Week9 Pre-Report: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

Euijin Hong[1]
[1]Department of Electrical and Electronic Engineering, Yonsei University.

1 Introduction

Gradient-weighted Class Activation Mapping(Grad-CAM) is a technique that produces a 'visual explanation' of CNN predictions, which makes them more transparent and explainable. This method produces a heatmap that highlights the important region of an image when predicting a concept, by making the use of gradients. One advantage of a Grad-CAM is that it can be applied in a wide variety of CNN familes, *without reconstructing or modifying their structure*. In this paper, Grad-CAM is combined with existing fine-grained visualizations for high-resolution class-discriminative visualization, Guided Grad-CAM, and apply it to a variety of computer vision tasks with ResNet-based architectures.

Contributions The main contributions of the introduction of Grad-CAM can be summarized as follows. The Grad-CAM technique, which generates visual explanations for CNN-based networks without requiring architectural changes or re-training. Grad-CAM outperforms baselines in localization and model faithfulness. It is applied to existing top-performing models in classification, captioning, and VQA tasks. Grad-CAM visualizations provide insights into failures of CNNs and expose that common CNN + LSTM models are good at localizing discriminative image regions. Grad-CAM visualizations help in diagnosing failure modes and uncovering biases in datasets. Grad-CAM is also used to obtain textual explanations for model decisions. Human studies show that Guided Grad-CAM explanations help establish trust and successfully discern a 'stronger' network from a 'weaker' one, even when both make identical predictions.

2 Theory and Configuration

Through previous studies, CNN's deeper presentation captures the higher-level visual construction. Moreover, we used the last Convolutional Layer to find a compromise between detailed spatial information and high-level semantic because the spatial information of Convolutional Feature is lost in the FC layer. Neurons of this layer contain information related to the meaning of the class in the image, such as an object. Grad-CAM tried to understand the importance of each neuron related to decision-making using gradient information flowing to the last CNN layer.

As shown in Figure 1, to obtain a class-discriminative localization map Grad-CAM $L^c_{Grad-CAM} \in \mathrm{R}^{u \times v}$, where u and v are width and height for any class c, gradient of the score for class c, y^c (before softmax) is computed, with respect to feature map activation A^k (feature map of kth channel of the final CNN layer) of a convolutional layer, *i.e.* $\frac{\partial y^c}{\partial A^k}$. The flowing back gradients are then global-average-pooled(GAP)[5] over the width and height dimensions to obtain the neuron importance weights α^c_k as follows:

$$\frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A^k} \tag{1}$$

The equation above means a *partial linearization* of a deep network, and shows the *importance* of the kth channel feature map with respect to target class c. Here, we can compute $\frac{\partial y^c}{\partial A^k}$ to backpropagation and the term $\frac{1}{Z}\sum_i\sum_j$ stands for the GAP where Z is a number of pixels in a feature map $(= u \times v)$.

Consequently, the algorithm performs a weighted combination of forward activation maps, by ReLU activation function to obtain:

$$L^c_{Grad-CAM} \in \mathrm{R}^{u \times v} = ReLU(\sum_j \alpha^c_k A^k) \tag{2}$$

The result is a coarse heatmap having the same size as the convolutional feature maps (14×14 for VGG[3] and AlexNet[2]). ReLU is used for confirming the features that have positive influence to class c. Hence, in order to figure out the pixels that have a negative impact on class c, we can use a *counterfactual explanation*, which can be written as $L^c_{Grad-CAM} \in \mathrm{R}^{u \times v} = ReLU(\sum_j \alpha^c_k A^k)$. The result of counterfactual explanation is shown in Figure 2. When ReLU is not applied, the localization map often highlights the region that we are not interested in, which results in decrease of localization performance.

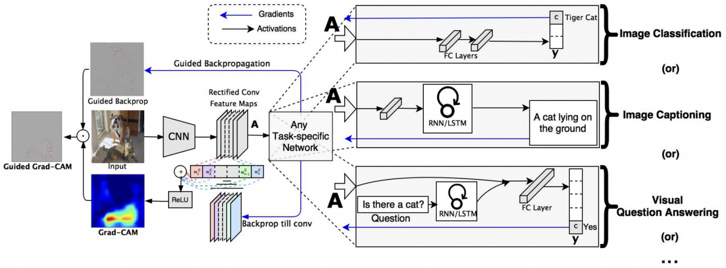


Figure 1: The overview of Grad-CAM. It takes an image and a category of interest as input. It forwards the image through the CNN model and task-specific computations to get a raw score for the category. The gradients are set to zero for all classes except the desired one, and this signal is backpropagated to the rectified convolutional feature maps of interest. These maps are combined to compute the coarse Grad-CAM localization, which shows where the model needs to look to make the decision. The blue heatmap shows this localization. The heatmap is then multiplied with guided backpropagation to create Guided Grad-CAM visualizations, which are high-resolution and concept-specific.

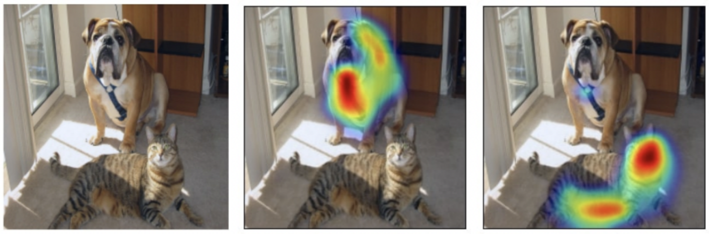

(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Figure 2: Counterfactual explanations with Grad-CAM

2.1 Grad-CAM generalizes CAM

In fact, Grad-CAM generalizes CAM for a wide variety of CNN-based architectures. CAM produces a localization map by a specific kind of architecture for an image classification CNN where global average pooled convolutional feature maps are directly applied into a softmax layer. This can be specified as a following equation.

$$Y^c = \sum_k w^c_k \frac{1}{Z}\sum_i\sum_j A^k_{ij} \tag{3}$$

In this equation, the term w^c_k stands for the kth class feature map weights, the term $\frac{1}{Z}\sum_i\sum_j$ is a GAP, and A^k_{ij} is a feature map. Then we can define F^k, the output of GAP, as follows:

$$F^k = \frac{1}{Z}\sum_i\sum_j A^k_{ij} \tag{4}$$

Then we can plug in the output F^k in equation (3), thus find out that CAM computes the final scores by:

$$Y^c = \sum_k w^c_k \cdot F^k \tag{5}$$

Here, the w^c_k means the weight connecting the kth fetarue map and class c. The gradient of feature map output F^k for class score Y^c as follows:

$$\frac{\partial Y^c}{\partial F^k} = \frac{\frac{\partial Y^c}{\partial A^k_{ij}}}{\frac{\partial F^k}{\partial A^k_{ij}}} \tag{6}$$

Since $\frac{\partial F^k}{\partial A^k_{ij}} = \frac{\partial}{\partial A^k_i j}\frac{1}{Z}\sum_i\sum_j A^k_{ij} = \frac{1}{Z}$, we can obtain the following equation:

This is an extended abstract. The full paper is available at the Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization.

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

Also, since $\frac{\partial Y^c}{\partial F^k} = \frac{\partial}{\partial F^k} \sum_k w^c_k F^k = w^c_k$, we can write the form as:

$$w^c_k = Z \frac{\partial Y^c}{\partial A^k_{ij}} \tag{8}$$

Then, by applying summation with respect to every pixel i and j,

$$\sum_i \sum_j w^c_k = \sum_i \sum_j Z \frac{\partial Y^c}{\partial A^k_{ij}} \tag{9}$$

Here, Z and w^c_k are not related with i and j, and $Z = \sum_i \sum_j 1$, we can obtain w^c_k as:

$$w^c_k = \sum_i \sum_j \frac{\partial Y^c}{\partial A^k_{ij}} \tag{10}$$

Since $\frac{1}{Z}$ is a normalization term, we can find out that $\alpha^c_k = w^c_k$ in Grad-CAM. Thus, the Grad-CAM is a generalized version of CAM.

2.2 Guided Grad-CAM

Grad-CAM has its limit in dealing with detailed factors in pixel-scale of an image. In order to resolve the problem, guided backpropagation visualization visualizes by discarding the negative gradients when it backpropagates through the ReLU layer. By doing so, we can intuitively figure out the pixels detected by neurons. Grad-CAM can easily localize an object, but because of its coarse heatmap, it is hard to know why the network had classified the specific object to a predicted class. The researchers could combine the advantages of guided backpropagation and Grad-CAM by performing element-wise multiplication and proceed the visualization ($L^c_G rad - CAM$ is first upsamples to the input image resolution using bilinear interpolation).

3 Evaluation

3.1 Evaluating Localization Ability of Grad-CAM

Weakly-Supervised Localization is a model that learns without an information of a bounding box. When an image is given, it first performs a class prediction from a network, then produces a Grad-CAM map for the predicted class. It then binarizes each pixel by using 15% of its maximum intensity as a threshold and draws a bounding box around the biggest single segment. Since Grad-CAM does not modify the network structure, it shows high localization performance with less loss of classification performance.

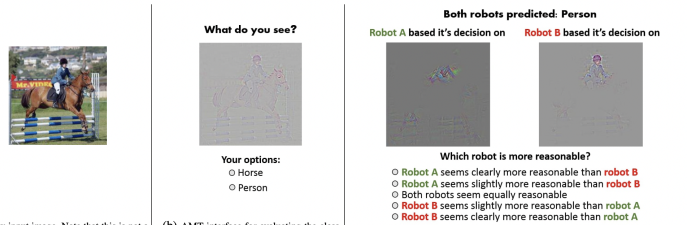
Weakly-Supervised Segmentation is done by performing classification to each pixel. In weakly-supervised segmentation, training is done without information of image-level given, thus can easily obtain the data from image classification dataset. However, the traditional algorithms show sensitive performance depending on the weak localization seed.

Pointing Game Experiment assesses the discriminativeness of localizing methods of an object in an image. It extracts the most activated region in the heatmap and compare with the real object label and obtains $Acc = \frac{\#Hits}{\#Hits+\#Misses}$. Since this method only measures precision, we calculate the localization map for top-5 class prediction in order to consider recall, and considered as 'hit' in cases when segments that are not ground-truth are predicted well. As a result, the performance of Grad-CAM (70.58%) exceeded that of c-MWP[4] (60.30%). Also, in case of Grad-CAM, heatmap was not generated when the class did not appear in an image, which is not the case in c-MWP.

3.2 Evaluating Visualizations

Class Discrimination: In order to evaluate whether Grad-CAM is helpful in discriminating the classes, researchers visualized a single class of a PASCAL VOC 2007 val set, which contains two categories. Four methodologies of 1) Deconvolution 2) Guided Backpropagation 3) Deconvolution Grad-CAM and 4) Guided Grad-CAM were used for VGG-16 and AlexNet CNN. Then they showed the visualizations to Amazon Mechanical Turk (AMT) workers and asked "Which of the two object categories is depicted in the image?". Intuitively, a good description is one that produces discriminative visualizations for the class of interest. And this shows the following result: 1) Deconvolution - 53.33% 2) Guided Backpropagation - 44.44% 3) Deconvolution Grad-CAM - 61.23% and 4) Guided Grad-CAM - 60.37%. The interesting part here is that Guided Backpropagation seems more aesthetically pleasing, but shows lower class discrimination result compared to that of Deconvolution.

Evaluating Trust: To compare the visualization performance of Guided Backpropagation and Guided Grad-CAM, VGG-16 and AlexNet are used, given the fact that the performance of VGG exceeds that of AlexNet. Then 54 AMT workers are inquired to relatively score (+2/+1/0/-1/-2) the reliability of descriptions as in Figure 3. VGG-16 and AlexNet are given a similar probability to produce an output for eliminating a bias. The subjects determined VGG-16 to be more accurate. Although Guided Backpropagation gained score 1 in average (meaning that VGG is relatively little more trustworthy than AlexNet), Grad-CAM got an average score of 1.27 (meaning that VGG is showing relatively more actual trust).


What do you see?
Your options:
○ Horse
○ Person
Both robots predicted: Person
Robot A based it's decision on
Robot B based it's decision on
Which robot is more reasonable?
○ Robot A seems clearly more reasonable than robot B
○ Robot A seems slightly more reasonable than robot B
○ Both robots seem equally reasonable
○ Robot B seems slightly more reasonable than robot A
○ Robot B seems clearly more reasonable than robot A

(a) Raw input image. Note that this is not a part of the tasks (b) and (c)

(b) AMT interface for evaluating the class-discriminative property

(c) AMT interface for evaluating if our visualizations instill trust in an end user

Figure 3: AMT interfaces for evaluating different visualizations for class discrimination (b) and trustworthiness (c). Guided Grad-CAM outperforms baseline approaches (Guided-backprop and Deconvolution) showing that our visualizations are more class-discriminative and help humans place trust in a more accurate classifier.

3.3 Diagnosing Image Calssification CNNS with Grad-CAM

Analyzing Failure Modes for VGG-16: The Guided Grad-CAM was applied for visualizing the prediction reference of VGG-16 classification failures. The result shows that there are justifiable reasons for wrong predictions. One of the advantages of Guided Grad-CAM high resolution and class-discriminativeness, which makes the analysis easier.

Effect of Adversarial Noise on VGG-16: Adversarial attack is a vulnerability that comtemporary deep networks possess. When an unidentifiable noise is introduced in an input image, the model performs a wrong classification in a high probability. The researchers applied Grad-CAM visualization to generated adversarial images where probabilities of existing categories are low and those of non-existing categories are high. As a result, Grad-CAM shows significant robustness against adversarial noise.

Identifying Bias in Dataset: Another useage of Grad-CAM is detecting the bias in the training dataset and diminishing it, since the dataset with bias fails to generalize the real world. This experiment shows that Grad-CAM can help find and reduce bias present in the dataset. This is very important not only because it improves generalization performance, but also because it must show fair and ethical results when making decisions in society.

3.4 Grad-CAM for Image Captioning and VQA

Image Captioning: Grad-CAM can be used to visualize the location for image captioning. Neuraltalk2 using fine-tuned VGG-16 model and LSTM-based language model is used in this study. When a caption is provided, the gradient of the final CNN layer for log probability is calculated and Grad-CAM visualization is then generated. When all images are evenly highlighted is set to baseline 1, Grad-CAM showed 3.27 ± 0.18, Guided Backpropagation showed 2.32 ± 0.08, and Guided Grad-CAM showed 6.38 ± 0.99. Thus we can see that Grad-CAM well localizes the image region of Dense Captioning (DenseCap) [1] model without training by bounding boxes.

Visual Question Answering: General VQA pipeline consists of a CNN for image processing and RNN language model for questioning. Images and questions are fused to predict the answer, typically with a 1000-way classification. Hence, we pick an answer and use its score y^c to compute Grad-CAM visualizations over the image to explain the answer. Although the task is complicated, the explanation is intuitive and informative. Then, in order to assess the Grad-CAM quantitatively, researchers obtained a correlation with the occlusion map. The rank correlation for the occlusion map were 0.42 ± 0.038 for Guided Backpropagation, and 0.60 ± 0.038 for Grad-CAM, which means that Grad-CAM shows higher faithfulness.

[1] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[4] Jianming Zhang, Sarah Adel Bargal, Zhe Lin, Jonathan Brandt, Xiaohui Shen, and Stan Sclaroff. Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10):1084–1102, 2018.

[5] 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.