Towards Multi-Grained Explainability for Graph Neural Networks

Xiang Wang, Yingxin Wu, An Zhang, Xiangnan He, Tat-Seng Chua

https://github.com/Wuyxin/ReFine.
**Graph Data is Ubiquitous**

Protein Structure  
COVID Graph  
Knowledge Graph  
Social Network

**Tasks of Graph Learning:**
- Node Classification
- Graph Classification
- Link Prediction ...

https://github.com/Wuyxin/ReFine.
Graph Neural Networks (GNNs)

**Core of GNNs**

- Graph Structure Guides Representation Learning
- Information Propagation & Aggregation

**Input Graph** $G$

**GNN Model** $f$

**Output Prediction** $\hat{y}$

**Explainability of GNNs**

Which fraction of the input graph is most influential to the model’s decision?

**Input Graph** $G$

**Output Prediction** $\hat{y}$

**Explanatory Subgraph** $G_s$

https://github.com/Wuyxin/ReFine.
**Local & Global Explainability**

**Case: Scene graph classification**

**Question:** Which fraction of the input graph is most influential to the model’s decision?

<table>
<thead>
<tr>
<th><strong>Input Scene Graph</strong></th>
<th><strong>Output Prediction</strong> ( \hat{y} : \text{Farm} )</th>
<th><strong>Explanation 1</strong></th>
<th><strong>Explanation 2</strong></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Related works</th>
<th>Potential drawback</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Local Explainability:</strong></td>
<td>GNNExplainer [Ying et al. 2019] PGM-Explainer [Vu et al. 2020]</td>
</tr>
<tr>
<td>Interprets each instance independently.</td>
<td></td>
</tr>
<tr>
<td><strong>Global Explainability:</strong></td>
<td>PGExplainer [Luo et al. 2020] XGNN [Yuan et al. 2020]</td>
</tr>
<tr>
<td>Systematizes the globally important patterns.</td>
<td></td>
</tr>
</tbody>
</table>

https://github.com/Wuyxin/ReFine.
Multi-Grained Explainability

What class-wise knowledge does the GNN leverage to make predictions in general?

Why the GNN model made the certain prediction for the instance at hand?

Before fine-tuning

After fine-tuning

Probability of being Farm

(a) 1.0
(b) 0.96
(c) 0.9
(d) 0.41
(e) 0.95

Pre-training towards Global Explainability

Fine-Tuning towards Local Explainability

https://github.com/Wuyxin/ReFine.
Pre-training & Fine-tuning (ReFine)

\[ \mathcal{L}_1 = MI(Y, M \oplus G_{\text{att}}) \]
Negative mutual information

\[ \min_{\theta} \mathcal{L}_1 + \gamma \mathcal{L}_{\text{cts}} \]
Contrastive Loss

\[ \theta_0' = \theta \]
User-defined ratio

\[ \min_{\theta'} \mathcal{L}_2 = MI(Y, G_{\text{exp}}) \]
where \( G_{\text{exp}} = \text{Top}_p(G_{\text{att}}) \)

https://github.com/Wuyxin/ReFine.
## Empirical Results of ReFine

Table 1: Comparison of our ReFine and other baseline explainers

<table>
<thead>
<tr>
<th></th>
<th>Mutagenicity</th>
<th>VG-5</th>
<th>MNIST</th>
<th>BA-3motif</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC-AUC</td>
<td>ACC-AUC</td>
<td>ACC-AUC</td>
<td>ACC-AUC</td>
</tr>
<tr>
<td>SA[9]</td>
<td>0.769</td>
<td>0.769</td>
<td>0.559</td>
<td>0.518</td>
</tr>
<tr>
<td>GNNExplainer[6]</td>
<td>0.895±0.010</td>
<td>0.895±0.003</td>
<td>0.535±0.013</td>
<td>0.528±0.005</td>
</tr>
<tr>
<td>PG-Explainer[7]</td>
<td>0.631±0.008</td>
<td>0.790±0.004</td>
<td>0.504±0.010</td>
<td>0.586±0.004</td>
</tr>
<tr>
<td>PGM-Explainer[19]</td>
<td>0.714±0.007</td>
<td>0.792±0.001</td>
<td>0.615±0.003</td>
<td>0.575±0.002</td>
</tr>
<tr>
<td>ReFine-CT</td>
<td>0.888±0.008</td>
<td>0.891±0.002</td>
<td>0.526±0.007</td>
<td>0.610±0.004</td>
</tr>
<tr>
<td>ReFine-FT</td>
<td>0.945±0.011</td>
<td>0.906±0.002</td>
<td>0.587±0.008</td>
<td>0.616±0.003</td>
</tr>
<tr>
<td>ReFine</td>
<td><strong>0.955±0.005</strong></td>
<td><strong>0.914±0.001</strong></td>
<td><strong>0.636±0.003</strong></td>
<td><strong>0.630±0.006</strong></td>
</tr>
<tr>
<td>Improvement</td>
<td>6.7%</td>
<td>2.1%</td>
<td>3.4%</td>
<td>7.5%</td>
</tr>
</tbody>
</table>

Table 2: Performance under different selection ratios before and after fine-tuning.

<table>
<thead>
<tr>
<th></th>
<th>Mutagenicity</th>
<th>VG-5</th>
<th>MNIST</th>
<th>BA-3motif</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ACC@ρ</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.4 0.6</td>
<td>0.4 0.6</td>
<td>0.4 0.6</td>
<td>0.4 0.6</td>
</tr>
<tr>
<td>ReFine-FT</td>
<td>96.8% 94.0%</td>
<td>91.3% 91.4%</td>
<td>41.4%</td>
<td>61.4%</td>
</tr>
<tr>
<td>ReFine</td>
<td>97.8% 96.2%</td>
<td>92.2% 93.4%</td>
<td>71.4%</td>
<td>82.0%</td>
</tr>
<tr>
<td>Improvement</td>
<td>+1.0% +2.2%</td>
<td>+0.9% +2.0%</td>
<td>+30.0%</td>
<td>+20.6%</td>
</tr>
</tbody>
</table>

https://github.com/Wuyxin/ReFine.
Empirical Results of ReFine

Selection Ratio

Pre-trained

Fine-tuned

ACC: 0.543 0.888 0.976

https://github.com/Wuyxin/ReFine.
Summary

• Local explainability & Global explainability present different views of GNN models.
• Multi-grained explainability can offer more reliable & faithful explanations.

Check out our code and models at
• https://github.com/Wuyxin/ReFine.

THANK YOU!