View-based Explanations for Graph Neural Networks (Extended Abstract)

Tingyang Chen
Zhejiang University
Zhejiang, China
chenty@zju.edu.cn

Dazhuo Qiu
Aalborg University
Aalborg, Denmark
dazhuoq@cs.aau.dk

Yinghui Wu
Case Western Reserve University
Cleveland, USA
yxw1650@case.edu

Arijit Khan
Aalborg University
Aalborg, Denmark
arijitk@cs.aau.dk

Xiangyu Ke
Zhejiang University
Zhejiang, China
xiangyu.ke@zju.edu.cn

Yunjun Gao
Zhejiang University
Zhejiang, China
gaoyj@zju.edu.cn

I. Extended Abstract

Generating explanations for graph neural networks (GNNs) is a crucial aspect to understand their decision-making processes, especially for complex analytical tasks such as graph classification [1]–[3]. Existing approaches [4]–[13] in this field are limited to providing explanations for individual instances or specific class labels. The main focus of these methods is on defining explanations as crucial input features, often in the shape of numerical encoding [14]. These methods often fall short in *providing targeted and configurable explanations for multiple class labels of interest*. Additionally, existing methods may return large or an excessive number of explanation structures, hence are not easily comprehensible. Moreover, these explanation structures often lack direct accessibility and cannot be queried easily, posing a challenge for expert users who seek to inspect the specific reasoning behind a GNN's decision based on domain knowledge.

To address these limitations, there is a growing need for more refined methodologies to explain the results of GNN-based decisions. Such methodologies should aim at offering "finer-grained" insights. Specifically, this would involve developing techniques that not only dissect the overall decision-making process of the GNN, but also *zoom in* on how certain features, nodes, or subgraphs contribute to specific classifications [14]. Moreover, enhancing the *accessibility*, *configurability*, and *queryability* of these explanations is paramount. Explanations should be presented in a *user-friendly* manner, possibly through visualizations or interactive tools that allow users to explore and interrogate the model's decisions. Such tools could enable other desirable capabilities such as highlighting critical substructures, providing interactive interfaces, and allow tunable parameters for domain experts to "query" the model about its decisions [15].

To this end, we propose GVEX [16], a novel framework that generates **G**raph **V**iews for GNN **EX**planation. **(1)** We design a two-tier explanation structure called *explanation views*. An explanation view comprises a collection of graph patterns along with a set of induced explanation subgraphs. Given a database $\mathcal{G}$ of multiple graphs and a specific class label l assigned by a GNN-based classifier $\mathcal{M}$, lower-tier subgraphs provide insights into the reasons behind the assignment of l by $\mathcal{M}$. They serve as both factual (that preserves the result of classification) and counterfactual explanations (which flips the result if removed). On the other hand, the higher-tier patterns summarize the subgraphs using common substructures for efficient search and exploration of these subgraphs. **(2)** We propose quality measures of an explanation view. Given multiple class labels of interest along with user-specified configuration parameters (e.g., specific sizes for each class label of interest), we formulate an optimization problem to compute the optimal explanation views for a GNN's explanation. We show that the problem is Σ_P^2-hard. **(3)** We present two algorithms. The first one adopts an *explain-and-summarize* approach, which begins by creating high-quality explanation subgraphs that effectively explain GNNs in terms of maximizing feature influence. Then it proceeds with a constrained graph pattern mining step to derive patterns [17], [18]. It is demonstrated that this approach yields an approximation ratio of $\frac{1}{2}$. The second algorithm works by processing an input node stream in batches in a single pass to incrementally maintain explanation views, ensuring an anytime quality guarantee with an approximation ratio of $\frac{1}{4}$. Our algorithms exhibit good performance across various graph types, including directed and undirected, sparse and dense, with or without node features. They are effective for both binary and multi-class classification tasks, in both static and streaming settings.

Evaluation. Using real-world benchmark data from a variety of domains and scales [19]–[22], we have experimentally validated the effectiveness, efficiency, and scalability of GVEX. GVEX-based methods outperform existing techniques in terms of conciseness, explanability, and efficiency. GVEX is designed to be "parallel-friendly" to manage graph instances at million-scale, and graphs having millions of nodes and edges. In particular, GVEX has the capability to handle both larger individual graphs and a large number of graph instances in few

hours. Moreover, through detailed case studies in chemistry, biology and social science domains, we demonstrate the potential wide range of practical applications of GVEX. This is particularly desirable for domain experts who seek for direct and intuitive explanations.

Outlook and Future Potential. Explainability is a key factor in the development of trustworthy artificial intelligence (AI) systems, particularly in ensuring transparency and reliability for interdisciplinary research [23]. To safely and trustfully deploy deep neural models, it is critical to provide human-intelligible explanations to end users and domain experts. Transparent integration with human-in-the-loop and explainability not only bolsters the reliability of AI applications, but also facilitates greater acceptance and understanding among users from diverse backgrounds. Our proposed GVEX paradigm is an important step in this direction by supporting *user-friendly, interactive, and configurable* explanations for GNNs, thereby closing the gap between intricate AI models and end users.

II. ACKNOWLEDGMENT

Tingyang Chen and Dazhuo Qiu contributed equally to this research. Tingyang Chen, Xiangyu Ke, and Yunjun Gao are supported in part by the NSFC under Grants No. (62025206, U23A20296) and Yongjiang Talent Introduction Programme (2022A-237-G). Dazhuo Qiu and Arijit Khan acknowledge support from the Novo Nordisk Foundation grant NNF22OC0072415. Yinghui Wu is supported in part by NSF under CNS-1932574, ECCS-1933279, CNS-2028748 and OAC-2104007. Xiangyu Ke is the corresponding author.

REFERENCES

[1] W. Hu, M. Fey, M. Zitnik, Y. Dong, H. Ren, B. Liu, M. Catasta, and J. Leskovec, "Open graph benchmark: Datasets for machine learning on graphs," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 22 118–22 133, 2020.

[2] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[3] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: A survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–37, 2022.

[4] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision (ECCV)*, 2014, pp. 818–833.

[5] R. Schwarzenberg, M. Hübner, D. Harbecke, C. Alt, and L. Hennig, "Layerwise relevance visualization in convolutional text graph classifiers," in *Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs@EMNLP)*, 2019, pp. 58–62.

[6] Q. Huang, M. Yamada, Y. Tian, D. Singh, and Y. Chang, "Graphlime: Local interpretable model explanations for graph neural networks," *IEEE Transactions on Knowledge and Data Engineering*, 2022.

[7] M. S. Schlichtkrull, N. De Cao, and I. Titov, "Interpreting graph neural networks for nlp with differentiable edge masking," in *International Conference on Learning Representations (ICLR)*, 2021.

[8] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, and X. Zhang, "Parameterized explainer for graph neural network," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 19 620–19 631, 2020.

[9] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec, "Gnnexplainer: Generating explanations for graph neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[10] H. Yuan, H. Yu, J. Wang, K. Li, and S. Ji, "On explainability of graph neural networks via subgraph explorations," in *International Conference on Machine Learning (ICML)*, 2021, pp. 12 241–12 252.

[11] H. Yuan, J. Tang, X. Hu, and S. Ji, "Xgnn: Towards model-level explanations of graph neural networks," in *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2020, pp. 430–438.

[12] Z. Huang, M. Kosan, S. Medya, S. Ranu, and A. Singh, "Global counterfactual explainer for graph neural networks," in *ACM International Conference on Web Search and Data Mining (WSDM)*, 2023, pp. 141–149.

[13] S. Zhang, Y. Liu, N. Shah, and Y. Sun, "Gstarx: Explaining graph neural networks with structure-aware cooperative games," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] H. Yuan, H. Yu, S. Gui, and S. Ji, "Explainability in graph neural networks: A taxonomic survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5782–5799, 2023.

[15] P. Gohel, P. Singh, and M. Mohanty, "Explainable AI: current status and future directions," *CoRR*, vol. abs/2107.07045, 2021.

[16] T. Chen, D. Qiu, Y. Wu, A. Khan, X. Ke, and Y. Gao, "View-based explanations for graph neural networks," in *ACM International Conference on Management of Data (SIGMOD)*, 2024.

[17] H. Shang, Y. Zhang, X. Lin, and J. X. Yu, "Taming verification hardness: an efficient algorithm for testing subgraph isomorphism," *Proc. VLDB Endow.*, vol. 1, no. 1, pp. 364–375, 2008.

[18] W.-S. Han, J. Lee, and J.-H. Lee, "Turboiso: Towards ultrafast and robust subgraph isomorphism search in large graph databases," in *ACM International Conference on Management of Data (SIGMOD)*, 2013, pp. 337–348.

[19] J. Kazius, R. McGuire, and R. Bursi, "Derivation and validation of toxicophores for mutagenicity prediction," *Journal of medicinal chemistry*, vol. 48, no. 1, pp. 312–320, 2005.

[20] K. M. Borgwardt, C. S. Ong, S. Schönauer, S. Vishwanathan, A. J. Smola, and H.-P. Kriegel, "Protein function prediction via graph kernels," *Bioinformatics*, vol. 21, no. suppl_1, pp. i47–i56, 2005.

[21] S. Freitas, Y. Dong, J. Neil, and D. H. Chau, "A large-scale database for graph representation learning," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

[22] W. Hu, M. Fey, H. Ren, M. Nakata, Y. Dong, and J. Leskovec, "Ogb-lsc: A large-scale challenge for machine learning on graphs," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks*, 2021.

[23] E. Dai, T. Zhao, H. Zhu, J. Xu, Z. Guo, H. Liu, J. Tang, and S. Wang, "A comprehensive survey on trustworthy graph neural networks: Privacy, robustness, fairness, and explainability," *CoRR*, vol. abs/2204.08570, 2022.