Thesis Project Portfolio

Explainability in GNNs: A Step Towards Global Self-Explainable GNNs

(Technical Report)

It Is Not My Responsibility: Failures in Preventing Malicious Deepfakes

(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Wendy Zheng

Spring, 2024

Department of Computer Science

Table of Contents

Sociotechnical Synthesis

The rapid evolution of artificial intelligence (AI) brings many new possibilities, but, at the same time, society is not prepared for this emerging technology. The fast development of AI outpaces our ability to understand it and adapt to it, exposing people to unintentional yet harmful outcomes. Use of AI in the real world without full understanding of how it works can result in incorrect and detrimental predictions. Furthermore, easy access to this powerful technology, along with lack of regulations against AI, allows bad actors to exploit it for personal benefits. The many unknowns of the technology call for action against the potential harms it has. Thus, the sociotechnical (STS) aspect of this study focuses on a specific use case of AI: deepfakes. I analyze the causes behind the rise of malicious deepfakes, which is necessary to determine the next step in effectively restricting the misuse of AI. The technical aspect focuses on understanding AI by finding explanations for its predictions. Knowing the reasoning behind its decisions makes it more transparent and prevents prejudice.

AI is often unexplainable; AI models learn by taking in large volumes of data and identifying possible correlations between the data and the desired predictions. Without further efforts, these correlations, to the human eye, are merely strings of numbers, so it is difficult to tell what the model learned and how it is applying that to make predictions. My technical project focuses on counterfactual explanations of Graph Neural Networks (GNNs), a class within AI models that perform on graphs. The counterfactual explanation of an input is the same input with minimal changes that causes the GNN to completely change its prediction. For example, if a model denies an application for a loan, a possible counterfactual could be the same application but with a slightly higher income and is approved by the model. Through code implementation, my group and I develop a generalized method to create counterfactuals during model training, meaning while the model is learning, we are simultaneously learning to generate global explanations. Global explanations are essentially common trends that are applicable to multiple inputs. Following the previous example, a global explanation could be increasing incomes to $50,000 generally causes the model to approve the application. The project is still in its development stage as our idea is being fine tuned.

My STS research studies a different perspective on the harms of AI: the misuse of deepfakes and the lack of action to minimize its negative effects. Deepfakes are fake media content generated by AI that manipulates a targeted person doing something that never happened. Victims of malicious deepfakes have no effective means to get assistance nor protect themselves, which enables people with malicious intent to harm others without consequence. Various actors take part in the network of malicious deepfakes. Thus, through document analysis, I uncover how their actions, or inaction, contribute to the proliferation of these deepfakes and what is needed to effectively counter the implications. Key actors, specifically the government, the deepfake software developers, and the deepfake creators, fail to take responsibility for this new technology, allowing for the effects of malicious deepfakes to spread in society.

While I accomplished much of my goals in both my technical and STS research, there are a few areas that need to be further examined. For my technical research, we use a rather heuristic method to find counterfactual explanations. As mentioned before, the model learning process is unexplainable, raising the need for counterfactual explanation. However, we use the same learning process to generate the explanations. Although they can provide insight on the model predictions, we do not know how the explanations are generated, circling back to the same problem. Thus, I think it would be interesting to discover an alternative method that is supported with theoretical proofs, which would give a concrete derivation of explanations. For my STS research, I focus on the causes for the continued development of malicious deepfakes, but I do not thoroughly cover the current attempts against malicious deepfakes. This is important because it indicates what methods are effective and influences the next steps in preventing malicious deepfakes.

I would like to thank Yinhan He and Professor Jundong Li for their mentorship and ideas. Additionally, I would like to acknowledge Alexander Shen and Carter Bassler for all their efforts on the technical project. Lastly, I would like to thank Professor Caitlin Wylie for her feedback throughout the STS project.