Demystifying Black Boxes: Analysis of Neural Network Transparency Research

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Parker Hutchinson
Fall 2023

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Lu Feng, Department of Computer Science

Demystifying Black Boxes: Analysis of Neural Network Transparency Research

CS4991 Capstone Report, 2023

Parker Hutchinson
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
pch6am@virginia.edu

ABSTRACT
Neural network usage has expanded greatly in recent years, with OpenAI's ChatGPT and its successors taking the world by storm as notable examples. However, many of these networks are hard to understand. When faced with the question of why a network produced the output it did, researchers may be hard-pressed to produce an answer other than a reference to the training data.

Despite this, many advancements have been made in clarifying the 'decision-making' process of neural networks. Such advancements will be crucial in the modern world, as reliance on such networks increases and more and more demands are placed on their outputs. Such progressions in demystifying the 'black-box' calculation process of neural networks will be important in understanding what guarantees can be made when deploying them.

1. INTRODUCTION

Neural networks are a revolutionary new form of statistical modeling. The first trainable neural network was Frank Rosenblatt's Perceptron, a single-layer model with adjustable weights and thresholds demonstrated in 1957 (Kay 2001). Since then, neural networks have undergone many advancements and iterations. By the 1980s, researchers had deduced algorithms for training networks with more than one layer, and in 1989, a new form of neural network specialized for image recognition, a convolutional neural network, was utilized to recognize hand-written characters. Already, however, "the networks' strategies were indeceipherable" to the researchers who had molded them.

In recent times, researchers have continued to make strides, with Microsoft producing a generative language model with 17B parameters and OpenAI releasing an AI system capable of producing images from text prompts (Karjian 2023). Although researchers have found a plethora of applications for neural networks and ways to improve their training algorithms, the work to understand the strategies used by the actual models themselves has not received as much publicity.

2. RELATED WORKS

Meta AI researchers Touvron, et al. (2023) discuss how they are able to generate neural networks that have state-of-the-art performance while simultaneously utilizing fewer parameters. Specifically, they note that their model with 13B parameters outperforms GPT-3, which has 175B parameters, on most benchmarks. The findings challenge the widespread assumption that "more parameters will lead to better performance", and the researchers verified these findings with standard benchmarks as well as some to analyze the "biases and toxicity" present in the final models. The researchers hope that

their findings will "democratize" the machine learning process, and found all their results "without resorting to proprietary and inaccessible datasets" (Touvron et al., 2023). The open-source approach and amazing results achieved by these researchers is encouraging for future developments and attempts to make machine learning both more accessible and more understandable. The fact that these researchers were able to produce such results while also using fewer parameters in their models is likewise encouraging, because models with fewer parameters are typically easier to understand and analyze than ones with more parameters and complexity.

Researchers Salahuddin, et al. (2021) analyze methods for making neural networks intepretable and understandable for clinical uses, where adoption "has been slow". They note that the findings are important for future application of machine learning models in medicine because the "partnership" between the models and practicing clinicians "requires trust on the side of the clinical experts". Interestingly, solving the network opacity problem in this case is incentivized by "legal and ethical" requirements in addition to the technical advancement it represents. One of the methods the researchers present includes training neural networks to provide textual justifications along with their predictions, giving the end user some insight into why the network may have chosen its output the way it did. Another method discussed is the use of an "attribution map", which highlights an input's most relevant parts when it is used in a model (Salahuddin et al., 2021). Such a tool can be used to see what a model 'pays attention to' before producing its output, and thus provides a benefit to the clinician.

3. META-STUDY DESIGN

This meta-study examines some prominent techniques for explaining AI models. Some explainability methods are domain-dependent, and the domain for each explainability method will be noted in its subsection.

3.1. PREDICTION DIFFERENCE ANALYSIS

Prediction Difference Analysis (PDA) is an explainability method most applicable to models which utilize image data for inputs (Soldatos and Kyriazis 2021). The principle behind this technique is that an input's relevance to a prediction can be estimated by comparing a model's prediction to the prediction reached when the model does not have access to that input. Thus, the general form for this analysis on a model f which takes in an input x which contains an attribute a looks as such :

$$probDiff(y|x) = p(y|x) - p(y|x\backslash a)$$

However, for some models an attribute cannot simply be "removed", and must be replaced with another value. In this case, $p(x\backslash A)$ can be estimated by replacing a with all possible values in its domain A, as such :

$$p(y|x\backslash A_i) = \Sigma_{s=1}^{m_i} p(A_i = a_s) p(y|x \leftarrow A_i = a_s)$$

Notably, this method comes with some constraints. In applying it, researchers found that one of its main weaknesses lies in its application to models where the classification remains the same when some parameters are constant. Thus, a separate method should be used when "change in more than one attribute value at once is needed to affect the predicted value". Applying prediction difference analysis in this case would require "an extensive search" of "pairs, triples, etc." of permutations of input attributes, which is "unfeasible" (Šikonja and Kononenko 2007).

3.2. SALIENCY MAPS

Saliency maps are used in models which utilize images in their inputs. They are matrices with the same dimensions as the original inputs, and every cell in the matrix denotes how relevant that pixel of a specific input was to the correlated prediction

generated by the model. These maps can be generated using gradients of the neural network; however, care must be taken with these "gradient-based approaches" because, for example, Rectified Linear Units (ReLUs) have 0 gradient when not firing, but "a ReLU that does not fire can still carry information". To address this, researchers have devised a method called DeepLIFT to improve saliency maps by "multiplying the gradient with the input signal". Saliency maps generated by a typical gradient method and the DeepLIFT method are shown in Figure 1 below (Shrikumar et al., 2017).

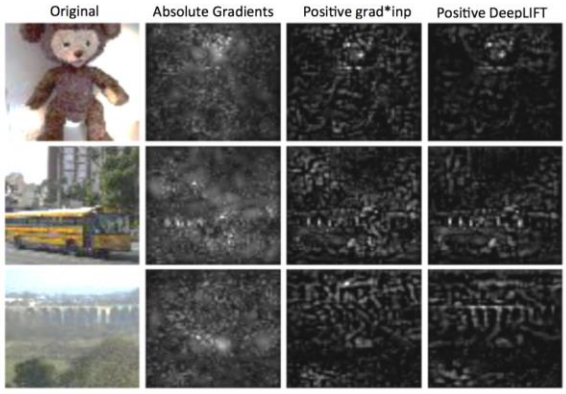

Figure 1. Comparison of saliency maps for different gradient-based methods.

3.3. SURROGATE MODELS

In trying to explain a machine learning model, a surrogate model can sometimes be generated to deduce insights about the rules the original model utilized to arrive at its predictions. One such method, TREPAN, is outlined by researchers as a way to extract "comprehensible, symbolic représentations from trained neural networks". This algorithm queries a neural network and generates a "decision tree" which is "accurate and comprehensible" and which holds "a high level of fidelity" to the network from which it was extracted (Craven and Shavlik 1995). TREPAN can be applied to models which take tabular data as inputs (Soldatos and Kyriazis 2021). After the TREPAN algorithm is conducted, the user has a decision tree of binary nodes available for making insights about their model.

Another extremely powerful surrogate model algorithm is LIME. The LIME algorithm "can explain the predictions of *any* classifier or regressor", making it more generalizable than methods which depend on the original model utilizing either tabular or image data for input (Ribeiro et al., 2016). LIME generates a model which seeks to match "the given feature vector and perturbed inputs" (Soldatos and Kyriazis 2021) to a particular prediction generated by the original model. An example of LIME applied to an image classification model is shown in Figure 2 below. In this case, the original model misclassified the husky image as an image of a wolf. LIME's generated explanation is shown on the right, which reveals that the original model was using snow in the background as a feature in its classification (Ribeiro et al., 2016).

(a) Husky classified as wolf (b) Explanation

Figure 2. LIME explanation for image classifier prediction of a husky input image.

4. OUTCOMES

The application and development of these explainable AI techniques marks a trend forward for the field. Some methods are tried and true, while some recent papers clearly push the envelope in explaining models in their respective areas.

5. CONCLUSION

Clearly, the field of AI explainability is one that warrants attention in the coming years. AI models continue to be deployed in more

and more domains, and the ability to understand as much as possible about their reasoning processes will be crucial as society increases in dependence on them.

6. FUTURE WORK

More domains can be analyzed for AI explainability than those discussed here. For example, an analysis of the models used in financial modeling and a discussion of how confident humans are in their reasoning would be beneficial to a meta-review of AI explainability.

Work should also be done to analyze the explainability of the quickly-growing large language models (LLMs) like ChatGPT. An analysis of how it is easier or more difficult to explain the outputs produced by these non-deterministic models would be both interesting and widely applicable for the many consumers of these tools in today's world.

REFERENCES

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. ArXiv:2302.13971 [Cs]. https://arxiv.org/abs/2302.13971

Karjian, R. (2023, September 22). *History and evolution of machine learning: A timeline*. TechTarget. https://www.techtarget.com/whatis/A-Timeline-of-Machine-Learning-History

Kay, A. (2001, February 12). *Artificial Neural Networks*. Computerworld. https://www.computerworld.com/article/2591759/artificial-neural-networks.html

Mark Craven and Jude Shavlik (1995). Extracting tree-structured representations of trained networks. Advances in neural information processing systems.

Marko Robnik-Šikonja and Igor Kononenko (2007). Explaining classifications for individual instances. IEEE Transactions on Knowledge and Data Engineering.

Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2021, October 31). Transparency of Deep Neural Networks for Medical Image

Analysis: A Review of Interpretability Methods. ArXiv.org. https://doi.org/10.48550/arXiv.2111.02398

Shrikumar, A., Greenside, P., Shcherbina, A., & Kundaje, A. (2017, April 11). Not Just a Black Box: Learning Important Features Through Propagating Activation Differences. ArXiv.org. https://doi.org/10.48550/arXiv.1605.01713

Soldatos, J. and Kyriazis, D. (Eds.) (2021). Trusted Artificial Intelligence in Manufacturing: A Review of the Emerging Wave of Ethical and Human Centric AI Technologies for Smart Production. Norwell, MA: Now Publishers.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. (2016). "Why Should I Trust You?" Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. https://doi.org/10.1145/2939672.2939778