ECE 381V: Fair and Transparent Machine Learning

Project Team 1
Anubhav Goel, Shray Mathur, Devyani Maladkar, Shagun Gupta

26 October 2022

Problem Statement

In this work, we look at the Disagreement Problem and extend it to the domain of Graph Neural Networks (GNNs).

The Disagreement Problem

The disagreement problem looks at how explanations of different models differ and formalizes the notion of disagreement to measure this difference. For example, the top-k most important features output by different methods may differ from each other. We look at [3] to study the different metrics that are used to measure disagreement.

Explainability in Graph Neural Networks

Recently, Graph Neural Networks (GNNs) have become increasingly popular since many real-world data are represented as graphs, such as social networks, chemical molecules, and financial data and therefore, it becomes critical to ensure that the relevant stakeholders can understand and trust their functionality. To this end, graph post-hoc explanations are increasingly used to understand the behavior of graph neural networks (GNNs). Several approaches have been proposed to explain the predictions of GNNs [5], such as XGNN [7], GNNExplainer [6], PGExplainer [4], and SubgraphX [8], etc. These methods are developed from different angles and provide different levels of explanations. However, explainability in graph machine learning is still a nascent area, and therefore, it is important to ensure that the explanations generated by these methods are reliable. While prior research has already taken the first steps toward analyzing the behavior of explanation methods for tabular datasets [3], several critical aspects pertaining to graph methods can be further explored.

We study the disagreement problem in the domain of GNNs, and define metrics to formalize the notion of disagreement in this domain. We aim to define

metrics measuring disagreement which are translations of the ones defined in [3] as well as come up with novel metrics leveraging graph structure that are more domain-specific. We measure the disagreement using these metrics across various GNN architectures and datasets.

Possible Approaches

In this project, we aim to extend the framework of the Disagreement Problem to Graph Neural Network Explanations. Our approach is to analyze instance-level Explanations obtained from popular explanation methods applied to GNN models. Some metrics that we can implement draw directly from metrics introduced in [3]. However, these do not exploit the graph structure. We aim to come up with a separate set of metrics that leverage graph structure and introduce novelty in our work. The input to the disagreement metrics will consist of outputs of the explainability methods which are in the form of node importance, edge importance, and node features.

We plan to look at a few popular GNN architectures and will extend our study to a broader set of architectures depending on the time and scope of the project. For each architecture that we study, we plan to train this architecture on a variety of datasets. For each instance of the trained model, we plan to apply a set of explainability methods to our GNN and then measure the disagreement between various approaches by our defined metrics.

Datasets

We look to leverage GRAPHXAI [1], to evaluate the reliability of graph explanation methods across synthetically generated datasets and real-world datasets. The real-world datasets that we plan to use are as follows:

MUTAG. The MUTAG [2] dataset contains 1,768 graph molecules labeled into two different classes according to their mutagenic properties, i.e., effect on the Gram-negative bacterium S. Typhimuriuma.

Recidivism. The Recidivism [1] dataset includes samples of bail outcomes collected from multiple state courts in the USA between 1990-2009. It contains past criminal records, demographic attributes, and other demographic details of 18,876 defendants (nodes) who got released on bail at the U.S. state courts. Defendants are connected based on the similarity of past criminal records and demographics, and the goal is to classify defendants into bail vs. no bail.

Additionally, **PyG (PyTorch Geometric)**, a library built upon PyTorch provides easy access to a variety of datasets for GNNs which we can use for our analysis as well.

References

[1] Chirag Agarwal, Owen Queen, Himabindu Lakkaraju, and Marinka Zitnik. Evaluating explainability for graph neural networks. *arXiv preprint arXiv:2208.09339*, 2022.

[2] Jeroen Kazius, Ross McGuire, and Roberta Bursi. Derivation and validation of toxicophores for mutagenicity prediction. *Journal of medicinal chemistry*, 48(1):312–320, 2005.

[3] Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective. *arXiv preprint arXiv:2202.01602*, 2022.

[4] Dongsheng Luo, Wei Cheng, Dongkuan Xu, Wenchao Yu, Bo Zong, Haifeng Chen, and Xiang Zhang. Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33:19620–19631, 2020.

[5] Yaochen Xie, Sumeet Katariya, Xianfeng Tang, Edward Huang, Nikhil Rao, Karthik Subbian, and Shuiwang Ji. Task-agnostic graph explanations. *arXiv preprint arXiv:2202.08335*, 2022.

[6] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[7] Hao Yuan, Jiliang Tang, Xia Hu, and Shuiwang Ji. Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 430–438, 2020.

[8] Hao Yuan, Haiyang Yu, Jie Wang, Kang Li, and Shuiwang Ji. On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*, pages 12241–12252. PMLR, 2021.