Explainability in Graph Neural Networks: An Experimental Survey

Peibo Li , Yixing Yang , Maurice Pagnucco and Yang Song

17 Mar 2022

概要

グラフニューラルネットワーク (GNN) は、様々な応用分野において、グラフ表現学習のために広く開発されている。しかし、他のニューラルネットワークモデルと同様に、GNN はその背後にあるメカニズムを理解できないというブラックボックス問題に悩まされている。この問題を解決するために、GNN が行う決定を説明するための GNN 説明可能手法がいくつか提案されている。本サーベイでは，最新の GNN 説明可能性手法の概要と，その評価方法について述べる．さらに、新しい評価指標を提案し、実世界のデータセットにおいて GNN の説明可能性手法を比較するための徹底的な実験を行う。また、GNN の説明可能性に関する将来の方向性を提案する。

1 どういう論文？

- 最新の GNN 説明手法について、その概要と評価方法を俯瞰
- 評価手法や評価用のデータセットを含む、最新の GNN 説明可能性手法を批判
- 新しい評価指標 Explanation Confidence（説明一致度）を提案
- 実世界データセットと複数アーキテクチャを用いて様々な GNN 説明可能性手法を比較した実験結果を提示
- 最後に、GNN の説明可能性に関する将来の方向性について議論

2 先行研究と比べてどこがすごい？

- GNN 説明可能性に関する唯一のレビュー論文 [Yuan et al., 2020b] は、実験的研究を欠いている。GNN 説明可能性の現在の最先端技術について、詳細な実験的評価を伴う批判的レビューを提示
- 既存の評価指標は、グランドトゥルースと閾値に人間の知識が必要であり、特定のドメインでしか利用できない。実世界のほとんどのデータセットで比較できるようにするために、我々は人手を必要としない評価指標を提案する。
- 既存の評価指標と我々の新しい評価指標に基づき、3 つの citation ネットワークを用いて既存の説明可能性評価手法をベンチマークする。
- 異なる集約関数と深い構造を持つ複数の高度な GNN モデルに対する異なる手法の説明可能性を実験的に研究している

2.1 準備

GNNExplainer[Ying et al., 2019]

- GNNExplainer は入力グラフの摂動に基づき GNN を説明し、GNN の構造情報を利用する。
- しかし、パラメータの大きさが入力グラフの大きさに比例するため、スケーラビリティの問題に悩まさ

れる。
- GNNExplainer はインスタンスレベルの説明しか提供しないため、事前予測のグローバルな理解に欠ける。

PGExplainer [Luo et al., 2020]
- GNNExplainer の最適化の枠組みを踏襲しつつ、エッジの重みをバイナリ変数から範囲 (0, 1) の連続変数に緩和し、目的関数を勾配ベースの手法で効率的に最適化
- 学習済みモデルが複数のインスタンスに対して行った予測をまとめて説明するため、GNN モデルの全体像を把握することができる
- グラフ内の全てのエッジの重要度を予測するモデルを単独で用いることで、グラフサイズに依存しないパラメータサイズを実現し、計算効率を向上させることができる

GraphMask [Schlichtkrull et al., 2020]
- PGExpaliner と同様に、GraphMask は学習された GNN モデルのグローバルな理解を提供することができる。
- 各層に異なるマスクを与えることで関連するパスを提供する。

PGM-Explainer [Vu and Thai, 2020]
- PGM-Explainer は説明される特徴間の依存関係を図示することができる。
- しかし、PGM-Explainer はインスタンスレベルの説明に限定され、ベイジアンネットワークの学習過程は非常に計算量が多い。

SubgraphX [Yuan et al., 2021]
- SubgraphX の欠点は PGM-Explainer と同様、インスタンスレベルの説明しか提供できず、探索木のサイズが指数関数的に増大するため、大規模グラフに適用できないことである。

XGNN [Yuan et al., 2020a]
- XGNN はノードの分類タスクには適用できない。
- あるブラックボックスを使って別のブラックボックスを説明するというパラドックスに陥っている。

2.2　評価方法

Accuracy
- Accuracy はデータセットが人間が定義したグランドトゥルースの説明パターンを含む場合に使用される [Ying et al., 2019, Luo et al., 2020, Yuan et al., 2020a, Vu and Thai, 2020]
- 生成された説明がどれだけ ground truth に適合しているかを測定する:

$$\frac{|\mathbf{GT} \cap \mathbf{E}|}{|\mathbf{E}|} \tag{1}$$

- GT は ground truth の説明におけるエッジの集合であり、E は生成された説明におけるエッジの集合
- accuracy は生成された説明のサイズに関係することに注意。
- それゆえ、多くの場合、ユーザーは ground truth の説明のサイズに基づいて密な説明を閾値する必要がある。
- このため、データセットに対して人間が定義した ground truth を必要とし、accuracy はほとんどの実世界のデータセットに適用できない。

Sparsity and Fidelity
- Sparsity（スパース性、疎密性）は通常、Fidelity（忠実性）と組み合わされる。
- Sparsity は元のグラフからどれだけ多くの冗長なエッジが削除されたかを測定する。直感的に説明部分グラフは意思決定に必要なすべての情報を持つ最小の部分グラフであるべきなので、Sparsity が高い方が望ましい。

$$\text{Sparsity} \ = 1 - \frac{m}{M} \tag{2}$$

- m は重要な部分グラフのサイズ（すなわち、エッジの数）、M は元のグラフのサイズ

- Fidelity は説明モデルが元の GNN モデルにどれだけ忠実であるかを測定する。

$$\begin{aligned} \text{Fidelity} \ =& P(Y=c \mid G)-P\left(Y=c \mid G \backslash G_S\right), \\ & c=\arg\max_{c \in C} P(Y=c \mid G) \end{aligned} \tag{3}$$

- P はモデルが出力する確率分布、Y は予測値、G は元のグラフ、GS は説明、C は全クラスの集合

Inverse Fidelity（逆忠実度）[Schlichtkrull et al., 2020]

- 重要な部分グラフを用いたモデルの accuracy と、元のグラフを用いたモデルの accuracy を、テストセットに対するタスクで比較する

$$\text{InvFidelity} \ =\frac{1}{N} \sum_{N}^{i=1}\left(\mathbb{1}\left(\hat{y}_i^{\prime}=y_i\right)-\mathbb{1}\left(\hat{y}_i=y_i\right)\right) \tag{4}$$

- y_i はラベル、N はサンプル数、$\hat{y}_i$ と $\hat{y}_i'$ は元のモデルと説明の予測値。
- Inverse Fidelity はモデルの根本的な推論プロセスを理解することを動機としていて、Ground truth を必要としない。そのため全てのタスクで適用可能。実世界タスクは多くの場合 ground truth を利用できないから使えそう
- 人間が定義したグランドトゥルースは、モデルがどのように推論するか既に分かっていれば説明する必要がないため、正しいことが保証されず、この指標は信頼性が低くなってしまう。
-
- 説明可能性の評価は faithfulness（忠実性）vs plausibility（もっともらしさ）
- faithfulness: 説明がいかに正確にモデルの真の推論プロセスを反映しているか
- plausibility: 説明が人間にとってどれだけ説得力があるか

3　技術や方法のポイントはどこ？

difficulty 1:

- ground truth を得られないタスクでは faithfulness を評価できない
- スパースサブグラフ以外の重要度スコアしか出力しない全ての説明可能性手法に対して、連続エッジの重要度スコアをバイナリ説明にスパース化する必要がある。
- しかしスパース化に閾値を手動で設けるとなると、生成された説明の品質が保証されない。

difficulty 2:

- Sparsity を利用して、異なる sparsity を持つ説明を生成し、その fidelity を計算し、異なるスパース性の平均 fidelity を得る
- しかし、fidelity の範囲はサンプルやモデルによって大きく異なるため、fidelity に基づいて異なるモデルに対する説明可能性手法の性能を直接比較することはできない。
- さらに、我々が評価したいのは、重要でないエッジをできるだけ取り除きながら、元のモデルと同様の振る舞いをする説明モデルを見つける手法の能力であるが、sparsity は説明の質を評価するのではなくエッジを取り除くことだけに着目しているため、独立した評価項目として用いることができない。

Proposal:

- **explanation confidence: 説明適合性**を提案

$$\begin{aligned} EC=1-\frac{\left|P(Y=c \mid G)-P\left(Y=c \mid G_S\right)\right|}{P(Y=c \mid G)} \\ c=\arg\max_{c \in C} P(Y=c \mid G). \end{aligned} \tag{5}$$

- 説明サブグラフとオリジナルグラフを用いた予測クラスの確率の差を、オリジナルグラフを用いた予測確率で正規化したもの
- この値が高いほど、説明部分グラフが GNN モデルの実際の学習過程を反映していることを意味し、高い信頼性が得られる
- 全てのサンプルとモデルで正規化されているため、独立変数として用いることができるため、異なる説明可能性手法の性能を直接比較することが可能である

4 どうやって有効と検証した？

現実的なシナリオで説明可能な方法を評価するために、我々は GNN のための最も一般的な標準引用ネットワークベンチマークを 3 つ選択する。Cora [McCallum et al., 2000]、Citeseer [Giles et al., 1998]、Pubmed [Sen et al., 2008] の 3 つのデータセットである。これらのデータセットでは、ノードは文書を表し、エッジは引用を表します。ノードは文書の bag-of-word 表現の要素であり、各ノードは特定のクラスに属している。GCN、GAT、GCNII モデルを用いた。sparsity を縦軸、EC を横軸に取る集計で評価。

各説明手法の特徴を反映した結果を示した

- GraphMASK は、生成される説明がモデルの実際の推論過程に近いため、EC が高いほど Sparsity を上回っている。
- PGExplainer は GCNII では GNNExplainer を上回った。これは GNNExplainer はより多くのパラメータと深い構造を持つモデルの説明には不向きであることを示す。

5 議論はある？

- 問題の定義：GNN のみならず XAI に対して、より明確な定義が必要
- 評価指標：グラフデータは可視化困難で理解しにくい。そのためタスク非依存な、わかりやすい定量的な評価指標が必要
- 従来のグラフ理論に基づく：従来の手法は論理学や数学に基づいているため、GNN に基づく手法の説明とアルゴリズムに基づく手法の背後にある論理の関係を研究することは興味深いこと

6 次に読むべき論文は？

- GNNExplainer [Ying et al., 2019]
- PGExplainer [Luo et al., 2020]
- GraphMask [Schlichtkrull et al., 2020]
- PGM-Explainer [Vu and Thai, 2020]
- SubgraphX [Yuan et al., 2021]
- XGN [Yuan et al., 2020a]

参考文献

[Luo et al., 2020] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., and Zhang, X. (2020). Parameterized explainer for graph neural network. *Advances in neural information processing systems*, 33:19620–19631.

[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*.

[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.

[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., 2020a] Yuan, H., Tang, J., Hu, X., and Ji, S. (2020a). Xgnn: Towards model-level explanations of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 430–438.

[Yuan et al., 2020b] Yuan, H., Yu, H., Gui, S., and Ji, S. (2020b). Explainability in graph neural networks: A taxonomic survey. *arXiv preprint arXiv:2012.15445*.

[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.