A New Concept for Explaining Graph Neural Networks

Anna Himmelhuber[1,2], Stephan Grimm[1], Sonja Zillner[1], Martin Ringsquandl[1], Mitchell Joblin[1], Thomas Runkler[1,2]

[1]Siemens AG, Munich, Germany.
[2]Technical University of Munich, Munich, Germany.

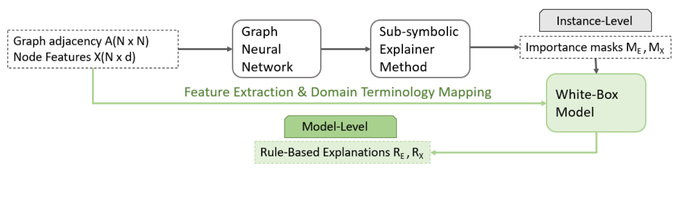
Introduction

Graph neural networks (GNNs), similarly to other connectionist models, lack transparency in their decision-making. A number of sub-symbolic approaches, such as generating importance masks, have been developed to provide insights into the decision making process of such GNNs. These are first important steps on the way to model explainability, but leaving the interpretation of these sub-symbolic explanations to human analysts can be problematic since humans naturally rely on their background knowledge and therefore also their biases about the data and its domain. To overcome this problem we introduce a conceptual approach by suggesting model-level explanation rule extraction through a standard white-box learning method from the generated importance masks.

Conceptual Schema

- We address this weakness of existing approaches by proposing a post-processing rule-based companion to such a sub-symbolic explainer method.
- Thereby we want to complement the sub-symbolic instance-level explanations with model-level rules. By extracting and aggregating global rule-based explanations through a standard white-box machine learning method from the generated explainer subgraph, we reduce the amount of additional interpretation needed by the user and provide a model-level explanation, that captures explanations about the global behavior of a model by investigating what input patterns can lead to a specific prediction.
- As an example of such an approach, we developed a novel method known as SUBGREX. We take the output of the state-of-the-art explainer method as input as well as graph-specific attributes such as node distances and network motifs and use decision trees to generate rule-based explanations.

Rule-Based Explanations of Subgraphs



- By combining the results of a sub-symbolic explainer method with a white-box rule generator, the representation needs for human comprehensibility and reasoning are satisfied.
- The rule-based explanation generation is not a stand-alone approach, but an add-on post-processing method in order to enhance the explanations and make them more user-centric.
- After training a GNN, the GNNs decision making process is interpreted by identifying a sparse receptive field containing influential elements. Our post-processing approach consists of taking these initial symbolic explanations and lifting them to the level of rules.

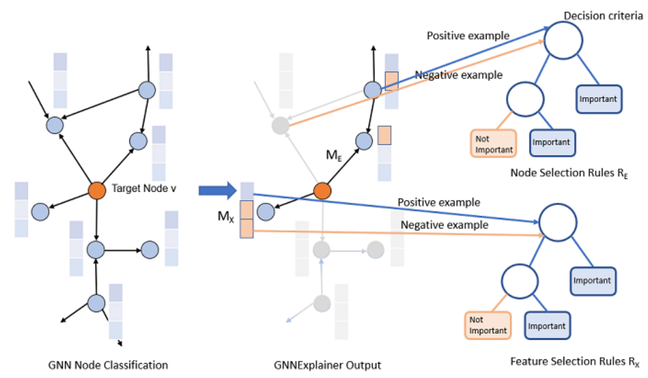
Explanation Generation

The proposed process for a node classification task, where an edge mask M_E and node feature mask M_X are generated by a sub-symbolic explainer model F_{ex}, and subsequently rules for edge and node features are created by the white-box models D_E and D_X. The rules are created through a classification process, where the individual edges and features are assigned binary labels “influential" or “not-influential" based on their masking value.

Algorithm 1 Rule-Based Explanation Generation

Inputs: Explanation Subgraph Model: F_{ex}; Graph adjacency: $A(N \times N)$; Node features: $X(N \times d)$; Set of attributes: $L = \{a_1, ...a_L\}$; Attribute mapping dictionary: $T = \{a_1 : t_1, ...a_L : t_T\}$; White-box model: D_E, D_X
for *category* $= 1, 2, \ldots, K$ do
  for *node* $= 1, 2, \ldots, N$ do
    M_X, $M_E = F_{ex}(node, A, X)$
    for support_node in M_E do
      If $M_E(support\_node) \rightarrow y_{support}$
      $x_{support} =$ Extract-Attributes($support\_node$, M_X, M_E, L)
      $x_{support_m} =$ Map-Attributes($x_{support}$, T)
      D_E.fit($x_{support_m}$, $y_{support}$)
    end for
    for support_node in M_X do
      If $M_E(support\_node) \rightarrow y_{support}$
      $x_{support} =$ Extract-Attributes($support\_node$, M_X, M_E, L)
      $x_{support_m} =$ Map-Attributes($x_{support}$, T)
      D_X.fit($x_{support_m}$, $y_{support}$)
    end for
  end for
end for

SUBGREX Model



To test this conceptual approach, we propose our SUBGREX model. We chose the GNNExplainer as the sub-symbolic explainer method F_{ex} for SUBGREX. As a method for extracting the rules, the standard machine learning mechanism decision tree is employed with $D_E = ID3_E, D_X = ID3_X$. The decision trees can be linearized into decision rules R_E and R_X.

Preliminary Results

$R_X^{Mutagen}$ ={If molecule contains atom C AND atom O; If molecule contains atom C AND atom S AND no atom O; If molecule contains atom H AND no atom C} Sensitivity: 0.98; Accuracy: 0.83

$R_E^{Mutagen}$ ={ If atom has more than 2 bonds AND if atom is part of an atom ring; If atom has only one bond AND is not part of an atom ring} Sensitivity: 0.72; Accuracy: 0.64

$R_X^{Nonmutagen}$ ={If molecule contains no atom N AND no atom H }

Sensitivity: 0.95: Accuracy: 0.94

Conclusions

- Conceptual vision for how to approach generating enhanced, more user-centric rule-based explanations from sub-symbolic instance-level explanations, which improve model-level understanding.
- Initial experiments demonstrate the validity of our method. Even with the rather simple SUBGREX method we show some surprisingly effective results in terms of meaningfulness of explanations and high sensitivity.
- In further research we plan to evaluate and compare the effectiveness of different white-box models including semantic web technologies such as inductive logic learning.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[2] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. 2016.

[3] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. In *Advances in neural information processing systems*, pages 6410–6421, 2018.

[4] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. In *Advances in neural information processing systems*, pages 9244–9255, 2019.

[5] D Doran, SC Schulz, and TR Besold. What does explainable ai really mean? a new conceptualization of perspectives. In *CEUR Workshop Proceedings*, volume 2071. CEUR, 2018.

[6] Phillip E Pope, Soheil Kolouri, Mohammad Rostami, Charles E Martin, and Heiko Hoffmann. Explainability methods for graph convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10772–10781, 2019.

[7] Asim Kumar Debnath, Rosa L Lopez de Compadre, Gargi Debnath, Alan J Shusterman, and Corwin Hansch. Structure-activity relationship of mutagenic aromatic and heteroaromatic nitro compounds. correlation with molecular orbital energies and hydrophobicity. *Journal of medicinal chemistry*, 34(2):786–797, 1991.

[8] Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*, 2018.

[9] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2018.

[10] Mark Craven and Jude W Shavlik. Extracting tree-structured representations of trained networks. In *Advances in neural information processing systems*, pages 24–30, 1996.

[11] Tim Miller, Piers Howe, and Liz Sonenberg. Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences. *IJCAI 2017 Workshop on Explainable Artificial Intelligence (XAI)*, 2017.

[12] Roberto Confalonieri, Tillman Weyde, Tarek R Besold, and Fermín Moscoso del Prado Martín. Trepan reloaded: A knowledge-driven approach to explaining artificial neural networks. 2020.

[13] Md Kamruzzaman Sarker, Ning Xie, Derek Doran, Michael Raymer, and Pascal Hitzler. Explaining trained neural networks with semantic web technologies: First steps. *arXiv preprint arXiv:1710.04324*, 2017.

SIEMENS

Technische Universität München

Supported by:
Federal Ministry for Economic Affairs and Energy
on the basis of a decision by the German Bundestag