial for applications where interpretability and trust in the model's predictions are essential.

[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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[2] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

[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.