Evaluating the explainability of graph neural networks for disease subnetwork detection

Sucharitha Rajesh (S.Rajesh-1@student.tudelft.nl)
Supervisors: Dr. Megha Khosla, Dr. Jana Weber

TU Delft

1. Introduction

- The explanations of **graph neural networks** highlight which parts of the graph were most important for the GNN's decision. These explanations can be used to **detect important subgraphs.**
- The quality of the explanations is important, however, there is a lack of direct empirical **evaluation** of the explanations.

- Integrating **standardized explainability evaluation metrics,** for example from the BAGEL benchmark [1] provides a fast and accurate way of evaluating both existing and new explainers with ease.

- Here we use **GNNSubNet** [2] as a case study: detecting the most important proteins for different types of cancer.

2. Research question

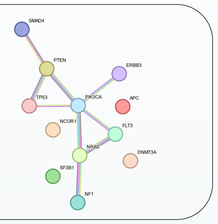
How do different explainability evaluation metrics evaluate GNN-SubNet?

3. Background

- **Graph Neural Networks**: take graphs as input and perform tasks like classification.
- **Protein-protein interaction networks (PPI):** graphs that model proteins as nodes and their interactions as edges.
- **Multi-omic datasets**: obtained by enriching the nodes of a PPI network with information like a patient's gene expression and DNA methylation.
- **GNNSubNet:** trains a GNN to classify the PPI graphs with data from cancer patients. Then finds an explanation of the GNN, revealing which subnetworks (sets of proteins) are most relevant in the development of a specific type of cancer.

4. GNNSubNet

PPI network
DNA methylation
Gene expression



Multi-omic dataset - patients with kidney cancer (cancer-specific); patients sampled randomly across other types of cancers (cancer-random)

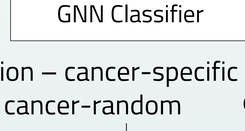
↓

GNN Classifier

Prediction – cancer-specific or cancer-random

↓

GNN Explainer



Node mask - global explanation for the specific type of cancer

↓

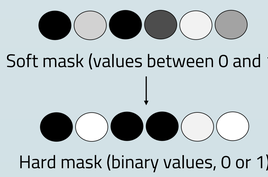
Clustering

Detecting disease subnetworks by using the explanation to weight the edges of the PPI graph

Goal: Evaluation

- Faithfulness
- Sparsity
- Validity

- GNNExplainer generates **soft** node masks, while validity requires **hard** masks.



Transform by taking the highest values [3]:

- Top-30% (S-0.7)
- Top-50% (S-0.5)

6. Results and discussion

Evaluating on synthetic dataset:

- All metrics have an ideal target score of 1, except validity+ that has an ideal target of 0.5.

Evaluating on KIRC cancer dataset:

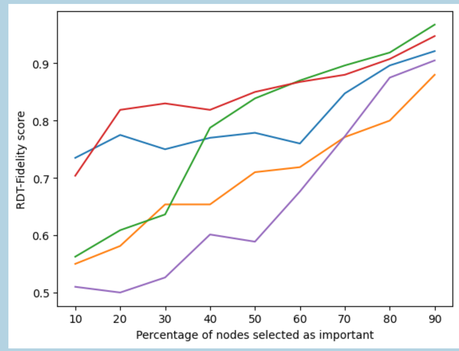
- High RDT-fidelity, low sparsity: the explanation is robust against perturbations of the graphs but is very dense (highlights many nodes as important).
- High validity- : the nodes highlighted by the explanation have high discriminative power
- Low validity+ with a high standard deviation: explanations found are very different from each other

Additional observations:

- Tradeoff between RDT-fidelity and sparsity: when enforcing a sparser node mask, RDT-fidelity decreases.
- Subclusters are unstable: different subclusters found every time the model is trained and explained.

Metric	Threshold	Average	Std.deviation
RDT-fidelity	NA	0.826	0.11
Sparsity	NA	0.040	0.02
Validity+	S-0.7	0.232	0.19
	S-0.5	0.292	0.20
Validity-	S-0.7	0.840	0.15
	S-0.5	0.843	0.01

Average metric scores over 10 evaluations of the global explanation of the KIRC dataset.



RDT-fidelity score computed with varying percentages of nodes selected as important in the hardmask (i.e. varying sparsity levels)

5. Methodology

Evaluate **explainability metrics** over 10 iterations of training and explaining:

- **RDT-fidelity**: A faithful explanation stays robust - small perturbations to unimportant nodes do not change the model's prediction.
- **Sparsity**: A sparse explanation needs to highlight a **small** number of important nodes, since the entire input is always a trivial explanation.
- **Validity-:** A valid explanation does not have a change in the prediction when unimportant nodes are set to average values
- **Validity+** : A valid explanation has a change in the prediction when the important nodes are set to average values.

7. Conclusions

- Metrics should be used in complement to each other for a more complete picture.
- For subnetwork detection, size and variability of subnetworks can be a useful measure.

8. Limitations

- Low model accuracy could bias metric scores.
- Metrics can be refined further and improved with domain knowledge.
- Clear metrics for size and variability of subnetworks can be defined and used.

References

[1] M. Rathee, T. Funke, A. Anand, and M. Khosla, "Bagel: A benchmark for assessing graph neural network explanations," arXiv preprint arXiv:2206.13983, 2022.

[2] B. Pfeifer, A. Secic, A. Saranti, and A. Holzinger, "GNN-SubNet: Disease subnetwork detection with explainable graph neural networks," Jan. 2022. doi: 10.1101/2022.01.12.475995.

[3] T. Funke, M. Khosla, M. Rathee, and A. Anand, Zorro: Valid, sparse, and stable explanations in graph neural networks, 2022. arXiv: 2105.08621.

Poster template provided by PosterNerd.