Enhancing Explainability in CNNs through Grad-CAM: Practical Applications and Project Implementation

Junghoon Lee
Aiffel Research 8th
Suwon, South Korea
coronarita1991@gmail.com

Abstract

This paper explores the enhancement of convolutional neural network (CNN) explainability using Gradient-weighted Class Activation Mapping (Grad-CAM). The study is divided into two main parts: practical application and project implementation. The first part details the theoretical foundation and practical execution of Grad-CAM on pre-trained models like ResNet. The second part focuses on a project that trains a pre-trained CNN model, applying CAM and Grad-CAM to visualize and interpret model predictions. Experimental results demonstrate how Grad-CAM shows model transparency, aiding in better understanding and trust in AI decisions.

1 Introduction

The field of deep learning has witnessed remarkable progress, particularly with the advent of Convolutional Neural Networks (CNNs) which have significantly advanced the capabilities of computer vision tasks. Despite their success, CNNs are often criticized for being "black boxes" due to their lack of interpretability, making it challenging to understand the rationale behind their predictions. This limitation hinders the deployment of CNNs in critical applications where trust and transparency are paramount.

Gradient-weighted Class Activation Mapping (Grad-CAM) has emerged as a powerful technique to address this challenge. Grad-CAM provides visual explanations for CNN decisions by highlighting the regions of an input image that are most influential for the model's predictions. This interpretability tool not only helps in understanding and debugging models but also enhances user trust by providing insights into the model's decision-making process.

Contribution of this paper This paper is divided into two sections. The first section focuses on the development history of Grad-CAM, including the theoretical foundation and implementation on pre-trained models like ResNet. The second section presents a project-based approach, where a pre-trained CNN model is trained and Grad-CAM is applied to visualize and interpret its predictions. Through experimental results, this study demonstrates the comparision of CAM and Grad-CAM in improving the transparency and explainability of CNN models, ultimately contributing to more valuable and interpretable AI systems.

2 Background

Convolutional Neural Networks (CNNs) have become the cornerstone of modern computer vision, excelling in tasks such as image classification, object detection, and segmentation. Their hierarchical

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

structure, inspired by the human visual system, allows them to learn intricate patterns from raw pixel data through multiple layers of convolutional operations. Despite their success, CNNs are often criticized for their opacity, as the complex transformations within the network obscure the rationale behind their predictions. This lack of transparency poses significant challenges, particularly in high-stakes applications like healthcare, autonomous driving, and security, where understanding the decision-making process is crucial.

To address this issue, various techniques have been developed to make CNNs more interpretable. Among these, Gradient-weighted Class Activation Mapping (Grad-CAM) has gained prominence. Grad-CAM generates visual explanations by leveraging the gradients of target concepts flowing into the final convolutional layer. This approach highlights the important regions in an input image that are most influential for the model's predictions, providing valuable insights into the model's inner workings.

3 Related Work

The concept of visualizing CNN activations was first introduced through Class Activation Mapping (CAM), which used global average pooling layers to create heatmaps indicating important image regions. However, CAM was limited to specific network architectures. Grad-CAM extended this idea by using gradient information, making it applicable to a wider range of models without architectural modifications.

Several advancements have been made in the field of model interpretability. LIME (Local Interpretable Model-agnostic Explanations) offers model-agnostic explanations by perturbing the input and observing changes in predictions. SHAP (SHapley Additive exPlanations) values provide a game-theoretic approach to interpreting model outputs by attributing the prediction to individual input features.

This paper builds upon these foundational works by applying conventional CAM and Grad-CAM to pre-trained and CNN models. Through detailed experiments and visual analyses, we aim to demonstrate the practical utility of Grad-CAM in enhancing model transparency and aiding in the interpretability of CNN decisions.

4 Method

4.1 Dataset Preparation

4.1.1 Image Dataset

- The dataset used is the **Stanford Dogs** dataset, loaded using the *TensorFlow Datasets (TFDS)* library.
- This dataset consists of images from **120 different dog breeds**, with approximately 100 to 150 images per breed.
- The dataset is pre-split into training, validation, and test sets, and appropriate preprocessing (resizing, normalization) was performed on the images before use.
- Data augmentation technique was not used.

4.2 Model Architecture

4.2.1 Pre-trained Models

- Fix the shape of input layers as (224, 224, 3).
- Utilize pre-trained CNN models such as ResNet50 as a backbone.[3]
- Modify the final layers to match the target classification tasks.

4.2.2 Grad-CAM

- Design a custom CNN architecture suitable for the specific dataset.
- Implement layers including convolutional, pooling, and fully connected layers.

Figure 1: Stanford Dogs dataset examples

4.3 Training Procedure

4.3.1 Hyperparameters

- Set learning rate, batch size, number of epochs, and optimizer (0.01, 16, 2, SGD).
- Use a validation set to avoid overfitting.

4.3.2 Training Process

- Train the models using the prepared dataset.
- Monitor training and validation loss to use model with CAM and Grad-CAM.

4.4 Implementation

4.4.1 CAM

- Traditional CAM was implemented to provide a baseline for comparison with Grad-CAM.
- CAM uses global average pooling layers to create class-specific activation maps.
- This method highlights important regions in an image by weighting the feature maps based on the output of the global average pooling layer.
- The implementation involved modifying the final layers of the CNN to include global average pooling, followed by a dense layer corresponding to the number of classes.
- While effective, CAM is limited to specific network architectures that include global average pooling

4.4.2 Gradient-weighted Class Activation Mapping (Grad-CAM)

- Grad-CAM was implemented to generate visual explanations applicable to a broader range of CNN architectures.

Figure 2: Original Image as a sample

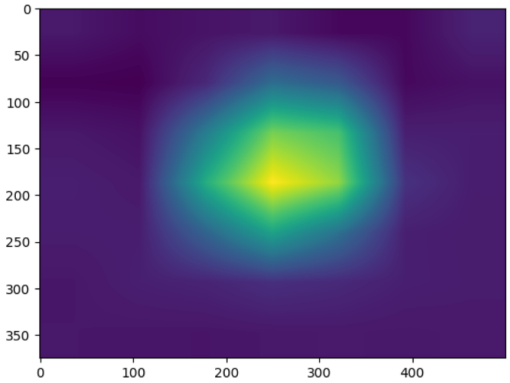

Figure 3: Heatmap Result of CAM as a last layer

- Grad-CAM calculates the gradients of the target class with respect to the feature maps of the last convolutional layer.
- These gradients are used to weight the feature maps, producing a heatmap that highlights important regions of the input image.
- The heatmap is then superimposed on the original image to create a visual explanation.
- This approach is more flexible and can be applied to models without global average pooling

4.4.3 Heatmap Generation

4.4.4 Overlay Heatmap

- Superimpose the heatmap onto the original image for visualization.

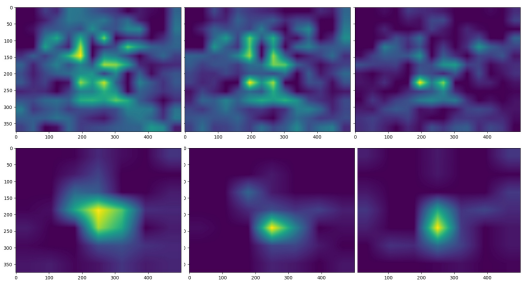

Figure 4: Heatmap Result of Grad-CAM as last 6 layers

4.5 Evaluation

- Use a sample image to compare two methods (as shown in Figure 2).
- Weight the feature maps by the computed gradients.
- Generate a heatmap highlighting important regions in the input image.
- Detail results are shown in Figure 3 and 4.

4.5.1 Performance Metrics

- Evaluate model with IoU(Intersection of Union).

4.5.2 Explainability Assessment

- Use Grad-CAM to generate visual explanations for model predictions.
- Analyze the heatmaps to assess the interpretability and reliability of the models.

4.6 Experimental Setup

4.6.1 Software and Tools

- Use Python, TensorFlow with Keras for model development and training.
- Employ visualization libraries (Matplotlib and openCV) for displaying heatmaps.

4.6.2 Hardware

- Utilize Nvidia T4 GPU to accelerate training and gradient computations.

4.6.3 Discussion

- Interpret the visual explanations provided by CAM and Grad-CAM.
- Discuss the strengths and limitations of the models in terms of explainability.

5 Results

5.1 Explainability Assessment

5.1.1 Visualizations

- Generate heatmaps for correctly classified images.
- Superimpose the heatmaps on the original images to visualize the regions of interest.

5.1.2 Analysis of Results

- Interpret the heatmaps to identify which regions of the images are most influential in the model's predictions.
- Discuss how the visualizations provide insights into the decision-making process of the models.

5.2 Comparative Analysis

5.2.1 Quantitative and Qualitative Insights

- Provide quantitative metrics (e.g., localization accuracy) to measure the quality of Grad-CAM explanations.
- Both results are overlaid using last layer of a pre-trained model (as shown in Figure 5 and 6).
- Include qualitative assessments based on human evaluation of the heatmaps.
- IoU results of both methods are visualized in Figure 7 and 8.

Figure 5: CAM Heatmap result overlaid with a sample image

Figure 6: Grad-CAM Heatmap result overlaid with a sample image

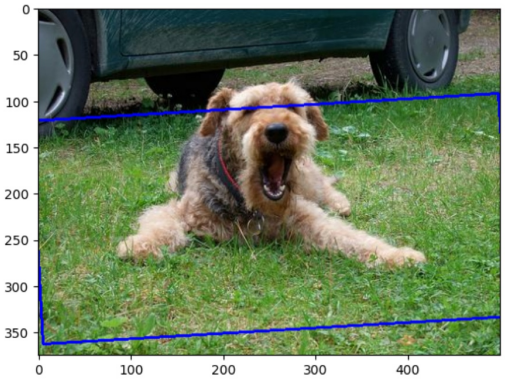

Figure 7: IoU result with Grad-CAM

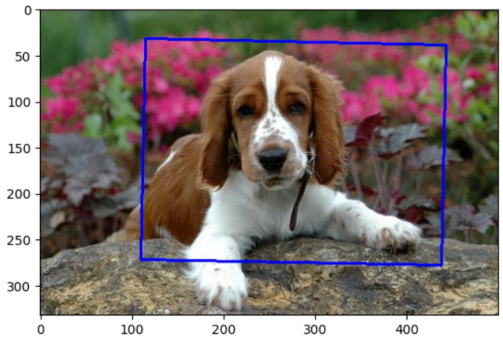

Figure 8: IoU result with CAM (other image)

6 Discussion

6.1 Interpretation of Results

The results of this study indicate that both traditional CAM and Grad-CAM provide valuable insights into the decision-making processes of CNN models. However, Grad-CAM offers several advantages due to its flexibility and applicability across different network architectures.

6.2 Advantages of Grad-CAM

One significant advantage of Grad-CAM is its ability to generate visual explanations for deeper layers of the network. This capability allows for a more detailed understanding of how complex features are learned and utilized by the model. By highlighting the regions of the input image that contribute most to the final prediction, Grad-CAM helps in interpreting the influence of deeper convolutional layers, which are often responsible for capturing high-level semantic information.

6.3 Impact on Model Trust and User Confidence

The visual explanations provided by Grad-CAM can significantly enhance model trust and user confidence. By making the model's decisions more transparent, users can better understand and trust the predictions made by the model. This is particularly important in critical applications where the stakes are high, such as healthcare and autonomous driving.

6.4 Limitations and IoU Performance

Despite its advantages, Grad-CAM's IoU (Intersection over Union) results were found to be suboptimal in our experiments. This indicates that while Grad-CAM is effective at highlighting relevant regions, its precision in accurately localizing these regions needs improvement. The suboptimal IoU performance suggests that Grad-CAM may struggle to produce fine-grained visualizations, particularly in complex images with multiple relevant areas.

6.5 Comparison with Traditional CAM

Traditional CAM, although limited to specific architectures with global average pooling, provided a solid baseline for comparison. It effectively highlighted important regions for class predictions but lacked the flexibility of Grad-CAM. Grad-CAM's ability to work with a wider range of architectures and provide deeper insights into the model's inner workings makes it a more versatile tool for model interpretability.

7 Conclusion

This study aimed to enhance the explainability of CNN models using Grad-CAM, alongside a comparison with traditional CAM. While both methods provided valuable insights into model decisions, several key findings emerged:

- Traditional CAM, though limited to specific architectures with global average pooling, effectively highlighted important image regions for class predictions.
- Grad-CAM demonstrated flexibility across various CNN architectures by utilizing gradient information to produce visual explanations.
- Despite its broader applicability, Grad-CAM's IoU (Intersection over Union) results were suboptimal in our experiments, indicating potential challenges in accurately localizing relevant regions.
- These findings suggest that while Grad-CAM is a powerful tool for model interpretability, further improvements and refinements are needed to enhance its localization accuracy and overall effectiveness.

Future work should focus on addressing these limitations by exploring advanced techniques such as Grad-CAM++ and integrating complementary methods to provide more robust and precise visual explanations. Improving the accuracy of these interpretability tools is crucial for their application in critical domains where understanding model decisions is paramount.

8 Acknowledgment

We would like to express our gratitude to the following individuals and sources for their invaluable contributions to this research:

- **AIFFEL Research Study Group**, for providing a collaborative platform and resources that facilitated this study.
- **ChatGPT4-Omni**, for granting access to make draft of this paper with LaTeX.

9 References

[1] K. Simonyan. & A. Vedaldi. & A. Zisserman (2013) "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps"., *arXiv:1312.6034*

[2] M. Lin. & Q. Chen & S. Yan, "Network In Network," *arXiv:1312.4400*, 2013.

[3] K. He. & X. Zhang. & S. Ren. & and J. Sun, "Deep Residual Learning for Image Recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[4] C. Szegedy *et al.*, "Going Deeper with Convolutions," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[5] R. R. Selvaraju *et al.*, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization," *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[6] B Zhou. & A Khosla. & A Lapedriza. & A Oliva. & A Torralba, "Learning Deep Features for Discriminative Localization," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.