The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

Prototype-Based Explanations for Graph Neural Networks (Student Abstract)

Yong-Min Shin, Sun-Woo Kim, Eun-Bi Yoon, Won-Yong Shin[*]

Yonsei University, Seoul 03722, Republic of Korea
{jordan3414, kswoo9977, yon6286, wy.shin}@yonsei.ac.kr

Abstract

Aside the high performance of graph neural networks (GNNs), considerable attention has recently been paid to *explanations* of black-box deep learning models. Unlike most studies focusing on model explanations based on a specific graph instance, we propose *Prototype-bAsed GNN-Explainer (PAGE)*, a novel *model-level* explanation method for graph-level classification that explains what the underlying model has learned by providing human-interpretable prototypes. Specifically, our method performs clustering on the *embedding space* of the underlying GNN model; extracts embeddings in each cluster; and discovers prototypes, which serve as model explanations, by estimating the maximum common subgraph (MCS) from the extracted embeddings. Experimental evaluation demonstrates that PAGE not only provides high-quality explanations but also outperforms the state-of-the-art model-level method in terms of consistency and faithfulness that are performance metrics for quantitative evaluations.

Introduction

Despite the great success of GNN, the GNN models do not inherently offer explanations that enable us to gain valuable insight into the underlying model and build trust in model decisions (Yuan et al. 2020). Although *explanation* methods of GNN models have recently been studied, most of them have focused on *instance-level* explanations, i.e., explanations for each given graph instance (Ying et al. 2019; Baldassarre and Azizpour 2019), which however require a sufficient amount of input instances to be evaluated in order to decide whether the underlying model is trustworthy. On the other hand, *model-level* explanations can be an alternative to solving this problem since they lead to more abstract and concise explanations without any instance-wise explanation and do not necessitate one-by-one evaluation. XGNN (Yuan et al. 2020) was presented as a state-of-the-art model-level GNN explanation approach built upon reinforcement learning, which has a limitation of requiring domain knowledge to provide appropriate rewards. As a more interpretable model-level method with no need of domain-specific knowledge, we propose PAGE for graph-level classification, which is inspired by empirical findings that graphs exhibiting similar patterns tend to be embedded closely to each other in the graph embedding space. As one of distinguishable characteristics, our method provides human-interpretable *prototypes* as explanation results, each of which is defined as a graph where features most important to model decisions are encoded. In other words, such a prototype shared by instances with similar semantics is used for a model explanation. As illustrated in Figure 1, PAGE first performs clustering on the embedding space using the Gaussian mixture model (GMM). Then, it discovers prototypes by estimating the MCS from the embeddings extracted in each cluster.

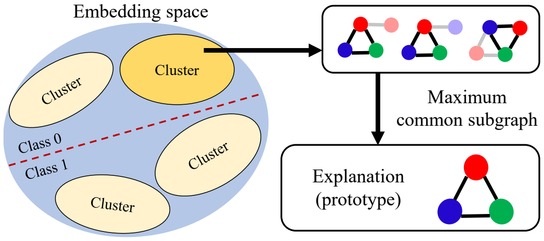

Figure 1: Overview of our proposed PAGE method.

Proposed Methodology

We present PAGE, a model-level explanation method for graph-level classification. As the first step of PAGE, we describe how to acquire clusters on the embedding space. We assume that a set of n input graphs, $\mathcal{G} = \{G_i\}_{i=1}^{n}$, and a GNN model f are given. By first feeding $\mathcal{G}$ into f to obtain node-level embedding vectors and passing through a readout function, we generate the set of graph-level embedding vectors, denoted as $\mathcal{H}_{\mathcal{G}} = \{\mathbf{h}_i\}_{i=1}^{n}$, at the penultimate layer of GNN. To discover groups of embeddings, each of which shares similar features learned by f, we fit the GMM on a subset of $\mathcal{H}_{\mathcal{G}}$ with the same class labels while estimating $\{(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)\}_{j=1}^{n_c}$ with mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$ for n_c clusters, where n_c is a pre-defined hyperparameter. For the l-th cluster, we select the k-nearest embeddings $\mathcal{K}_l = \{\mathbf{h}_{\pi(l,i)}\}_{i=1}^{k}$ using the Mahalanobis distance from each cluster's $\boldsymbol{\mu}_l$, where $\pi(l,i)$ is the index of the i-th nearest embedding in the l-th cluster.

As the second step of PAGE, we estimate the MCS for

[*] Corresponding Author.

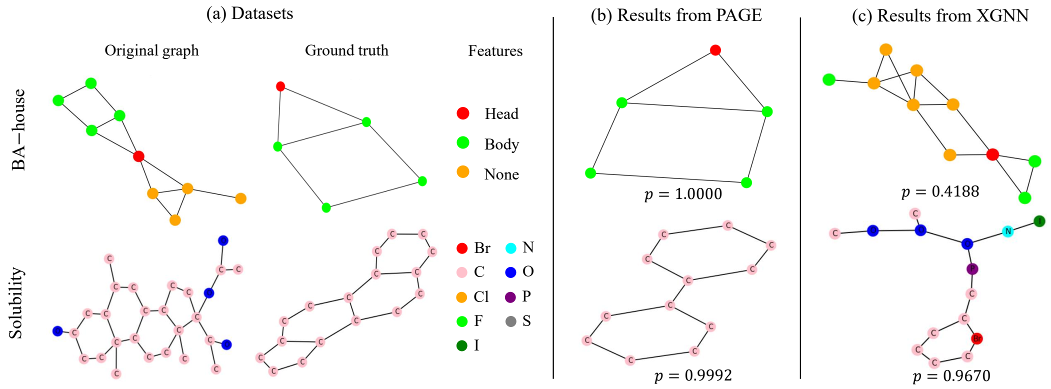

Figure 2: Qualitative results for PAGE and XGNN. (a) Datasets with the ground truth sets. (b) Prototypes discovered from PAGE and their model output probabilities. (c) Explanation results from XGNN and their model output probabilities.

Data	Consistency		Faithfulness	
	PAGE	XGNN	PAGE	XGNN
BA-house	**0.048**	0.312	**0.733**	0.328
Solubility	**0.109**	0.348	**0.591**	0.085

Table 1: Quantitative results for PAGE and XGNN with respect to consistency (the lower the better) and faithfulness (the higher the better).

each cluster in order to discover human-interpretable prototypes. To this end, with a modification for better convergence stability, we apply NeuralMCS (Ma et al. 2021) for all pairs in the set $\mathcal{K}_l$, which calculates the MCS of two graphs given their node embeddings by iteratively selecting node pairs exhibiting the highest embedding similarities.

Experimental Evaluation

Datasets

(1) **BA-house** (Ying et al. 2019): A graph is labeled as zero if it contains a house-shaped subgraph, corresponding to the ground truth set, and one otherwise.

(2) **Solubility** (Baldassarre and Azizpour 2019): The dataset is composed of real-world molecules, labeled by their solubility levels. We follow domain knowledge-aided explanations for the ground truth set (see Figure 2).

Model Settings and Performance Metrics

We employ graph convolutional network (GCN) (Kipf and Welling 2017) as a benchmark GNN model. In our study, we carry out both qualitative and quantitative evaluations. For the quantitative evaluation, we adopt two performance metrics: consistency and faithfulness (Sanchez-Lengeling et al. 2020). Consistency is the robustness of explanations across different GCN hyperparameters, which is measured by the standard deviation of model output probabilities of explanation results (e.g., prototypes in PAGE). Faithfulness is the quality of explanations versus the model performance, which is measured by the Kendall's tau coefficient between the model output probability of explanation results and the GCN's test accuracy.

Experimental Results

We compare our method with XGNN (Yuan et al. 2020), the state-of-the-art model-level explanation method. Figure 2 illustrates qualitative results for PAGE and XGNN. It is shown that, in contrast to the case of XGNN, PAGE successfully produces prototypes similar or identical to the ground truth for both datasets. It is also seen that the output probabilities, denoted by p, from PAGE are higher than those from XGNN. Table 1 shows quantitative results with respect to consistency and faithfulness. From the table, the superiority of PAGE is empirically verified.

Future Work

Our study is being extended to two cases including the node classification task and a more scalable solution to explanations.

Acknowledgements

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C3004345) and by the Yonsei University Research Fund of 2021 (2021-22-0083).

References

Baldassarre, F.; and Azizpour, H. 2019. Explainability techniques for graph convolutional networks. In *ICML Workshop*.

Kipf, T. N.; and Welling, M. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

Ma, G.; Ahmed, N. K.; Willke, T. L.; and Yu, P. S. 2021. Deep graph similarity learning: A survey. *Data Min. Knowl. Discov.*, 35(3): 688–725.

Sanchez-Lengeling, B.; Wei, J. N.; Lee, B. K.; Reif, E.; Wang, P.; Qian, W. W.; McCloskey, K.; Colwell, L. J.; and Wiltschko, A. B. 2020. Evaluating attribution for graph neural networks. In *NeurIPS*.

Ying, Z.; Bourgeois, D.; You, J.; Zitnik, M.; and Leskovec, J. 2019. GNNExplainer: Generating explanations for graph neural networks. In *NeurIPS*.

Yuan, H.; Tang, J.; Hu, X.; and Ji, S. 2020. XGNN: Towards model-level explanations of graph neural networks. In *KDD*.