Explainability in Graph Neural Networks: An Experimental Survey[Georgiev et al., 2021]

Peibo Li , Yixing Yang , Maurice Pagnucco and Yang Song

17 Mar 2022, Under review

概要

グラフニューラルネットワーク（GNN）モデルに関する最近の研究では、古典的なグラフアルゴリズムや組合せ最適化問題への GNN の適用に成功した。これには、前提条件が満たされない場合のアルゴリズムの適用や、十分な学習データが得られない、あるいは生成できない場合の学習済みモデルの再利用など、数多くの利点があります。しかし、GNN はブラックボックスモデルであり、直接解釈することができないため、これらのアプローチの主な障害は、説明可能性に欠けることである。本研究では、概念に基づく説明に関する既存の研究を GNN モデルに適用することで、この制限を解決する。GNN の読み出し機構を修正した概念ボトルネック GNN を導入する。3 つのケーススタディを用いて、我々は以下のことを実証する。(i) 提案モデルが、各ターゲットクラスに対して、正確に概念を学習し、学習した概念に基づく命題式を抽出できること、(ii) 提案概念ベース GNN モデルが、最先端モデルとの比較性能を達成できること、(iii) グラフレベルの概念に対して明示的にスーパービジョンを与えずに、グローバルグラフ概念を導出できること、である。

concept: 概念

！ イントロの書き方が自分の研究の参考になりそう

1 どういう論文？

conclusion から

- グラフアルゴリズムにおける概念に基づく推論機構を持つ Concept Bottleneck Graph Neural Networks を提案
- これを通して、性能に影響を与えることなく、ノードレベルの概念を正確に学習できることを示した
- 学習データとモデルの重みを調べることで、定義された概念に基づき、各ノードレベルの出力クラスを数式で説明することが可能である
- 概念によって、特定のグラフレベルのタスク（終了時期の決定など）の教師なしルール抽出を行うことができる
- 抽出されたルールは解釈可能であり、ルールを適用しても精度に大きな影響を与えない

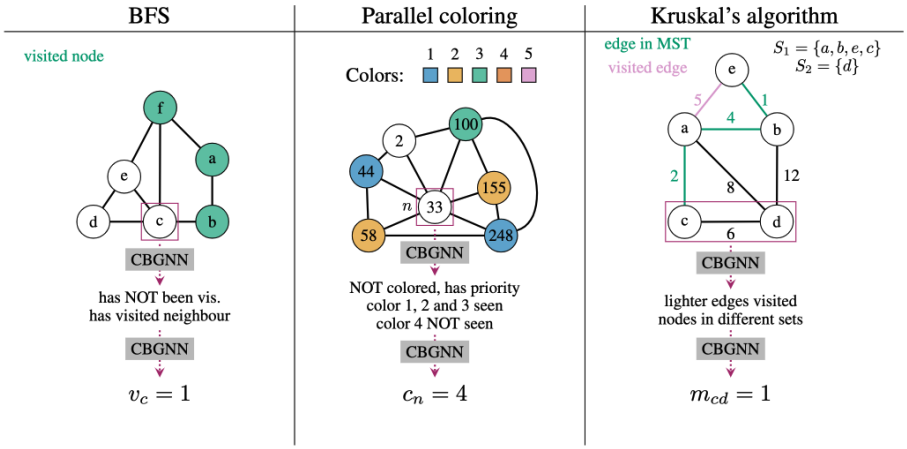

図 1: 概念ボトルネックグラフニューラルネットワーク (CBGNN) アプローチの概要。重要なことは、CBGNN モデルは、アルゴリズムのルールだけでなく、与えられたタスクのための概念情報を抽出するために訓練することができることです。3 つのアルゴリズムの例を挙げ、CBGNN がどのように入力データから概念を抽出し、それを使って出力をどのように計算するのかを示す。

2　先行研究と比べてどこがすごい？

GNN Explainability

- 既存法：
  - [Pope et al., 2019, Baldassarre and Azizpour, 2019, Schnake et al., 2020] の研究は、個々の予測を担う最も重要なノード／サブグラフを特定するために、CNN アプリケーションに用いられる特徴重要度勾配ベースのアプローチ（Class Activation Mappings や Layer-wise Relevance Propagation など）を GNNs に適応
  - [Ying et al., 2019, Vu and Thai, 2020, Luo et al., 2020] の研究は、相互情報量の最大化に基づくものや、特徴説明のマルコフブランケット条件付き確率など、GNN 説明可能性に特有の、より複雑なアプローチに焦点を当てている

  これらの研究は、
  - 本研究の焦点である組み合わせ最適化タスクに焦点を当てるのではなく、ソーシャルネットワーク、化学、または創薬に関わる GNN タスクとベンチマークに焦点を当てていることである。
  - 事前に訓練された GNN を事後的に説明することに焦点を当てている
  - 特徴量に基づく説明アプローチ（すなわち、入力ノード／サブグラフの相対的重要性を保持すること）に焦点を当てている
- 我々は、解釈可能な GNN モデルを構築することに焦点を当てている。
- 我々は代わりに概念に基づく説明アプローチに依存している。

Concept-based Explainability

- 既存の concept base の説明手法はは CNN の文脈でのみ概念を探求している
- RNN モデルの文脈で概念を探求している研究は [Kazhdan et al., 2020] だけであることである
- この研究では、GNN のための概念に基づく説明可能性に焦点を当て、[Koh et al., 2020] と同様に、概念は人間が指定したものとする

Combinatorial Optimisation for GNNs

- 提案法は [Veličković et al., 2020, Veličković et al., 2019] の拡張
- [Veličković et al., 2020] だけが [Ying et al., 2019] で説明を試みている
- しかし、彼らのモデルは、

1. モデル構造による説明可能性がなく
2. 局所的な説明を与えるために、単一のサンプルに対してさらなる最適化を必要とした。

- 他のすべての先行研究はブラックボックス方式で動作し、学習されたモデルの説明可能性を考慮しなかった。

3　技術や方法のポイントはどこ？

- GNN モデルの出力の前にコンセプトボトルネック層を適用する
- 中間概念処理に依存した新しいタイプの GNN である概念ボトルネックグラフニューラルネットワーク (CBGNN) を提案する。本論文は、概念ボトルネック法を GNN に適用した最初の研究である。
- 適切な概念の集合に依存し、それを監督することで、幅優先探索、クラスカルのアルゴリズムなどの古典的アルゴリズムのルールや、並列グラフコーディングなどのより進んだヒューリスティクスを導くことができること

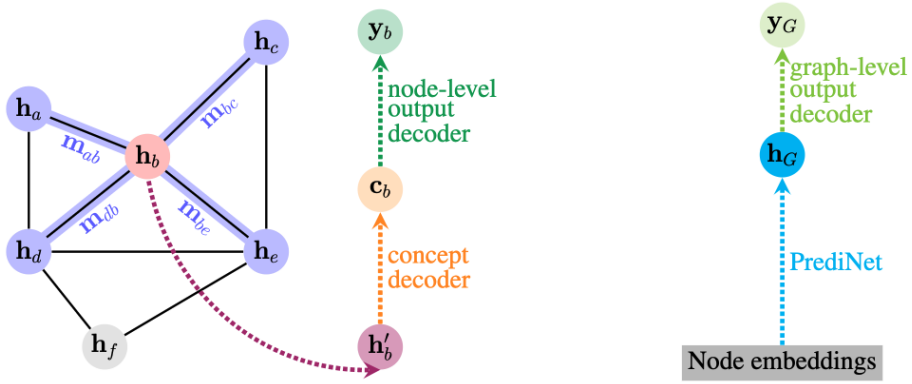

図 2: 左：ノードレベルの出力を生成するために、隣接ノードからのメッセージ m_{ij} は現在のノード表現 h_b と結合され、結果として更新された表現 h'_b となる。右：ノード埋め込みを PrediNet に通すことで、グラフレベルの埋め込み h_G を得ることができる。グラフレベルの出力 y_G（ここでは終了確率）を潜在状態 h_G から直接抽出する。グラフレベルの概念は、ノード概念に対する完全な列挙アプローチにより抽出される。

$$
\begin{aligned}
\mathbf{z}_i^{(t)} &= f_A\left(\mathbf{x}_i^{(t)}, \mathbf{h}_i^{(t-1)}\right) \\
\mathbf{z}_i'^{(t)} &= f_A\left(\mathbf{y}_i^{(t)}, \mathbf{h}_i^{(t)}\right) \\
\mathbf{H}^{(t)} &= P\left(\mathbf{Z}^{(t)}, E\right) \\
\mathbf{H}'^{(t)} &= P\left(\mathbf{Z}'^{(t)}, E\right) \\
\mathbf{c}_i^{(t)} &= \sigma\left(g'_A\left(\mathbf{z}_i^{(t)}, \mathbf{h}_i^{(t)}\right)\right) \\
\overline{\mathbf{H}^{(t)}} &= \text{PrediNet}\left(\mathbf{H}'(t)\right) \\
\mathbf{y}_i^{(t)} &= g_A\left(\mathbf{c}_i^{(t)}\right) \\
\tau^{(t)} &= \sigma\left(T_A\left(\overline{\mathbf{H}^{(t)}}\right)\right)
\end{aligned}
\tag{1}
$$

where σ is a logistic sigmoid function.

4 どうやって有効と検証した？

- 3 種類のケーススタディ（BFS、グラフカラー化、クラスカール）を用いて本アプローチを定量的に評価し、CBGNN アプローチが既存の最先端技術と同等の性能を達成できることを示す。
- CBGNN モデルによって利用される概念が、CBGNN が学習したヒューリスティックを要約するルールを提供するためにどのように利用できるかを示すことによって、我々のアプローチを定性的に評価する。

5 議論はある？

- GNN はこれらのアルゴリズム課題に対して、ノードまたは近傍情報を含む高レベルの概念を生成することができ、学習された概念抽出器は強く汎化されることがわかる（5 倍大きいグラフに対しても概念の精度は落ちない）。
- ボトルネックは最終的なモデルの精度に大きな影響を与えない

6 次に読むべき論文は？

参考文献

[Baldassarre and Azizpour, 2019] Baldassarre, F. and Azizpour, H. (2019). Explainability techniques for graph convolutional networks. *arXiv preprint arXiv:1905.13686*.

[Georgiev et al., 2021] Georgiev, D., Barbiero, P., Kazhdan, D., Veličković, P., and Liò, P. (2021). Algorithmic concept-based explainable reasoning. *arXiv preprint arXiv:2107.07493*.

[Kazhdan et al., 2020] Kazhdan, D., Dimanov, B., Jamnik, M., and Liò, P. (2020). Meme: generating rnn model explanations via model extraction. *arXiv preprint arXiv:2012.06954*.

[Koh et al., 2020] Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. (2020). Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR.

[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.

[Pope et al., 2019] Pope, P. E., Kolouri, S., Rostami, M., Martin, C. E., and Hoffmann, H. (2019). Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10772–10781.

[Schnake et al., 2020] Schnake, T., Eberle, O., Lederer, J., Nakajima, S., Schütt, K. T., Müller, K.-R., and Montavon, G. (2020). Higher-order explanations of graph neural networks via relevant walks. *arXiv preprint arXiv:2006.03589*.

[Veličković et al., 2020] Veličković, P., Buesing, L., Overlan, M., Pascanu, R., Vinyals, O., and Blundell, C. (2020). Pointer graph networks. *Advances in Neural Information Processing Systems*, 33:2232–2244.

[Veličković et al., 2019] Veličković, P., Ying, R., Padovano, M., Hadsell, R., and Blundell, C. (2019). Neural execution of graph algorithms. *arXiv preprint arXiv:1910.10593*.

[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). Gnnex-

plainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32.