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.

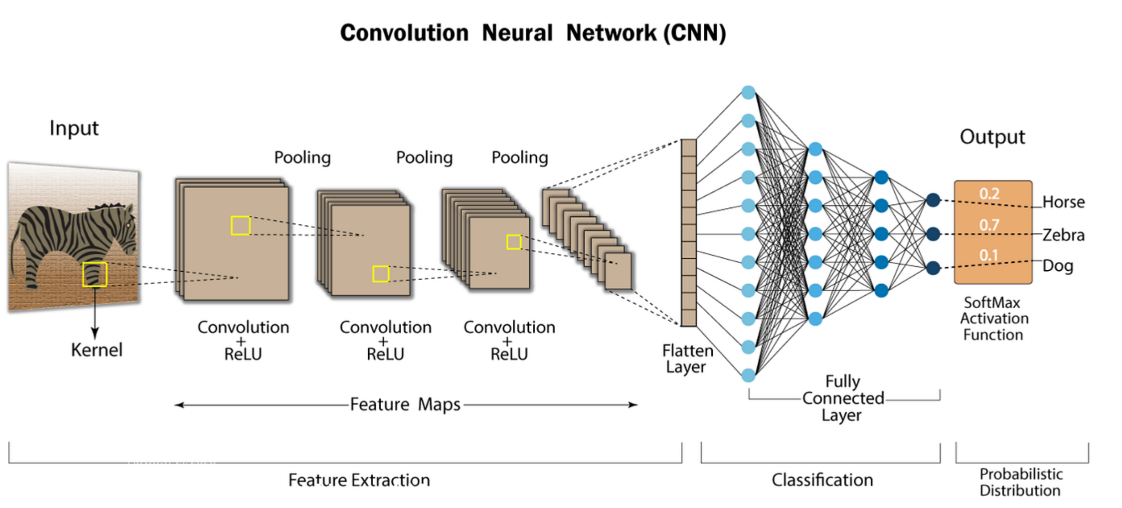

Figure 1. Convolutional Neural Network Architecture

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.

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

Vivamus congue volutpat elit non semper. Praesent molestie nec erat ac interdum suscipit erat. **Phasellus mauris felis, molestie ac pharetra quis**, tempus nec ante. Don ante vel purus mollis fermentum. Sed felis mi, pharetra eget nibh a, feugiat eleifend do mollis condimentum purus quis sodales. Nullam eu felis eu nulla eleifend bibendu lorem. Vivamus felis velit, volutpat ut facilisis ac, commodo in metus.

1. **Morbi mauris purus**, egestas at vehicula et, convallis accumsan orci. Orci varius penatibus et magnis dis parturient montes, nascetur ridiculus mus.
2. **Cras vehicula blandit urna ut maximus**. Aliquam blandit nec massa ac sollicitudi Curabitur cursus, metus nec imperdiet bibendum, velit lectus faucibus dolor, qui metus mauris gravida turpis.
3. **Vestibulum et massa diam**. Phasellus fermentum augue non nulla accumsan, no rhoncus lectus condimentum.

Experiments

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

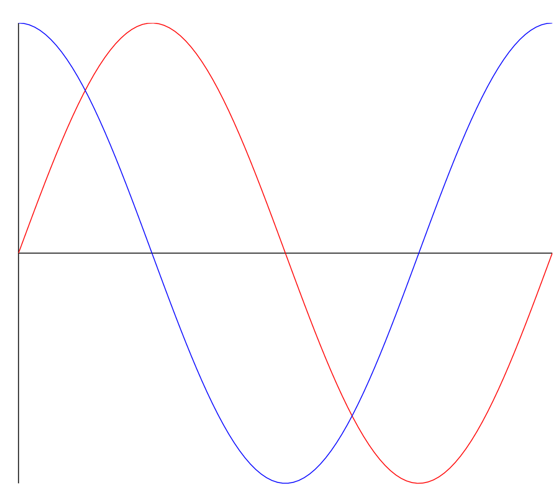

Figure 2. Another figure caption.

Lorem ipsum dolor sit amet, consectetur adip Aliquam vel dapibus erat. Morbi quis leo cong tis augue bibendum, malesuada neque. Duis per quis orci sed consequat. Nam pellentesq corper tempor. Duis eget nulla blandit, vulputa tae, ullamcorper ligula. Mauris a urna ac massa scelerisque sed et augue. Donec eget urna vi elementum pellentesque et eget enim. Praes mentum nibh. Nullam eu nibh neque.

Conclusion

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

Etiam sit amet tempus lorem, aliquet condimentum velit. Donec et nibh consequat, s eget, dictum orci. Etiam quis semper ante. Ut eu mauris purus. Proin nec consecte Mauris pretium molestie ullamcorper. Integer nisi neque, aliquet et odio non, sagi 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 e 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 erat id turpis.
- **Pellentesque facilisis dolor in leo** bibendum congue. Maecenas congue finibus j vitae eleifend urna facilisis at.

References

[1] C. Molnar. Interpretable machine learning: A guide for making black box models able. https://christophm.github.io/interpretable-ml-book/. Accessed: 2 26.

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

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

[4] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra cam: Visual explanations from deep networks via gradient-based localization. *tional Journal of Computer Vision*, 128(2):336–359, Oct. 2019. ISSN 1573-14 10.1007/s11263-019-01228-7. URL http://dx.doi.org/10.1007/s11263-019-012