Building a Phishing Link Detector

Exploring Explainability versus Accuracy in Artificial Intelligence Models

A Thesis Prospectus
In STS 4500
Presented to
The Faculty of the
School of Engineering and Applied Science
University of Virginia
In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science in Computer Science

By
Austin Huang

December 1, 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.

ADVISORS

Joshua Earle, Department of Engineering and Society

Briana Morrison, Department of Computer Science

Introduction

Artificial Intelligence (AI) has become increasingly accessible and used across a variety of areas. The average person can now spin up an AI model neural network with the many websites available online. One specific branch, Generative AI models, are now able to produce creative output. A well-known example is ChatGPT, which took the world by storm last year with its ability to code, produce essays, and argue. AI has slotted itself into more real world applications ranging from everyday tasks, such as optimizing our feeds on social media, predicting the weather, to even life-impacting activities, such as assisting with medical diagnosis.

Being able to utilize AI in solving more complex problems that can sometimes be met with skepticism. Although AI can be used to solve problems humans would not be capable of solving by themselves, the method by which they produce an answer is frequently hidden behind a black box of numbers and computations. The issue of the decision being hidden, somewhat ironically, only increases for more important uses of AI (Xu et al., 2019). It would be prudent that its users be able to trust it first. One prominent area of work towards increasing trust is the development of Explainable AI (XAI). XAI is a developing field that desires for AI models, in addition to giving an informed decision, also supplying a logical reasoning that a user would find reasonable and capable of following. However, for an AI model to be more accurate, intuitively, the model would need to be more complex, lessening the ability for its decision-making process to be explained (Bell et al. 2022). It is precisely this nuance that I want to research due to a curiosity that stemmed from this past summer.

This previous summer, I had the opportunity to develop my own AI model to detect phishing links. To do this, I applied the models I had learned about in school, combined with rudimentary Natural Language Processing. The model that we built had a high accuracy, but left me wondering the exact reasoning as to why our model flagged certain links and left others

alone. In this capstone project, I will elaborate on the development of the phishing detector in my technical project section. Subsequently, I will explore the nuances of explainability-accuracy tradeoff of AI for my STS project.

Technical Project

During the summer of 2023, I had the opportunity to intern as a Systems Engineer. While this was not aligned with my major, I was exposed to varieties of projects, specifically working with the Veterans Affairs to create a tool to help with their statement of benefits. However, the project I wanted to highlight from the internship for my technical project was my experience with a team of four other interns coding a Phishing Machine Learning model from scratch. My project was an application of what I had learned in school and introduced me to relevant applications in modern-day industry.

Phishing is the most common form of cybersecurity attack (Griffith, 2023). Usually, those with malintent distribute a link by email or other way of communication to attempt to trick the end user to click on it. These links have a variety of unfortunate consequences, such as stealing user data, identifying fraud, spreading viruses, or downloading malware or spyware. Sometimes a convincingly mimicked email disguises the link and at times the link itself will have a similar name to the intended link, maybe a letter off. Thus, for a human eye, detecting whether a link is malicious or not, is difficult; it is designed to be. Therefore, I thought that developing a model that could do the work of filtering out suspicious links would be valuable and interesting to explore.

During my project, I reviewed and learned about common Machine Learning models and also incorporated a brief amount of Natural Language Processing to the URLs, the more challenging aspect of our project. When trained and tested on a small dataset, our model

achieved a high F1-score, a single encompassing measure of accuracy that accounts for both false negatives and false positives, and the model had good promise to extend well to a broader application.

STS Project

Accuracy can be found mathematically and reported merely with a single percentage. It is this measure of accuracy that many would consider to be the ultimate teller of a better-trained, more appropriate model. For example, a weather forecast model that accurately predicts when it rains 90% of the time is better than an AI model that accurately predicts when it rains only half of the time. Less prominent is the explainability of an AI model. When an AI model makes a decision, who is to say why it made the decision? Some models are designed to inherently be statistically explainable, such as linear regression, while others are black-box models that are not easily understood. Unfortunately, many scenarios can not be adequately modeled by those that are inherently explainable. Thus, the explainability of an AI is a measure of how transparent it is to its users to be able to provide a reasoning adequate on the level of usefulness a user desires (Xu et al., 2019).

Now, in a time when AI is becoming more commonplace at the forefront of many decisions, it is increasingly becoming important for an AI to explain how it makes them (Ribiero et al., 2016). AI is being used to help with the judicial process, deliver medical diagnoses, making hiring decisions, managing finances, automate driving, and more (Xu et al., 2019). An individual's lives and livelihoods are impacted in all those examples, and the explainability of AI becomes an important factor in establishing trust in making accountable decisions in these serious matters.

When XAI is used to make decisions that impact people's lives, I foresee two main parties of people being affected; those that use XAI and those that XAI is used on. Generally, it may be that the subject experts who choose to use AI find that their social positions and worth are lowered now that their work could be replicated and explained by a machine. A doctor using XAI might only need to train as much as to repeat a model's justifications, bypassing the need to consider options and make decisions themselves. A judge might only need to review the specific laws and explanations XAI presents in determining criminal justice instead of carefully remembering and considering these laws first themselves. With less control over their tools, and greater public accessibility, one could predict that the public would disvalue subject experts, stifling further development and research.

Another looming problem exists not for systems that choose to use XAI. but rather those that are affected by XAI's justifications. In a society that becomes normalized to XAI, how would people respond to a machine's decisions over them? Many decisions by nature are subjective. Would it be the case that a machine designed to make a subjective decision might be seen as more objective and validating an opinion, offering some sort of truth not previously revealed? A danger of XAI might be hidden in a mask of an accessibility idealism that would reveal itself in a not-so-easy to explain way to those who are victims, judged by the machine and not the human: the medical patient who is diagnosed terminally ill by a machine, the applicant who is rejected by a machine, or the criminal who is guilty by a machine. The use of XAI would give power to those who are to judge against those who are to be judged.

Pivoting away from affected parties due to XAI, the particular topic I wanted to explore in this field is the supposed tradeoff between accuracy and explainability of AI. The general notion is that the more accurate an AI model is, the more complex it is and thus, the less

explainable it becomes. The reverse is also true, where the less complex an AI is, while it becomes more explainable, it will also yield less accurate results (Bell et al., 2022). This hypothesis has merit as it has been shown humans on average can only understand models with up to 7 different factors; any more and humans would render the model "functionally impossible to explain" (Candelon et al., 2023). This is in contrast to the magnitudes times more of factors for any typical neural network.

I want to focus on this tradeoff for my STS project. To explore the relationship between explainability and accuracy in AI, I will use Actor Network Theory. Actor Network Theory is a famous STS framework that views all entities in a considered situation as part of a network, establishing connections between entities or actors, that all contribute individualistically between actors and somehow contribute to the whole network as well. The Actor Network Theory has generally been effective in analyzing technologies and their different interactions in the world. To effectively utilize ANT, I plan to delve into a specific instance of an XAI algorithm, such as Local Interpretable Model-Agnostic Explanations (LIME) models, first developed by Marco Ribeiro, Sameer Singh, and Carlos Guestrin in 2016, to develop a network specific to that algorithm.

I will go about research in two main steps. First, I plan to analyze the context and validity of the explainability and accuracy tradeoff in AI models by looking at previous literary texts. Then, I will utilize Actor-Network Theory to explore the value of explainability in AI on a specific application of XAI, such as with the development of LIME in 2016, and contexts in which it is more important than others.

Key Texts

In an article titled, “It’s just not that simple: An empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy,” Andrew Bell, Ian Solano-Kamaiko, Oded Nov, and Julia Stoyanovich conducted a study that examined two common beliefs about the relationship between explainability and accuracy: the first hypothesis they examined was whether explainability and accuracy were tradeoffs. The second hypothesis they researched was whether there was any significant difference between the results of a black-box model or an interpretable model. For both these hypotheses, they were inconclusive (hence the article title). This article provided context towards the tradeoff I want to research and a way to quantify explainability.

Feiyu Xu, Hans Uszkoreit, Yangzhou Du, Wei Fan, Dongyan Zhao, and Jun Zhu write a brief overview of XAI in their article entry titled, “Explainable AI: A brief survey on history, research areas, approaches and challenges.” The title explains much as the article overviews XAI’s history, motivations and approaches, taking into account traditional and expert systems as well, ending with a discussion of anticipated challenges in the future. This paper was useful to introduce many of the technical terms regarding XAI and a little bit about how it connects to machine learning. Because it discusses the history of XAI, the paper contextualizes XAI and discusses its relevance. Specifically, this paper discussed the current challenges of XAI related to deep neural networks, and included references to some other potentially good resources.

Marco Ribiero, Sameer Singh, and Carlos Guestrin, in their article titled, “Why should I trust you? Explaining the predictions of any classifier,” details the development of a new algorithm called local interpretable model-agnostic explanations (LIME). LIME is an effort to offer an explanation of a model’s prediction, independent of the type of model used. Some

interesting questions they brought up were if the explanation offered by LIME were accurate to the model itself and how to verify it if it was. For example, if the original model decided that a flower was a tulip because of its color, would LIME be able to identify that it was due to color or would it claim some other factor led the model to predict the flower was a tulip. This study was particularly useful to identify trust as an important factor of many AI applications and the necessity of it in further improvements in the model.

In the article titled, “AI can be both accurate and transparent,” Francois Candelon, Theodoros Evgeniou, and David Martens write about how they conducted a study examining the validity of explainability and accuracy tradeoff in AI. They find that, in many cases, using a more interpretable model only accounted for a slight decrease in its accuracy. However, there would still exist some cases where it was necessary to use a complex model because of the nature of the problem, for example in computer vision problems where images and videos are the entities examined. This article was useful in examining firsthand results regarding this tradeoff and identifying some cases where it is necessary to have a less interpretable model in order to output results.

References

Bell, A., Solano-Kamaiko, I., Nov, O., & Stoyanovich, J. (2022). It's just not that simple: An empirical study of the accuracy-explainability trade-off in machine learning for public policy. *2022 ACM Conference on Fairness, Accountability, and Transparency*. https://doi.org/10.1145/3531146.3533090

Candelon, F., Evgeniou , T., & Martens, D. (2023, May 15). *AI can be both accurate and transparent*. Harvard Business Review. https://hbr.org/2023/05/ai-can-be-both-accurate-and-transparent

Griffith, C. (2023, October 6). *The latest phishing statistics (updated October 2023): Aag it support*. AAG IT Services. https://aag-it.com/the-latest-phishing-statistics/#:~:text=Phishing%20is%20the%20most%20common,100%20million%20phishing%20emails%20daily.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., & Müller, K.-R. (2019, March 11). *Unmasking Clever Hans Predictors and Assessing what Machines Really Learn*. Nature. https://www.nature.com/articles/s41467-019-08987-4

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135-1144).

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019). Explainable AI: A brief survey on history, research areas, approaches and challenges. *Natural Language*

Processing and Chinese Computing, 563–574.

https://doi.org/10.1007/978-3-030-32236-6_51