UCLA Samueli School of Engineering

To Explain Better: Structured Attention Graphs for Graph Classification

Dylan Kupsh, Wenhan Yang, Zongyang Yue, Baiting Zhu
Course Project of CS 249 – Graph Neural Network (GNN)

Background

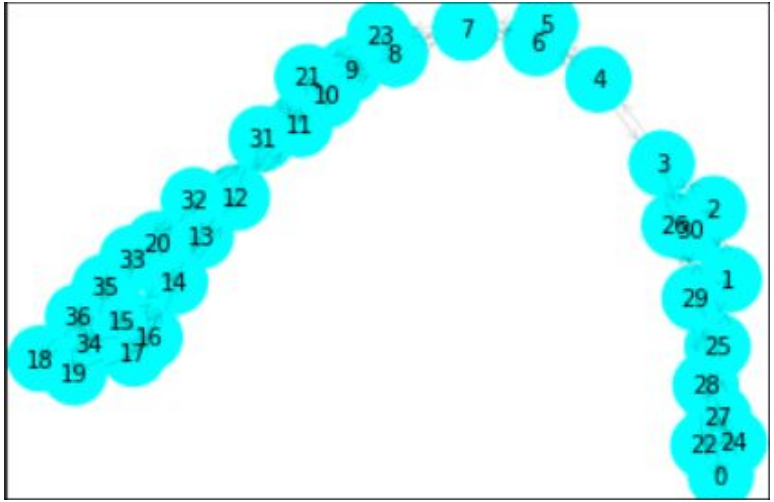
- Graph: powerful format of representation that can represent complex relational data
  - social network, biological compounds, etc.
- To process such rich data efficiently: Graph Neural Networks
  - Incorporate neighborhood information
  - Capture graph structure
- Problem: Transparency
  - No clear explanation for the predictions
  - Which node/link determines the result?
- Why does it matter?
  - Important to our understanding on the deep GNNs.
  - Increase trust in the models

Existing Works

- GNNExplainer
  - Initialize random soft masks for edges & node features to learn the important components of the graph
- SA
  - Use gradients as the indicator of input features' importance (including edges, nodes, node features, etc).
  - Assume that higher gradient values is equivalent to more important features
- XGNN
  - Explain GNNs via graph generation
  - Trains a graph generator so that the generated graph can maximize the target graph prediction.

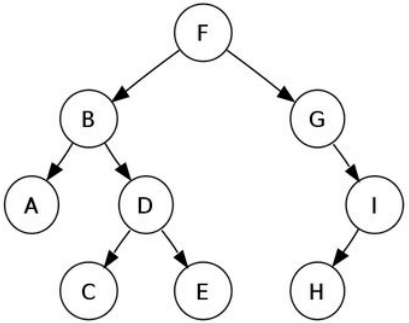
Our proposal: one explanation is not enough

- Problem: The visualization & logical inference is not intuitive, especially for the larger graphs
  - Assigning weights to nodes/features
  - Graph classification scenario: difficult to see directly what is the essential component of the graph.
  - Lack of logical structure
  - Limited visualization capacity



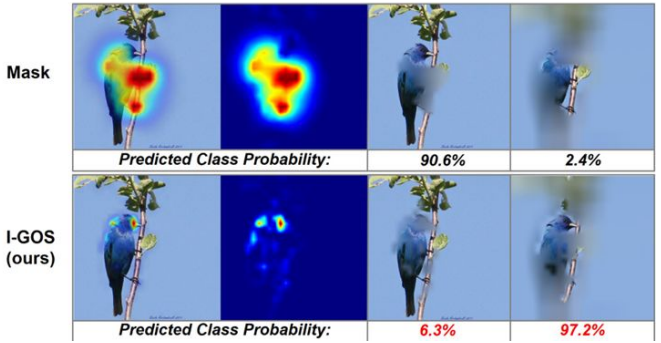
Our proposal: one explanation is not enough

- Our solution: Construct logical graph to make the explanation more intuitive and interpretable.
  - Pick the confident components and present the information in disjunctive normal form.
  - SAG: Structure Attention Graph
    - Directed, acyclic graph over attention maps
    - Connected based on containment relationship

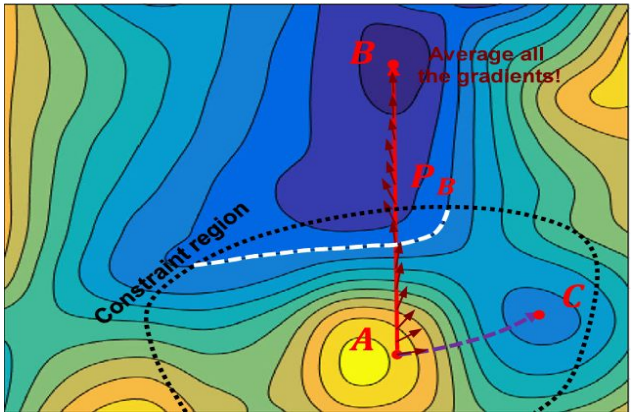


Related Works

I-GOS (Qi et al.) tries to find the region, which after being masked, the original classification score maximally drops. They also hope that if the mask are “reversed”, the classification score is high.



I-GOS imagines a line from starting point (original image) to the global optima (masked image) and try to use use the Integrated Gradient to approach the answer.



Related Works

CAM (Zhou et. al) and **Grad-CAM (Shitole et. al)** both try to explain predictions of CNN-based models. This model tries to identify important neurons in the last Conv Layer before flatten. CAM looks at activation map only. Grad-CAM also claims the gradient of a class *c* captures the "importance" of feature map.

Class Activation Mapping(CAM)

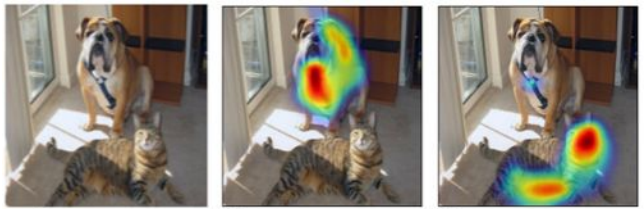
$$M_c(x,y) = \sum_k w_k^c f_k(x,y)$$

k: unit at (x,y) f_k: activation of k

c: class y^c: score for c before softmax

Grad-CAM

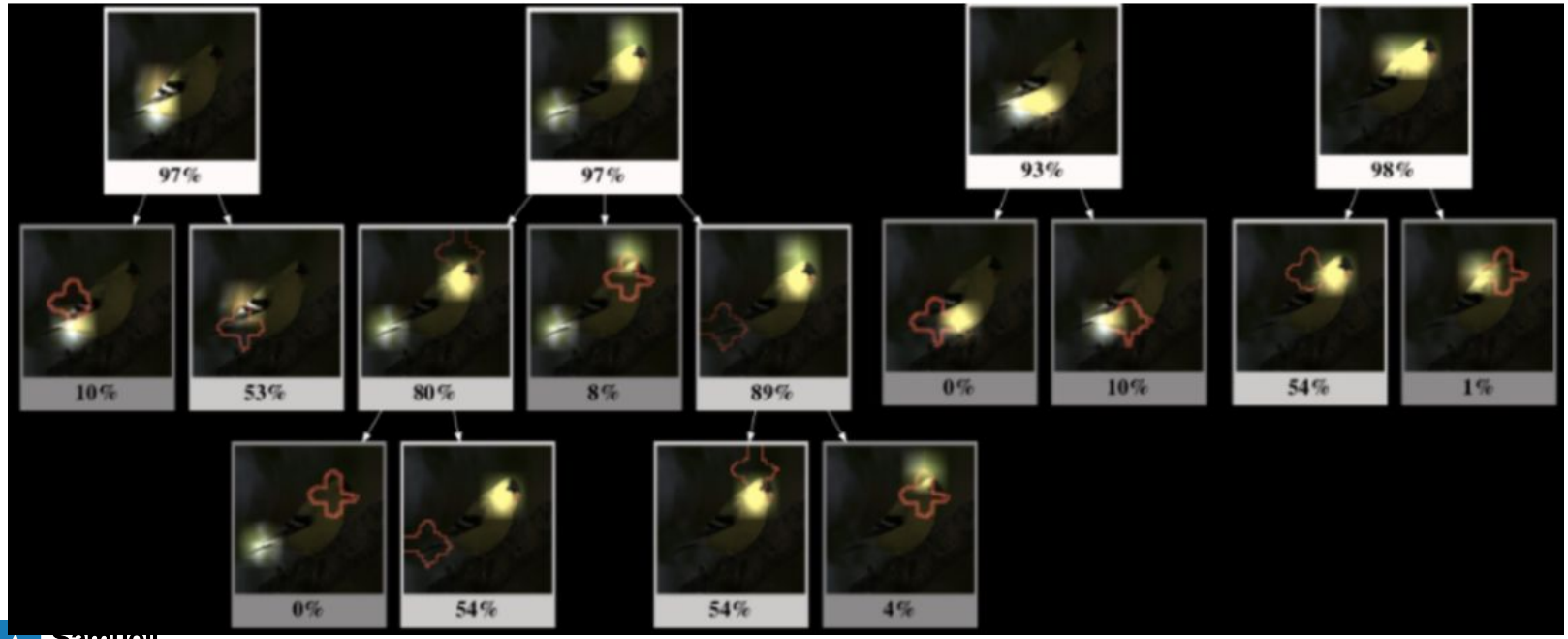
$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

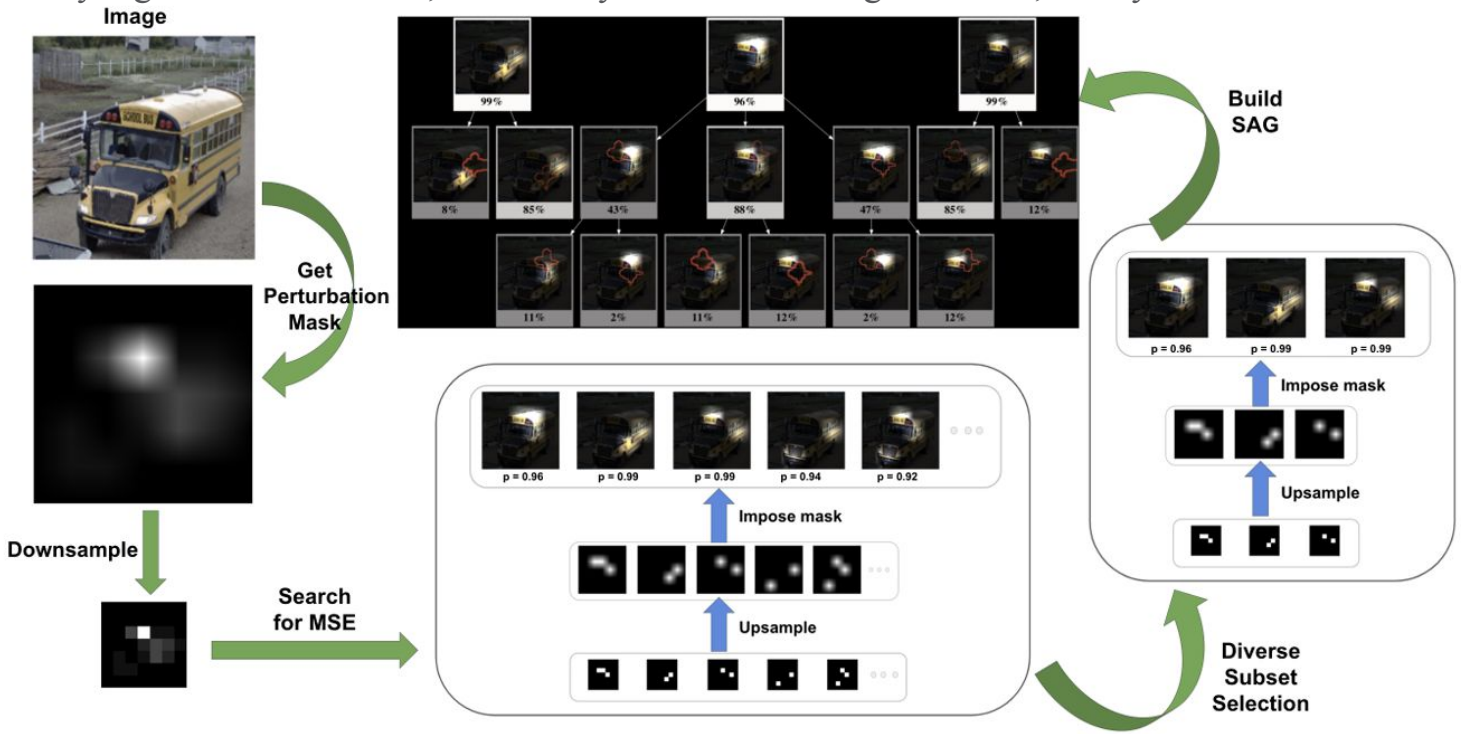
Related Works

SAG (Shitole et. al) claims that a single saliency map is not enough for explanation, and thus tries to build a hierarchical explanation, a tree.

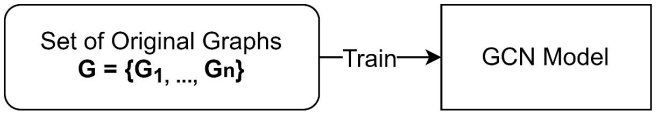


Related Works

SAG (Shitole et. al)'s structure and idea is simple. We impose mask, do a down-scaling, search for the most likely region based on MSE, then finally revert the scaling and mask, finally build a SAG.

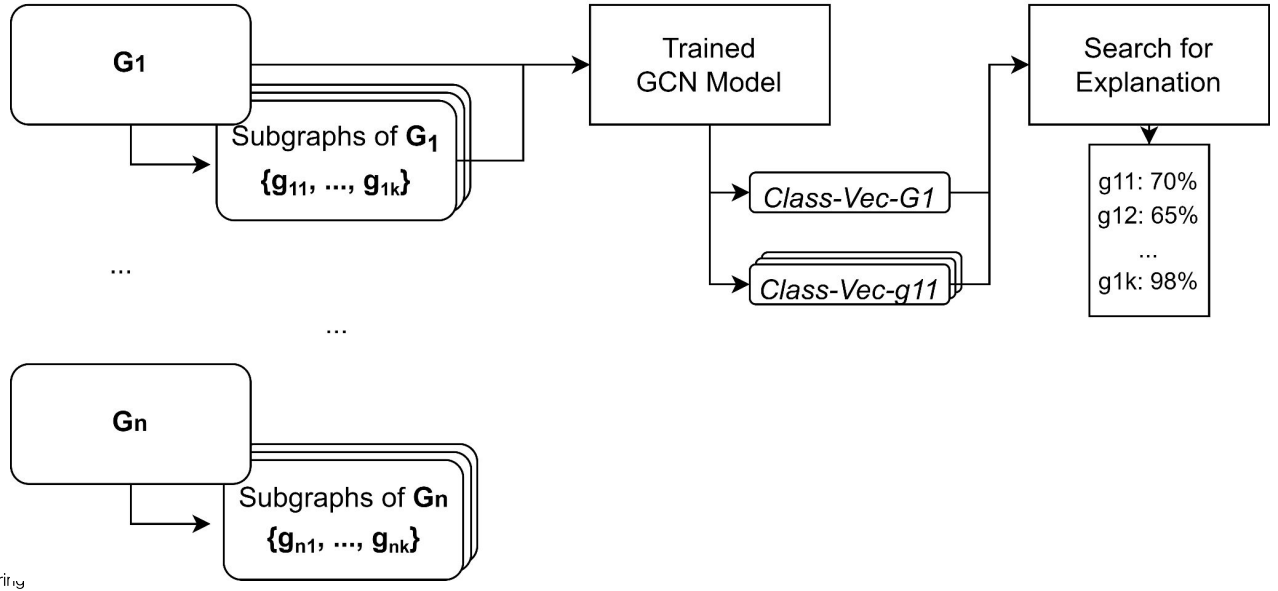


Objective & Model



We want to learn from the approach of SAG, and apply similar concept to Graphs instead of images. To achieve this, we need to:

1. pretrain a GNN model
2. find a sub-graph method



Dataset

We use published datasets in the PyG / torch_geometric library.

First dataset: MUTAG

- A collection of nitroaromatic compounds represented by each graph
- 188 graphs, 2 classes
- nodes are atoms and labeled by atom type
- edges are bonds between corresponding atoms

Second dataset: ENZYMES

- Protein tertiary structures obtained from the BRENDA enzyme database
- 600 graphs, 6 classes
- nodes represent secondary structure elements
- an edge connects two nodes if they are neighbors along the amino acid sequence or one of three nearest neighbors in space.

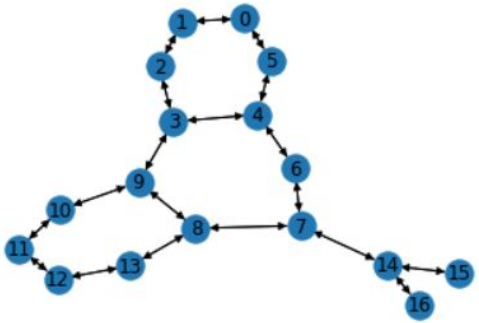
Experiment Setting

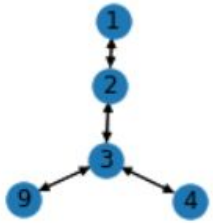
Minimal Sufficient Explanations: any one of the subgraph that is sufficient for high confidence.

High confidence: with a threshold value < 1.

Full Graph goes through pre-trained GCN: class C with 90% confidence.

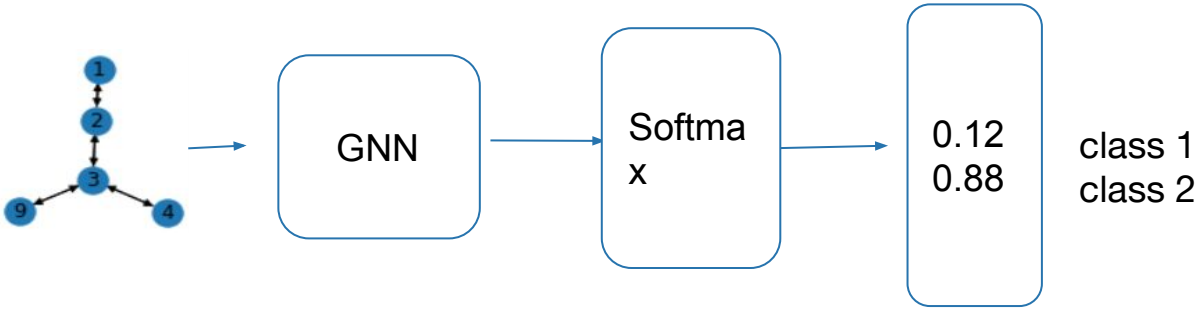
Subgraph goes through pre-trained GCN: class C with >= 90% * threshold confidence.
Yeah! it is a Minimal Sufficient Explanation.
Use the value of the softmax result of that class to measure confidence.

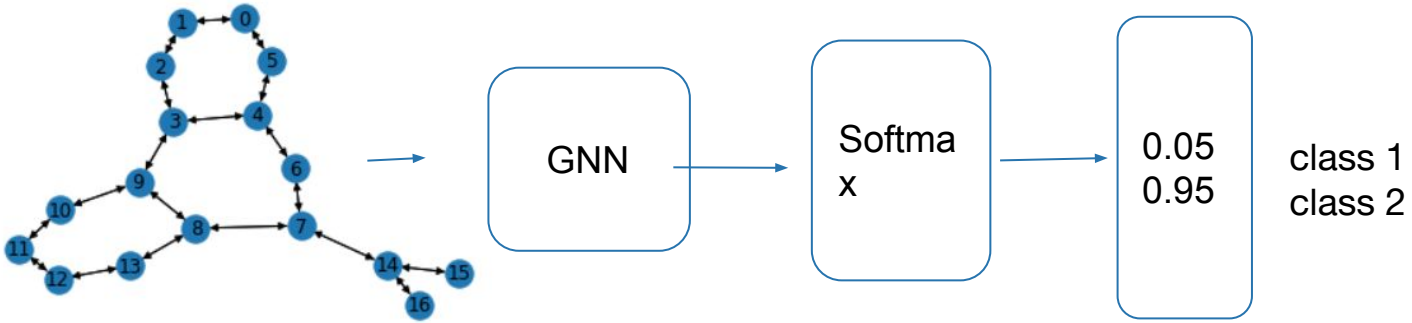




Experiment Setting

Use the value of the softmax result of that class to measure confidence.



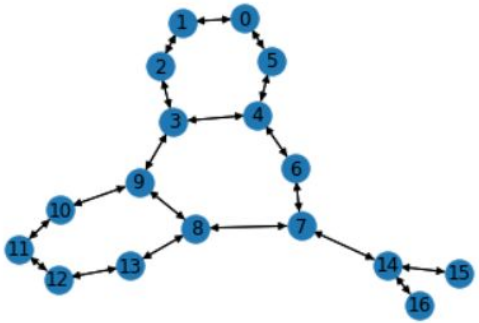


Experiment Setting

We aim to find the minimal sufficient explanations for the full graphs.

Sub-graph methods, BFS with Maximum Depth.

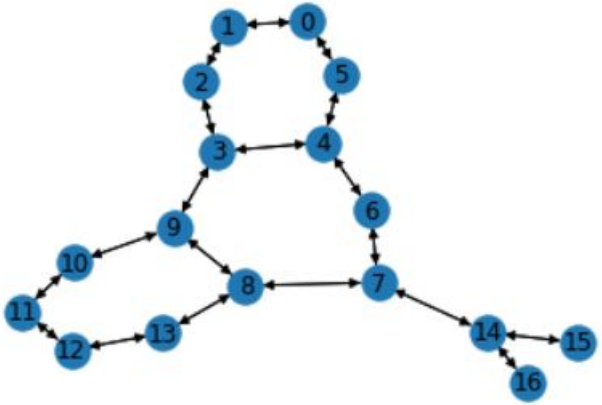
For pre-training GCN, we applied a simple GCN with three convolutional layers. The pretrained GCN has an overall prediction accuracy of 59%



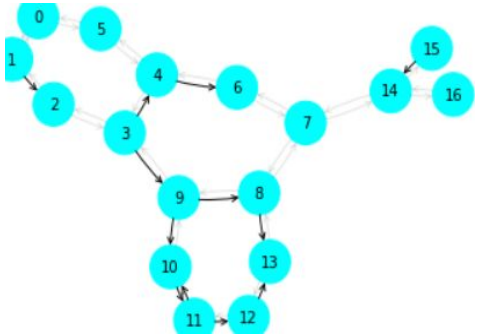
BFS with max_depth:
source node 3 max_depth 1: 2, 3, 4, 9
source node 4 max_depth 2: 0, 5, 4, 3, 9, 6, 7

Results & Analysis - Mutag

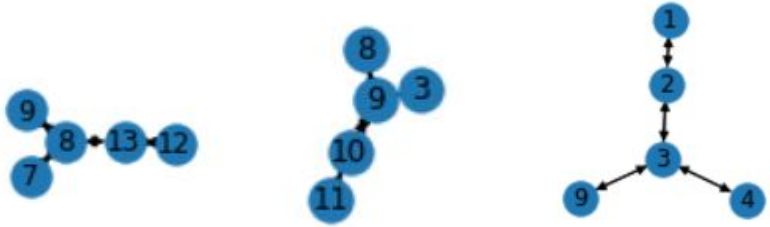
Original



GNN Explainer



Our Explanation

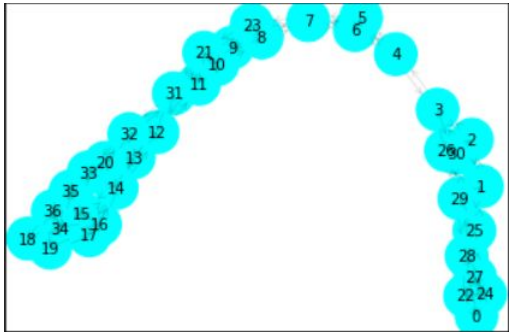


1 **2** **3**

Results & Analysis - Enzymes

Original

GNN Explainer



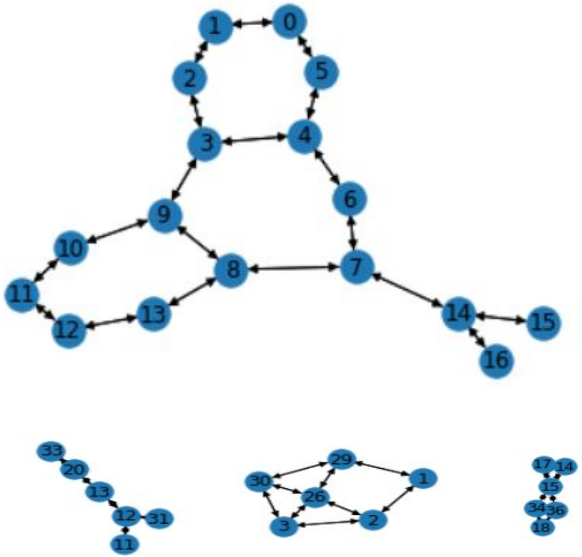
Our Explanation

1 2 3

Conclusion

Multiple possible subgraphs provide additional context to graph neural network predictions.

Multiple smaller subgraphs are more readable, allowing better insight into large graph predictions.



Future Works

- Structured Graph Trees, exploring subgraphs of candidate explorations
- Smarter sub-graph generation, especially for larger baseline graphs
- Exploration of lager graph sizes
- User study on usability of different explainability methods
- Application to other tasks

Q&A