Project Proposal

Ostbayerische
Technische Hochschule
Amberg-Weiden

"Explainable AI through Visualization of Layer-Specific Embeddings and Class Activation Mapping (CAM) in PyTorch "

Submitted by

Kashif Riyaz, Nitesh Morem

Supervised by

Prof. Dr.-Ing. Christian Bergler

Ostbayerische Technische Hochschule Amberg-Weiden
Department of Electrical Engineering, Media and Computer Science

June 29, 2024

Abstract

In the technological era of deep learning, neural network architecture has become incredibly well-known for its ability to look for patterns in data and use those patterns. However, the way to discover these patterns is still unknown to most people. Many people are still unable to comprehend the mechanisms underlying these discoveries. Our project aims to shed light into the decision-making process of the pretrained models showing the transparency and interpretability of AI systems using class activation mapping (CAM). CAM highlights important regions in input data for specific class predictions, offering visual explanations for the model's decisions.

We will start with implementation of Grad-CAM on simple CNN models and visualize layer specific embeddings to analyze the model's behavior and outcome, subsequently the framework will then be modified to fit and further can be extended with pretrained models and identify valuable insights in data that influence the model's predictions also other class activation mappings such as Grad-CAM++, Eigen-CAM and XGrad-CAM will be implemented,enabling more insights into model's decision-making procedure.

1 Introduction and Motivation

The deep neural networks have revolutionized various domains with their remarkable performance. However, the nature of these neural networks, "black-box" often raises concerns about transparency and interpret-ability, hindering trust and broader adoption in critical applications.In order to understand and address this, explainable AI has become crucial, aiming to unveil the inner workings of these complex models.

This project primarily investigates the explain-ability of deep neural networks with two approaches: layer-wise embedding visualization and class activation mapping. The former gives insight into the learned representations at each layer and, hence, specific information processed by the model. On the other hand, CAM highlights important regions in input data for specific predictions of classes, thus giving a visual explanation for the model's decisions.

- **Background Information:** The ability of deep learning models to extract the features from the raw data has made significant changes in computer vision. Two important papers "Learning Deep Features for Discriminative Localization" and "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization" have played important role in enhancing the transparency and interpretability of CNNs through innovative techniques, Class activation mapping (CAM) and Gradient-Weighted class activation mapping (Grad-CAM).

- **Research Gap Identification:** Current research does not fully grasp the literature on how complex Convolutional Neural Networks (CNNs) make decisions, especially when used with pretrained models. Because lack of understanding makes it difficult to trust and use these models in real-world situations. Furthermore, there is also a lack of standardized methods and evaluations for interpreting these models, which adds to the challenge.These issues have been important because, up until now, AI technologies have had very few users as people are Concerned about their reliability and possible bias. We hope that an easily understandable AI will make these systems more trustworthy. We are solving real-world problems and developing a framework to address these gaps in knowledge.

- **Research Objectives or Hypotheses:** Our main research objectives include the following: execute CAM on pre-trained CNN models with a view to managing visual explanations for an estimation; evaluate the effectiveness of CAM across a wide variety of image data; set up a standard framework to evaluate model interpretability; know how CAM may influence user trust in AI reliability; and finally, solve the real concerns in the world pertaining to bias and transparency of AI. These objectives are meant to improve the interpretability and trustworthiness of models, thereby filling the current gaps in literature.

- **Significance of the Study:** It will be very important research in improving the theoretical understanding of AI and also its practical applications.

Our approach further gives practical insights through complex CNN decisions with visual explanations. Contributing valuable insights for future studies in Explainable AI by our standardized framework of model interpretability evaluation, our research informs and inspires further research and applications. In a nutshell, our research will try to fill in the gaps which exist and raise AI trust to make a strong impact that lasts in the field of AI.

2 Related Work

A significant breakthrough in bridging this gap came with the introduction of Class Activation Mapping (CAM) by Zhou et al. (2016) [1] Zhou et al. introduced CAM for CNNs with global average pooling. This enabled the pre-trained CNN to perform object localization. The class activation mapping allowed them to predict the score of the given image and highlight the data features detected by CNN.

Expanding on this Selvaraju et al. (2017) [2] introduced Grad-CAM to make CNN models more transparent by producing a visual explanation. They also combined Grad-Cam with existing fine-grained visualization to create class discriminative visualization called Guided Grad-CAM and applied it to image classification, and image captioning including ResNet-Based architecture further improving deep neural network interpretability.

3 Data Material and Preprocessing

- **Data Collection:** We will collect the data needed for our study, which includes models pre-trained and datasets used for testing. The models could also be obtained via PyTorch using the torchvision.models library. Our study uses some of the pre-trained models on the ImageNet dataset. The testing data will be shared in terms of online URLs for ease of use in all environments.

- **Data Description:** The data used will all be images, usually in JPEG or PNG format. How the data is going to be pre-processed is described based on specific model requirements to maintain the consistency of input data.

- **Data Preprocessing Steps:** The raw data for images will initially be the raw pixel values. To feed this to our models, we will use PyTorch to first convert these pixel values into tensors of model-specific dimensions. We then feed these processed tensors to the pre-trained models.

- **Software and Tools:** The whole project will be carried out in Jupyter Notebook using the Python environment. It will make use of some of the core libraries that include PyTorch, torchvision, matplotlib, numpy, pandas, and most critically, the Jacob gill GitHub repository "pytorch-gradcam".

4 Methodology

We set out to create a framework that will help us to understand how pre-trained AI models make decisions. We will focus on CNNs and use CAM to visualize parts of an image which are important for the model's predictions.Our work will be carried out in three parts:

- **Creating a simple framework:** Firstly, we will create a basic framework. It will take a pre-trained model and some images as input.Then the pre-trained models can be imported from Python libraries while the images can be downloaded from anywhere.The framework should be able to identify regions and key features from the data that the model is making some predictions on.
- **Improving the Framework:**Further improvements will be made to the framework in order to enhance its performance. It is expected to work on many other pre-trained models available for from a library which is "torchvision.models".
- **Testing the Framework:** At last, we will test the framework and run it on different pre-trained models and multiple image sets to check that it has done a good job of telling us what is of importance in the input images.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[2] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.