A Survey of Explainable AI (XAI) Methods for Convolutional Neural Networks

Antonio Fernando Silva e Cruz Filho [1] João Gabriel Andrade de Araujo Josephik[1] Prof. Dr. Nina S. T. Hirata

[1]Institute of Mathematics and Statistics

Introduction

As artificial intelligence (AI) becomes part of critical fields like healthcare, finance, and autonomous vehicles, it's important to understand how these systems make decisions. This is where Explainable AI (XAI) comes in. XAI helps make AI models, which are often complex, easier to understand and interpret. This ensures that AI systems are trusted and used responsibly.

Neural networks are powerful tools for tasks like recognizing images or making predictions. However, they are often seen as "black boxes" because it's hard to explain how they reach their decisions. This lack of clarity can be a problem in areas where understanding the reason behind a decision is as important as the result itself.

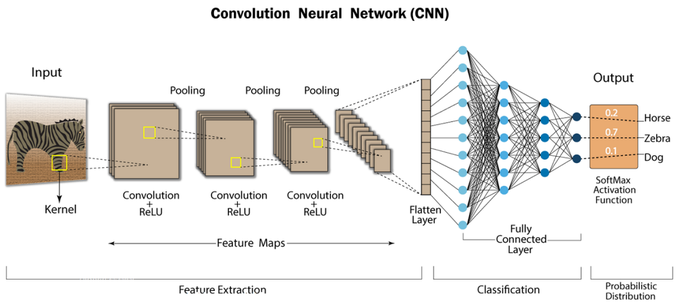

Figure 1. Convolutional Neural Network Architecture

Convolutional Neural Networks (CNNs) are neural networks designed for image and spatial data. They process parts of an image (like edges or textures) and combine these features to make decisions, making them ideal for tasks like face or object recognition. However, CNNs are also complex, and it can be hard to tell what they are focusing on and why. XAI tools can help with this by showing how the CNN works step by step.

We will explore methods such as Feature Visualization, Saliency Maps, and LIME focused on explaining black-box image models like CNNs, aiming to create more robust, reliable, and less biased networks.

Feature Visualization

CNNs can learn abstract features and concepts from images. One can use techniques such as Feature Visualization to visualize the learned features by maximizing a network's neuron (or a set of neurons) value. This technique, called *Activation Maximization*, can be modeled by the formula below, using the *Gradient Ascent* method:

$$x_{t+1} = x_t + \mu \frac{\partial a(\theta, x_t)}{\partial x_t}$$

Where x_t represents an image at iteration t, μ represents a tunable hyperparameter and the function a represents the forward pass of a unit in a Neural Network with parameters θ.

By defining x_0 as a specific image or just random noise, one can create *dreamy-like* [2] images representing that will maximize a certain set of neurons of a Network.

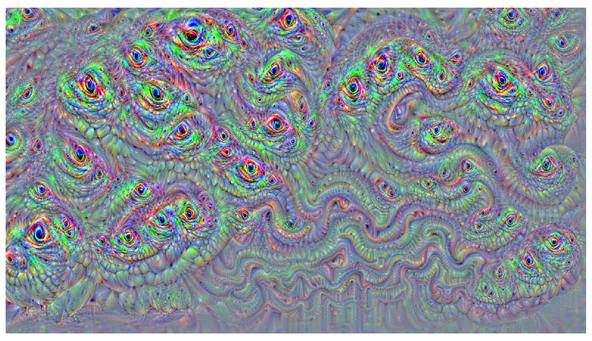

Figure 2. Feature Visualization of CNN VGG16 using a random image

Figure 3. Feature Visualization of CNN VGG16 using a photo of IME-USP

By close inspection in both images, one can notice that certain characteristic features like animal's eyes and dog's faces emerge, showing that the network learned those representations and certain layers are maximized by the presence of those features in images.

Saliency Maps

GRADCAM OMG!!!!

A highlighted block

This block catches your eye, so **important stuff** should probably go here.

Curabitur eu libero vehicula, cursus est fringilla, luctus est. Morbi consectetur mauris quam, at finibus elit auctor ac. Aliquam erat volutpat. Aenean at nisl ut ex ullamcorper eleifend et eu augue. Aenean quis velit tristique odio convallis ultrices a ac odio.

LIME in Images

LIME (Local interpretable model-agnostic explanations) is a technique to explain a complex model's decisions by creating a simpler explainable model based on the complex model. A LIME model can be defined by the expression bellow:

$$g^*(x) = \arg\min_g L(f, g, \pi_x) + \Omega(g)$$

Where L is a loss function to compare the simpler model performance with the complex model performance in the region close to x, defined by π_x. Ω is a function that returns a *complexity metric* for a simple model g.

LIME can be used in images to create masks that represent areas that were more important for a model's decision.

Figure 4. Original Image

Figure 5. Masked image using LIME

Experiments

Et rutrum ex euismod vel. Pellentesque ultricies, velit in fermentum vestibulum, lectus nisi pretium nibh, sit amet aliquam lectus augue vel velit. Suspendisse rhoncus massa porttitor augue feugiat molestie. Sed molestie ut orci nec malesuada. Sed ultricies feugiat est fringilla posuere.

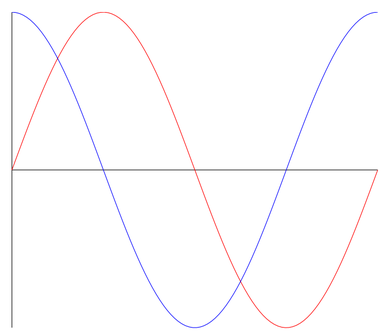

Figure 6. Another figure caption.

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aliquam vel dapibus erat. Morbi quis leo congue, lobortis augue bibendum, malesuada neque. Duis ullamcorper quis orci sed consequat. Nam pellentesque ullamcorper tempor. Duis eget nulla blandit, vulputate orci vitae, ullamcorper ligula. Mauris a urna ac massa dignissim scelerisque sed et augue. Donec eget urna vitae neque elementum pellentesque et eget enim. Praesent a fermentum nibh. Nullam eu nibh neque.

Conclusion

Nulla eget sem quam. Ut aliquam volutpat nisi vestibulum convallis. Nunc a lectus et eros facilisis hendrerit eu non urna. Interdum et malesuada fames ac ante *ipsum primis* in faucibus. Etiam sit amet velit eget sem euismod tristique. Praesent enim erat, porta vel mattis sed, pharetra sed ipsum. Morbi commodo condimentum massa, *tempus venenatis* massa hendrerit quis. Maecenas sed porta est. Praesent mollis interdum lectus, sit amet sollicitudin risus tincidunt non.

Etiam sit amet tempus lorem, aliquet condimentum velit. Donec et nibh consequat, sagittis ex eget, dictum orci. Etiam quis semper ante. Ut eu mauris purus. Proin nec consectetur ligula. Mauris pretium molestie ullamcorper. Integer nisi neque, aliquet et odio non, sagittis porta justo.

- **Sed consequat** id ante vel efficitur. Praesent congue massa sed est scelerisque, elementum mollis augue iaculis.
  - In sed est finibus, vulputate nunc gravida, pulvinar lorem. In maximus nunc dolor, sed auctor eros porttitor quis.
  - Fusce ornare dignissim nisi. Nam sit amet risus vel lacus tempor tincidunt eu a arcu.
  - Donec rhoncus vestibulum erat, quis aliquam leo gravida egestas.
- **Sed luctus, elit sit amet** dictum maximus, diam dolor faucibus purus, sed lobortis justo erat id turpis.
- **Pellentesque facilisis dolor in leo** bibendum congue. Maecenas congue finibus justo, vitae eleifend urna facilisis at.

References

[1] C. Molnar. Interpretable machine learning: A guide for making black box models explainable. `https://christophm.github.io/interpretable-ml-book/`. Accessed: 2024-11-26.

[2] A. Mordvintsev, C. Olah, and M. Tyka. Inceptionism: Going deeper into neural networks. `https://web.archive.org/web/20150703064823/http://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html`. Accessed: 2024-11-26.

[3] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier, 2016. URL `https://arxiv.org/abs/1602.04938`.