Lab 7: GradCAM pre-report

CHI YEONG HEO[1],
[1]School of Computer Science and Engineering, Pusan National University, Busan 46241 Republic of Korea

1. Introduction

This report provides a summary of a paper in the field of computer vision: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization [1].

2. Overview of Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) is a method developed to provide visual explanations for the decisions of convolutional neural networks (CNNs). Grad-CAM computes the gradients of a specific target output and backpropagates them to the final convolutional layer to generate coarse heatmaps, which localize the regions of the input image that are most relevant to the model's prediction. This technique is applicable to a wide range of CNN architectures without requiring modifications to the network's structure or additional training. By enhancing the interpretability of model predictions, Grad-CAM contributes to improving the transparency and explainability of computer vision systems.

3. Grad-CAM

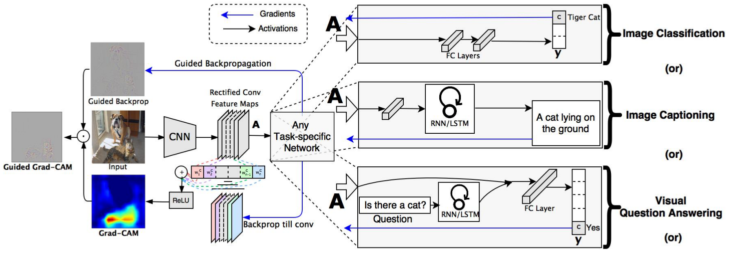

Figure 1: Overview of Grad-CAM: Gradients for a target class are backpropagated to the final convolutional layer, producing a coarse localization map (blue heatmap) that highlights regions relevant to the model's prediction. Combining this map with Guided Backpropagation yields high-resolution, class-specific visualizations.

Grad-CAM generates class-discriminative localization maps by utilizing the gradients of the target class score y^c , with respect to feature map activations A^k of a convolutional layer, $\frac{\partial y^c}{\partial A^k}$. The gradients are globally average-pooled to compute importance weights α_k^c, representing the contribution of each feature map to the target class.

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

A linear combination of the feature maps weighted by α_k^c, followed by ReLU, produces the localization map:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A^k\right).$$

ReLU ensures only positive contributions are considered, resulting in a coarse heatmap highlighting key regions influencing the prediction.

Grad-CAM generalizes CAM: Grad-CAM extends CAM to work with a wider range of CNN architectures, including those used for complex tasks like image captioning and VQA.

Guided Grad-CAM: Combining Grad-CAM with Guided Backpropagation provides high-resolution, class-specific visualizations by integrating global and fine-grained details.

Counterfactual explanations: Modifying Grad-CAM approach allows models to identify regions that, if removed, would increase the model's confidence in its current prediction. This is achieved by altering the importance weights:

$$\alpha_k^c = -\frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}.$$

4. Evaluation

4.1. Localization

Grad-CAM demonstrates strong performance in weakly-supervised localization and segmentation, enabling models to localize important regions without the need for bounding box annotations or pixel-wise labels. As shown in Table 1, Grad-CAM achieves competitive localization error rates while maintaining classification accuracy. Furthermore, Figure 2 highlights its ability to provide effective seeds for segmentation tasks.

		Classification		Localization	
		Top-1	**Top-**5	**Top-**1	**Top-**5
VGG-16	Backprop	30.38	10.89	61.12	51.46
	c-MWP	30.38	10.89	70.92	63.04
	Grad-CAM (ours)	30.38	10.89	**56.51**	46.41
	CAM	33.40	12.20	57.20	**45.14**
AlexNet	c-MWP	44.2	20.8	92.6	89.2
	Grad-CAM (ours)	44.2	20.8	68.3	56.6
GoogleNet	Grad-CAM (ours)	31.9	11.3	60.09	49.34
	CAM	31.9	11.3	60.09	49.34

Table 1: Classification and localization error % on ILSVRC-15 val (lower is better) for VGG-16, AlexNet and GoogleNet. Grad-CAM achieves superior localization errors without compromising on classification performance.

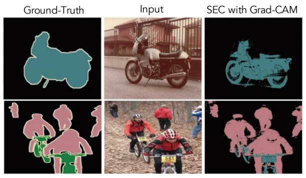

Figure 2: PASCAL VOC 2012 Segmentation results with Grad-CAM as seed for SEC.

4.2. Visualization

Grad-CAM (Grad-CAM) visualizations enhance interpretability by highlighting class-specific features, as demonstrated in human studies where Guided Grad-CAM achieved higher accuracy (61.23%) in identifying object categories compared to Guided Backpropagation (44.44%). Additionally, Grad-CAM visualizations helped users trust better-performing models like VGG-16 over AlexNet, even with identical predictions. By correlating well with occlusion-based methods (rank correlation 0.261), Grad-CAM balances faithfulness to model behavior and interpretability, making it a reliable tool for understanding CNN predictions.

4.3. Diagnosing Image Classification CNNs

Grad-CAM assists in diagnosing errors in image classification models by highlighting regions contributing to incorrect predictions. This insight helps identify potential biases or deficiencies in training data, facilitating model debugging and improvement.

4.4. Textual Explanations with Grad-CAM

Grad-CAM (Grad-CAM) enhances model interpretability by generating visual and textual explanations through neuron importance scores (α_k) in convolutional layers. Using neuron naming techniques, the top-5 positive and negative neurons provide descriptive textual explanations. Despite occasional misclassifications or thresholding issues, Grad-CAM effectively aligns intuitive concepts with model predictions for greater transparency.

4.5. Image Captioning and VQA

Grad-CAM extends its utility to image captioning and Visual Question Answering (VQA) by localizing regions relevant to generated captions or answers. Its class-discriminative heatmaps provide visual explanations that align with semantic content, offering transparency in these complex tasks.

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.