Group 6 Project Proposal

Project Title: Explainability of Graph Neural Networks using Structured Deepwalks

Group Information: Zongyang Yue (Group Leader), Wenhan Yang, Dylan Kupsh, Baiting Zhu

Problem Statement:
Machine learning explainability facilitates user trust, and is crucial for furthering the machine learning application space into more sensitive areas, like medical treatment. As graph neural networks are an emerging subfield in machine learning, recent literature on explainability, although expanding, is insufficient. Our project team aims to transfer recent machine learning explainability models from image recognition to graph neural networks, aiming to increase the explainability of machine learning through structured confidence trees. In specific, we want to show how GNN models perform tasks to humans. Current graph explanation methods typically output a single set of important scores on edges and node features. However, inspired by the recent development in image explanation with structured attention graphs, we argue that there may be multiple explanations for a single graph classification tasks, and that logically combining them will improve the explainability of the model.

Dataset:
Since most of the GNN explainability methods are evaluated on synthetic datasets, we plan to start from here. There are existing synthetic datasets such as BA-SHAPES and BA-COMMUNITY, or we can generate datasets ourselves following the approach proposed by [Ying et al., 2019]. We also propose using a subset of the OGN-Products dataset, with a node classification task. As this dataset tries to predict item classification based on related Amazon product purchases, we believe the predictions are naturally situated to help human interpretability. As we are limited on computational resources, we aim to take a subset of this dataset, only 10 categories for instance instead of the full 47, which will hopefully yield shorter and less computationally intensive training times. Additionally, as this dataset serves as a popular dataset for classifying graph-neural network effectiveness, we believe the GNN explainability granted from our networks will help readers gain further intuition for more complicated GNNs.

Solution Plan:
Some of the previous models include: GNNExplainer: Given a graph G_0, node features X_0, and a trained GNN, GNNExplainer provides “explanations” on any node of a graph. [Ying et al., 2019], and GraphMask: predicts if an edge can be discarded. [Schlichtkrull et al., 2021]
First, we use different instance-level explainers or a single explainer with different hyperparameters to come up with different important scores. If the distributions of important scores are similar, we plan to modify the loss function to generate a different set of important scores that match with the other ones.
Second, we rank the node importance based on its scores and do a deep walk on the top three nodes in the pruned graph to generate subgraphs of high importance. Then, we continue to apply the explainers on the subgraphs to match the prediction result on the original graph. Then, we prune the graph and rank the nodes again. We plan to do two to three repetitions of the same procedure to come up with a structured attention graph for the original graph.

Evaluation Plan: We plan to evaluate by using:

1. Fidelity+: the difference of accuracy (or predicted probability) between the original predictions and the new predictions after masking out important input features. [Yuan et al., 2021]
2. Sparsity: good explanations should be sparse, which means they should capture the most important input features and ignore the irrelevant ones. [Yuan et al., 2021]
3. Stability: good explanations should be stable. Intuitively, when small changes are applied to the input without affecting the predictions, the explanations should remain similar
    a. Sparsity and Fidelity+/Fidelity- are highly correlated, we suggest comparing their Fidelity+ scores with the same level of Sparsity scores. [Yuan et al., 2021]
4. Plausibility:
    a. How convincing the explanations are to humans [Selvaraju et al., 2017] [Woo et al., 2018]

Schedule:

Week 5 and 6:

- Setup GNN for prediction score among datasets. For a given node, get the confidence score for important nodes. For this task, we will experiment using different GNN's generated by other papers

Week 7 and 8:

- Implement DeepWalk technique. For nodes with high importance, we will generate subnodes that significantly contribute to accuracy.

Week 9:

- Assemble DeepWalk + GNN to generate a new method for interpreting Graph Neural Networks

Week 10:

- Create interesting case-studies using this technique. Evaluate effectiveness of our technique and future improvements.

Workload:

Wenhan Yang: Implement the explainer models and the pruning method.
Dylan Kupsh: Implement explainer model, survey research
Baiting Zhu: Background research, model implementation and output analysis
Zongyang Yue: Setup GNN for prediction score among datasets, Evaluate effectiveness of our technique.

References:

1. Explainability in Graph Neural Networks: A Taxonomic Survey (https://arxiv.org/pdf/2012.15445.pdf)
2. One Explanation is Not Enough: Structured Attention Graphs for Image Classification (https://arxiv.org/pdf/2011.06733.pdf)
3. Explainability in Graph Neural Networks: An Experimental Survey (https://arxiv.org/pdf/2203.09258.pdf)