GNNExplainer: Generating Explanations for Graph Neural Networks[Ying et al., 2019]

Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, Jure Leskovec

2019 NeurIPS

概要

グラフニューラルネットワーク (GNN) は、グラフ上の機械学習のための強力なツールである。GNN は入力グラフのエッジに沿って再帰的にニューラルメッセージを渡すことにより、ノードの特徴情報とグラフ構造を結合する。しかし、グラフ構造と特徴情報の両方を取り込むと複雑なモデルになり、GNN による予測を説明することは未解決である。本論文では、GNNEXPLAINER を提案する。これは、あらゆるグラフベースの機械学習タスクにおいて、あらゆる GNN ベースのモデルの予測に対して、相互に予測可能な説明を提供する、初の一般的でモデルにとらわれないアプローチである。GNNEXPLAINER は、GNN の予測に重要な役割を持つコンパクトな部分グラフ構造とノード特徴の小さなサブセットを特定する。さらに、GNNEXPLAINER はインスタンスのクラス全体に対して一貫性のある簡潔な説明を生成することができる。我々は GNNEXPLAINER を、GNN の予測と可能な部分グラフ構造の分布との間の相互情報を最大化する最適化タスクとして定式化する。合成グラフと実グラフを用いた実験により、我々のアプローチはノードの特徴だけでなく、重要なグラフ構造も特定できることが示され、説明精度において代替ベースラインアプローチを最大 43.0% 上回ることが示された。GNNEXPLAINER は、意味的に関連する構造を可視化する能力から、解釈可能性、欠陥のある GNN のエラーに対する洞察に至るまで、様々な利益を提供する。

1　どういう論文？

[Li et al., 2022]:「GNNExplainer は重要でないエッジ/ノードの特徴をマスクすることで GNN を説明することを提案した．具体的には、GNNExplainer は、個々のサンプルごとに入力グラフの学習可能なマスクを学習する。」

- GNN の予測を説明するためのアプローチである GNNEXPLAINER を提案
- 重要なサブグラフが GNN の予測との相互情報量を最大化するように最適化する
- 重要でないエッジとノードがマスクされ、予測に重要なサブグラフが強調される
- GNN のための最初の説明手法であると主張

そもそもの GNN の説明が簡潔でわかりやすかった

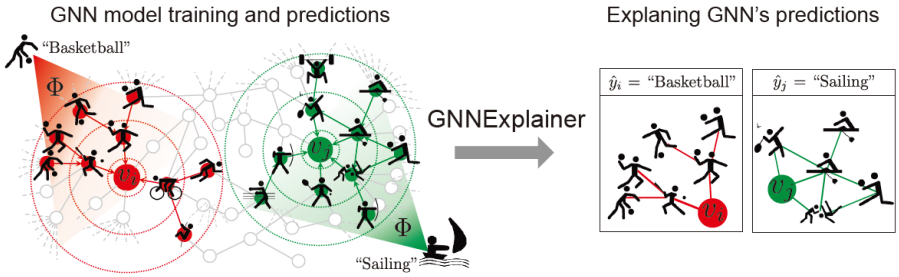

図 1: GNNEXPLAINER provides interpretable explanations for predictions made by any GNN model on any graph-based machine learning task. Shown is a hypothetical node classification task where a GNN model Φ is trained on a social interaction graph to predict future sport activities. Given a trained GNN Φ and a prediction $\hat{y}_i$ = "Basketball" for person v_i, GNNEXPLAINER generates an explanation by identifying a small subgraph of the input graph together with a small subset of node features (shown on the right) that are most influential for $\hat{y}_i$. Examining explanation for $\hat{y}_i$, we see that many friends in one part of v_i 's social circle enjoy ball games, and so the GNN predicts that v_i will like basketball. Similarly, examining explanation for $\hat{y}_j$, we see that v_j 's friends and friends of his friends enjoy water and beach sports, and so the GNN predicts $\hat{y}_j$ = "Sailing."

2 先行研究と比べてどこがすごい？

- CNN 等の GNN 以外のタイプの NN の説明手法は、グラフの本質である関係性情報を取り込む能力に欠けている。GNN の予測の説明には、ノードの特徴だけでなく、グラフが提供する豊富な関係情報を活用することが必要である。
- 既存法
  - 解釈可能な簡単なモデルに近似する手法 [Ribeiro et al., 2016, Augasta and Kathirvalavakumar, 2012, Lakkaraju et al., 2017, Zilke et al., 2016]
  - 勾配計算等で計算の重要な側面を識別する手法
    - 特徴勾配 [Erhan et al., 2009, Zeiler and Fergus, 2014]
    - 入力特徴に対するニューロンの寄与の back-propagation[Chen et al., 2018, Shrikumar et al., 2017, Sundararajan et al., 2017]
    - 反実仮想 [reference が不審？]
  -

で生成される saliency maps（顕著性マップ）[Zeiler and Fergus, 2014] はいくつかの事例で誤解を招く恐れが報告されており [Adebayo et al., 2018]、勾配飽和などの問題もある [Shrikumar et al., 2017, Sundararajan et al., 2017]
- グラフの隣接行列のような離散的な入力において悪化する。勾配値が非常に大きくても非常に小さな区間しかないため。
- GNN 向きではない

以下、よくわからなかった：

最近の GNN モデルは注目メカニズムを介して解釈可能性を増大させる [Schlichtkrull et al., 2020, Veličković et al., 2017, Xie and Grossman, 2018]。しかし、学習されたエッジの注目値は重要なグラフ構造を示すことができるが、その値は全てのノードに渡る予測に対して同じである。したがって、これは、あるエッジがあるノードのラベルを予測するためには不可欠であるが、他のノードのラベルを予測するためには不可欠でない多くのアプリケーションと矛盾する。さらに、これらのアプローチ

は特定の GNN アーキテクチャに限定されているか、グラフ構造とノードの特徴情報の両方を共同で考慮することで予測を説明することができない。

Finally, recent GNN models augment interpretability via attention mechanisms [Schlichtkrull et al., 2020, Veličković et al., 2017, Xie and Grossman, 2018]. However, although the learned edge attention values can indicate important graph structure, the values are the same for predictions across all nodes. Thus, this contradicts with many applications where an edge is essential for predicting the label of one node but not the label of another node. Furthermore, these approaches are either limited to specific GNN architectures or cannot explain predictions by jointly considering both graph structure and node feature information.

3 技術や方法のポイントはどこ？

- GNNEXPLAINER は学習した GNN とその予測結果を入力にとり、予測に最も影響を与えるノード特徴の小さなサブセットを持つ、入力グラフの小さなサブグラフの形式で説明を返す。
- GNNEXPLAINER は説明を、GNN が学習したグラフ全体のリッチな部分グラフとして指定し、その部分グラフが GNN の予測との相互情報を最大化するようにする。
- 結果、重要なグラフ経路を特定し、経路のエッジに沿って渡される関連ノードの特徴情報を強調できることを示す。
- モデルに依存せず、ノード分類、リンク予測、グラフ分類など、グラフに関するあらゆる機械学習タスクにおいて使える。
- シングルインスタンス説明の場合、GNNEXPLAINER は 1 つの特定のインスタンス（すなわち、ノードラベル、新しいリンク、グラフレベルラベル）に対する GNN の予測を説明し、
- マルチインスタンス説明の場合、GNNEXPLAINER はインスタンスの集合（例えば、与えられたクラスのノード）を一貫して説明する説明を提供する。
-
- ノード v が与えられた時、GNN の予測 $\hat{y}$ に影響を与える重要なグラフ構造 G_s とノードの特徴 $X_s = \{x_j | v_j \in G_s\}$ を識別する。重要性の概念を次のように相互情報量で定式化：

$$\max_{G_S} MI\left(Y, (G_S, X_S)\right) = H(Y) - H\left(Y \mid G = G_S, X = X_S\right) \tag{1}$$

- predicted label distribution Y. which can be expressed as follows:

$$H\left(Y \mid G = G_S, X = X_S\right) = -\mathbb{E}_{Y|G_S,X_S}\left[\log P_\Phi\left(Y \mid G = G_S, X = X_S\right)\right] \tag{2}$$

- 重要なサブグラフの隣接行列を fractional adjacency matrix（分数隣接行列？）として連続値に緩和。objective は次のように変形

$$\min_{\mathcal{G}} \mathbb{E}_{G_S \sim \mathcal{G}} H\left(Y \mid G = G_S, X = X_S\right) \tag{3}$$

- 凸の仮定により、Jensen の不等式は以下の上限を与える。

$$\min_{\mathcal{G}} H\left(Y \mid G = \mathbb{E}_{\mathcal{G}}\left[G_S\right], X = X_S\right) \tag{4}$$

- 実際には、ニューラルネットワークは複雑であるため、凸性の仮定は成立しないが、実験的に正則化を用いてこの目的を最小化すると、質の高い説明に対応する局所最小が得られることが多いことが分かった。
- $\mathbb{E}_{\mathcal{G}}$ の推定のために、多変量ベルヌーイ分布に分解された G の表現

$$P_{\mathcal{G}}\left(G_S\right) = \prod_{(j,k) \in G_c} A_S[j,k] \tag{5}$$

により、簡単に期待値を推定する

- 最終的に計算効率の良い変形として以下を勾配降下法で最適化

$$\min_{M} -\sum_{c=1}^{C} \mathbb{1}[y=c] \log P_{\Phi}\left(Y=y \mid G=A_S \odot \sigma(M), X=X_S\right) \tag{6}$$

- where $M \in \mathbb{R}^{n \times n}$ denotes the mask that we need to learn, $\odot$ denotes element-wise multiplication, and σ denotes the sigmoid that maps the mask to $[0,1]^{n \times n}$.
- Lastly, we remove low values in M through thresholding and compute the element-wise multiplication of $\sigma(M)$ and A_c to arrive at the explanation G_S for GNN's prediction $\hat{y}$ at node v.

4 どうやって有効と検証した？

- GNNEXPLAINER を合成グラフと実世界のグラフで評価した
- GNNEXPLAINER は GNN の予測に対して一貫性のある簡潔な説明を与える
- 2 つの実世界のデータセットを用いて、GNN の予測 $\hat{y}$ に影響を与える重要なグラフ構造 G_s とノードの特徴 $X_s = \{x_j | v_j \in G_s\}$ を識別することにより、GNNEXPLAINER がいかに重要なドメインの洞察を提供できるかを示す

5 議論はある？

- 最適化におけるパラメータ数は、予測を説明することを目的としているノード v の計算グラフ Gc のサイズに依存するが、計算グラフは一般に比較的小さいため、GNNEXPLAINER は入力グラフが大きくても効果的に説明を生成することが可能
- GNNEXPLAINER can be applied to: Graph Convolutional Networks [21], Gated Graph Sequence Neural Networks [26], Jumping Knowledge Networks [36], Attention Networks [33], Graph Networks [4], GNNs with various node aggregation schemes [7, 5, 18, 16, 40, 39, 35], Line-Graph NNs [8], position-aware GNN [42], and many other GNN architectures.

6 次に読むべき論文は？

- 最近の論文から読んだ方が流れ掴めそうなので、[Vu and Thai, 2020]
- その次に [Yuan et al., 2021]

参考文献

[Adebayo et al., 2018] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. (2018). Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.

[Augasta and Kathirvalavakumar, 2012] Augasta, M. G. and Kathirvalavakumar, T. (2012). Reverse engineering the neural networks for rule extraction in classification problems. *Neural processing letters*, 35(2):131–150.

[Chen et al., 2018] Chen, J., Song, L., Wainwright, M., and Jordan, M. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pages 883–892. PMLR.

[Erhan et al., 2009] Erhan, D., Bengio, Y., Courville, A., and Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, 1341(3):1.

[Lakkaraju et al., 2017] Lakkaraju, H., Kamar, E., Caruana, R., and Leskovec, J. (2017). Interpretable & explorable approximations of black box models. *arXiv preprint arXiv:1707.01154*.

[Li et al., 2022] Li, P., Yang, Y., Pagnucco, M., and Song, Y. (2022). Explainability in graph neural networks: An experimental survey. *arXiv preprint arXiv:2203.09258*.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

[Schlichtkrull et al., 2020] Schlichtkrull, M. S., De Cao, N., and Titov, I. (2020). Interpreting graph neural networks for nlp with differentiable edge masking. *arXiv preprint arXiv:2010.00577*.

[Shrikumar et al., 2017] Shrikumar, A., Greenside, P., and Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

[Sundararajan et al., 2017] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

[Veličković et al., 2017] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

[Vu and Thai, 2020] Vu, M. and Thai, M. T. (2020). Pgm-explainer: Probabilistic graphical model explanations for graph neural networks. *Advances in neural information processing systems*, 33:12225–12235.

[Xie and Grossman, 2018] Xie, T. and Grossman, J. C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters*, 120(14):145301.

[Ying et al., 2019] Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.

[Yuan et al., 2021] Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. (2021). On explainability of graph neural networks via subgraph explorations. In *International Conference on Machine Learning*, pages 12241–12252. PMLR.

[Zeiler and Fergus, 2014] Zeiler, M. D. and Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer.

[Zilke et al., 2016] Zilke, J. R., Loza Mencía, E., and Janssen, F. (2016). Deepred–rule extraction from deep neural networks. In *International Conference on Discovery Science*, pages 457–473. Springer.