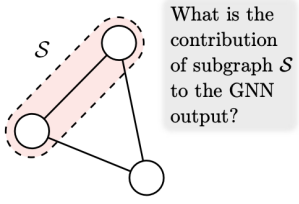
Efficient Higher-Order Subgraph Attribution via Message Passing

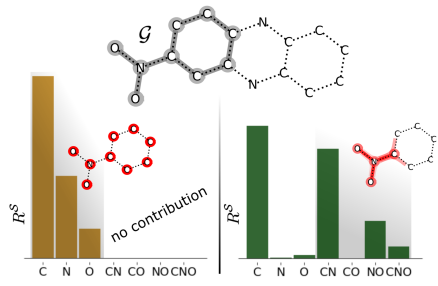
Ping Xiong, Thomas Schnake, Grégoire Montavon, Klaus-Robert Müller, Shinichi Nakajima

ICML 2022

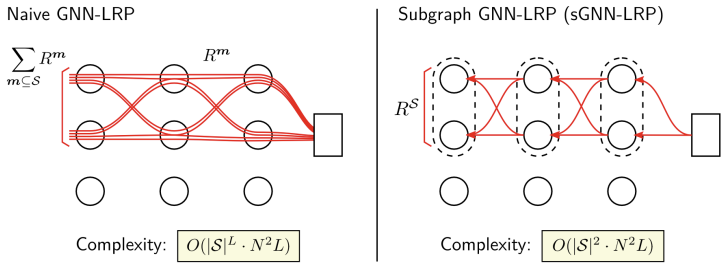
XAI on GNN

- Explain GNN at the level of node, edge, walk, subgraph [5].
- GNN-LRP [1, 4]
  - Walk-level (high-order) explanation.
  - Decomposition and layer-wise backpropagation.
  - Better performance than lower-order methods.





Naive GNN-LRP vs. sGNN-LRP



- Relevance propagation as Markov chain: $R^{\mathbf{m},\mathbf{n}} = \left(\prod_{l=0}^{L-1} T^{l,m_l,m_{l+1}}_{n_l,n_{l+1}}\right) r^{L,m_L}_{n_L} = p(\mathbf{m},\mathbf{n})$.
- $R^{\mathbf{m}} = \sum_{\mathbf{n}} \left(\prod_{l=0}^{L-1} T^{l,m_l,m_{l+1}}_{n_l,n_{l+1}}\right) r^{L,m_L}_{n_L} = p(\mathbf{m})$,
  $R^{\mathcal{S}} = \sum_{\mathbf{m} \subseteq \mathcal{S}} \sum_{\mathbf{n}} \left(\prod_{l=0}^{L-1} T^{l,m_l,m_{l+1}}_{n_l,n_{l+1}}\right) r^{L,m_L}_{n_L} = p(\mathbf{m} \subseteq \mathcal{S})$.
- **Sum-product** algorithm [2] & Forward-hook trick [3, 4].

Computational Efficiency Evaluation

Dataset	Model	Naive	sGNN-LRP
BA-2motif	GIN-3	224.22	**4.22**
	GIN-5	6.07×10^3	**6.44**
	GIN-7	1.42×10^5	**9.81**
MUTAG	GIN-3	4.23×10^3	**28.90**
Mutagenicity	GIN-3	4.28×10^3	**26.68**
REDDIT-B	GIN-5	—	**195.43**
Graph-SST2	GCN-3	3.16×10^5	**29.94**

Table: Computation time (in msec) comparison. '—' means 'failed'.The subgraph size is $|\mathcal{S}| = 5$.

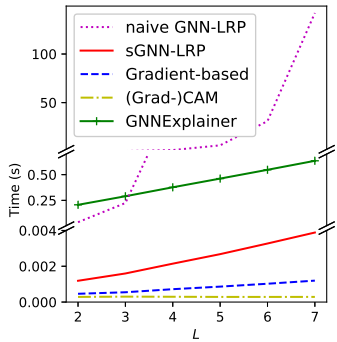


Figure: Computation time on BA-2motif dataset as a function of model depth. GIN-L for $L = 2, \ldots, 7$ with $|\mathcal{S}| = 5$.

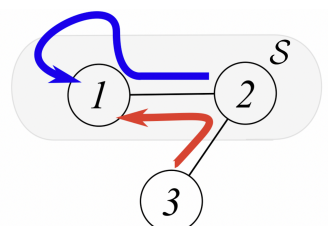
Generalized Subgraph Relevance Definition

$\mathcal{S}$ is important iff

1. the model predictions for $\mathcal{G}$ and $\mathcal{S}$ are **almost same**, and
2. the model predictions for $\mathcal{G} \setminus \mathcal{S}$ and $\mathcal{G}$ **diverge drastically**.

Generalized subgraph relevance: $R_\alpha^{\mathcal{S}} = \sum_{\mathbf{m} \in \mathcal{G}} g_\alpha^{\mathcal{S}}(\mathbf{m}) R^{\mathbf{m}}$,

$$g_\alpha^{\mathcal{S}}(\mathbf{m}) = \begin{cases} 0 & \text{if } m_l \notin \mathcal{S},\ \forall\ l = 0, \ldots, L, \\ \alpha^{\sum_{l=0}^{L} \mathbf{1}(m_l \notin \mathcal{S})} & \text{otherwise.} \end{cases}$$



Evaluation of Generalized Subgraph Attribution

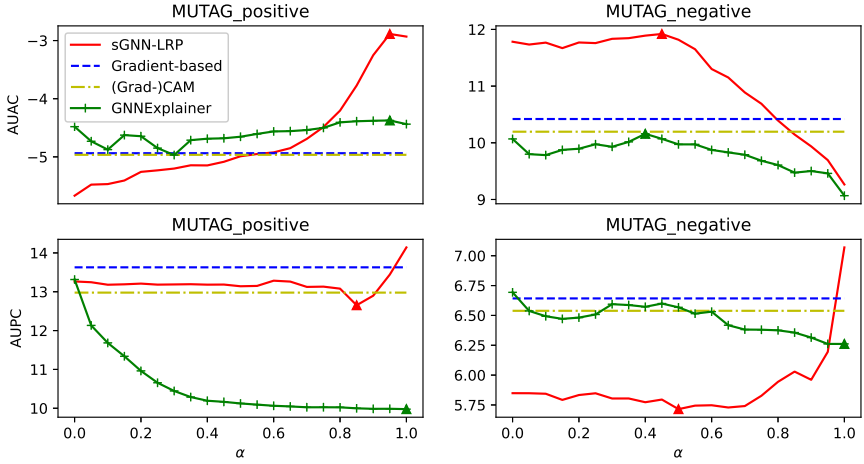


Figure: The optima are marked with ▲. AUAC↑ and AUPC↓ is better.

References

[1] Sebastian Bach et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation". In: *PloS one* 10.7 (2015), e0130140.

[2] Christopher M. Bishop. "Pattern Recognition and Machine Learning (Information Science and Statistics)". In: Berlin, Heidelberg: Springer-Verlag, 2006, pp. 402–411. ISBN: 0387310738.

[3] Wojciech Samek et al. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications". In: *Proc. IEEE* 109.3 (2021), pp. 247–278.

[4] Thomas Schnake et al. "Higher-Order Explanations of Graph Neural Networks via Relevant Walks". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021), pp. 1–1.

[5] Hao Yuan et al. "Explainability in Graph Neural Networks: A Taxonomic Survey". In: *CoRR* abs/2012.15445 (2020). arXiv: 2012.15445. URL: https://arxiv.org/abs/2012.15445.