Qualifying Exam

Art Tay

Introduction

- Overview of the problem area.
  - Graphs are an important data structure.
    - GRAPHS provide an incredibly flexible structure for modeling complex data. Data can naturally appear as graphs, like molecules. We can reduce data to a graph, such as the key points of a image. We can even use graphs to add structure, such as grammatical relationships.
  - GNN models are good at prediction and inference on graph data.
    - Graph Neural Networks (GNNs) have become a popular choice for prediction and inference on graph data. At their core, GNNs work by iteratively updating node embeddings based on information from neighboring nodes. The idea is to use the graph's structure to engineer better features. This message passing scheme allows GNNs to capture complex dependencies and patterns present within the graph structure. GNN architectures typically consist of multiple layers, each performing message passing and aggregation operations to refine the embeddings. These layers are often followed by pooling and dense prediction layers to produce the final output.
  - There are many important applications for graph classification models.
    - Some important applications of graph classification include predicting chemical toxicity (Bai et al. 2019), classifying proteins (Gallicchio and Micheli 2019), and even detecting cancer from pathology slides (Xiao et al. 2023).
  - **Problem:** While GNNs achieve remarkable predictive power, their complexity prevents the exaction of the scientific rationale.

- Why is the problem important?
  - Explaining or interpreting GNN predictions would
    - help with the adoption of such models for critical applications,
    - prevent adversarial attacks,
    - detect potential implicit discrimination,
    - guide scientific as well as machine learning research.

- How does the problem relate to the fundamentals areas of Statistics?
  - Explain-ability vs Interpretability
    - Yuan et al. (2022)
    - A model is interpretable if the models decision process can be readily understood by humans. For example, a linear regression model is interpretable because the coefficient clearly define how any prediction get made.

        - * A model is explainable if the models prediction can be reasoned post-hoc. Permuting each variable and measuring the variation in the predictions can be used to estimate each variables marginal effect [cite].
    - One goal would be to create a GNN type model whose decision process is human interpretable. A straight translation from statistics would be a circuit type analysis [cite]. For graphs, this would mean some form of coefficients on subgraphs producing the prediction.
    - Another goal might be to develope a method that determines if a feature is statistical significant to the GNN model. The challenge is that the graph features that matter to researchers aren't necessarily tabular.
- What is the impact of solving this problem?
    - In the application where GNNs have shown strong predictive power, we can exact a testable scientific hypothesis for the nature of the classification.
    - In the application where GNNs have weak predictive power, highlight the potential misunderstandings the model is having.

Notation

Analysis of Core Papers

GNNInterpreter

(Wang and Shen 2024)

- Note on model-level explanations.
- Prediction objective.
- Embedding objective.
- Intuitive explanation of concrete distribution.
- Regularization terms.

D4Explainer

(Chen et al. 2023)

- Note on counter-factual explanations.
- Graph diffusion.

ProtGNN

(Zhang et al. 2021)

Synthesis of Core Papers

- Comparison of generation methods.
    - GNNInterpreter uses continuously relaxed discrete distributions.

  - D4Explainer uses diffusion.
  - Diffusion is slower, but can be more realistic. Probably because diffusion is less subject to the **out-of-distribution (OOD) problem**.
  - Prototype projection are like generative methods. Restricted to in distribution, but realism is all but guaranteed.

Technical Details

- Minimal reproduction of each method on MUTAG.

Future Directions

References

Bai, Yunsheng, Hao Ding, Yang Qiao, Agustin Marinovic, Ken Gu, Ting Chen, Yizhou Sun, and Wei Wang. 2019. "Unsupervised Inductive Graph-Level Representation Learning via Graph-Graph Proximity." https://arxiv.org/abs/1904.01098.

Chen, Jialin, Shirley Wu, Abhijit Gupta, and Rex Ying. 2023. "D4Explainer: In-Distribution GNN Explanations via Discrete Denoising Diffusion," no. arXiv:2310.19321 (October). https://doi.org/10.48550/arXiv.2310.19321.

Gallicchio, Claudio, and Alessio Micheli. 2019. "Fast and Deep Graph Neural Networks." https://arxiv.org/abs/1911.08941.

Wang, Xiaoqi, and Han-Wei Shen. 2024. "GNNInterpreter: A Probabilistic Generative Model-Level Explanation for Graph Neural Networks," no. arXiv:2209.07924 (February). https://doi.org/10.48550/arXiv.2209.07924.

Xiao, Guanghua, Shidan Wang, Ruichen Rong, Donghan Yang, Xinyi Zhang, Xiaowei Zhan, Justin Bishop, et al. 2023. *Deep Learning of Cell Spatial Organizations Identifies Clinically Relevant Insights in Tissue Images.* Preprint. In Review. https://doi.org/10.21203/rs.3.rs-2928838/v1.

Yuan, Hao, Haiyang Yu, Shurui Gui, and Shuiwang Ji. 2022. "Explainability in Graph Neural Networks: A Taxonomic Survey," no. arXiv:2012.15445 (July). https://doi.org/10.48550/arXiv.2012.15445.

Zhang, Zaixi, Qi Liu, Hao Wang, Chengqiang Lu, and Cheekong Lee. 2021. "ProtGNN: Towards Self-Explaining Graph Neural Networks," no. arXiv:2112.00911 (December). https://doi.org/10.48550/arXiv.2112.00911.