PRE-REPORT: Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization

SongSeungWu

1 ABSTRACT

The paper introduces Gradient-weighted Class Activation Mapping, a technique for generating visual explanations from CNN-based models, enhancing their transparency and explainability. Grad-CAM uses gradients of a target concept flowing into the final convolutional layer to create a localization map that highlights significant regions in the image. This method is applicable to various CNN architectures without requiring modifications or retraining, including models for image classification, captioning, multimodal tasks, and reinforcement learning.
Grad-CAM is combined with fine-grained visualizations to form Guided Grad-CAM, which produces high-resolution class-discriminative visualizations. It is applied to tasks like image classification, image captioning, and visual question answering.

2 INTRODUCTION

The paper addresses the need for interpretability in deep neural networks, particularly CNN-based models, which excel in tasks like classification and object detection but lack transparency. Grad-CAM, a generalized version of Class Activation Mapping, is proposed to create visual explanations without altering model architectures or requiring retraining. Grad-CAM localizes discriminative image regions relevant to model predictions, making it applicable across diverse CNNs, including those for structured outputs and multi-modal inputs.
The technique combines pixel-space gradient visualizations with Grad-CAM to produce Guided Grad-CAM, which is both high-resolution and class-discriminative.
Applications include:

- Diagnosing model failures by identifying biases in datasets.
- Enhancing trust through faithful and insightful visualizations.
- Providing textual explanations for model decisions using neuron importance.
- Conducting human studies to validate Grad-CAM's utility in establishing trust and distinguishing model strengths.

3 GRAD-CAM

The Grad-CAM section introduces a method for generating class-discriminative visual explanations by leveraging gradient information from the last convolutional layer of CNNs. Grad-CAM calculates importance weights for each feature map using gradients of the target class score and creates a coarse localization map. This approach is generalizable to any CNN architecture without modifications, including structured output and multi-modal tasks.

- Generalization of CAM: Grad-CAM extends CAM to handle more complex architectures while preserving model performance and transparency.
- Guided Grad-CAM: Combines pixel-space visualization methods like Guided Backpropagation with Grad-CAM to enhance resolution and class-discrimination.
- Counterfactual Explanations: Modifies Grad-CAM to highlight regions that, if removed, would increase model confidence in its predictions, providing actionable insights.
- Applications: Grad-CAM effectively visualizes important regions for class predictions and is applied in classification, captioning, and VQA tasks, demonstrating its versatility.

4 EVALUATING LOCALIZATION ABILITY OF GRAD-CAM

- Weakly-supervised Localization: Grad-CAM is tested on the ImageNet localization challenge, where it predicts bounding boxes for top-1 and top-5 classes without using bounding box annotations during training. Grad-CAM achieves superior localization accuracy compared to CAM and other methods while maintaining classification performance, as shown with VGG-16, AlexNet, and GoogleNet models.
- Weakly-supervised Segmentation: Grad-CAM is applied to the PASCAL VOC 2012 segmentation task, replacing CAM maps to generate weak localization seeds. This results in an improved Intersection over Union score of 49.6, compared to CAM's 44.6.
- Pointing Game: Grad-CAM is evaluated using the "Pointing Game," which measures the accuracy of visual explanations in localizing target objects. Grad-CAM achieves higher accuracy compared to c-MWP by improving both precision and recall.

5 DIAGNOSING IMAGE CLASSIFICATION CNNS WITH GRAD-CAM

This section demonstrates how Grad-CAM can analyze failure modes, assess robustness to adversarial noise, and identify and reduce biases in datasets using a VGG-16 model pretrained on ImageNet.

- Analyzing Failure Modes: Grad-CAM visualizations of misclassified examples highlight both the predicted and correct classes. These visualizations reveal that seemingly unreasonable predictions often have rational explanations, leveraging Grad-CAM's high resolution and class-discriminative properties.
- Effect of Adversarial Noise: Grad-CAM proves robust to adversarial examples by correctly localizing true categories despite adversarial perturbations causing high-confidence misclassifications.
- Identifying Dataset Bias: Grad-CAM identified gender stereotypes in a doctor vs nurse classification model, showing that it relied on facial features instead of task-relevant features. By adding balanced gender representations to the training set, the model's generalization improved , and it began focusing on relevant features. This demonstrates Grad-CAM's role in detecting and mitigating dataset biases for ethical AI.

6 CONCLUSION

This work introduced Grad-CAM, a class-discriminative localization technique for generating visual explanations and enhancing the transparency of CNN-based models. By combining Grad-CAM with high-resolution visualization methods, the study developed Guided Grad-CAM, which achieves superior interpretability and faithfulness compared to prior methods. Human studies confirmed its ability to improve class discrimination, expose classifier trustworthiness, and identify dataset biases. Grad-CAM also enables identification of important neurons and provides textual explanations for model decisions. Its versatility was demonstrated across tasks like image classification, captioning, and visual question answering. The authors emphasize that trustworthy AI must reason about its actions for human users, suggesting future applications in reinforcement learning, NLP, and video processing.