Faithful and Accurate Self-Attention Attribution for Message Passing Neural Networks via the Computation Tree Viewpoint

Yong-Min Shin[1], Siqing Li[2], Xin Cao[2], Won-Yong Shin[1]
[1]Yonsei University, [2]University of New South Wales

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

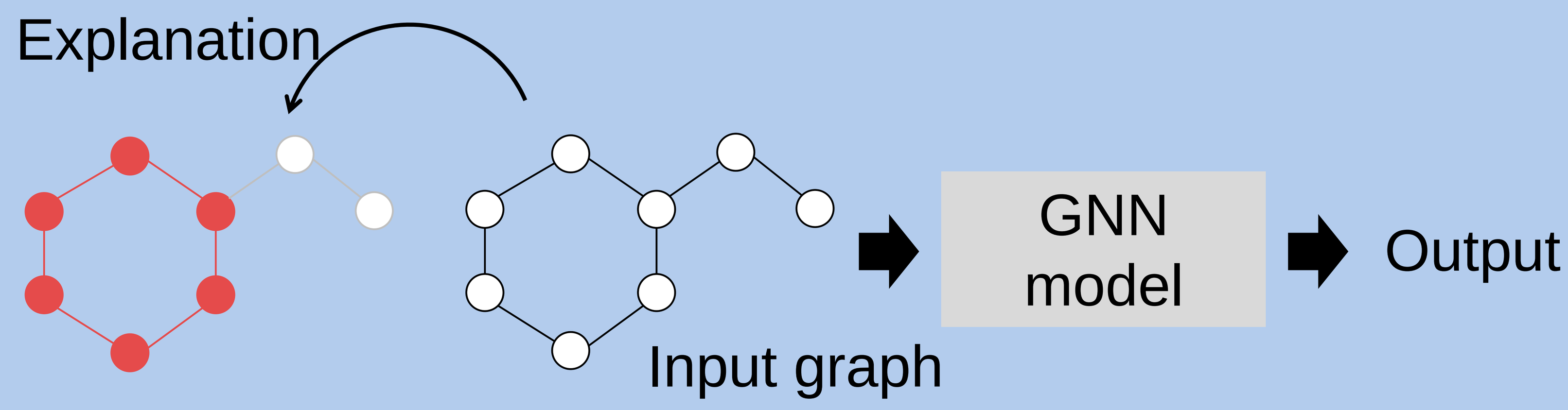
1. Vast discussion on Attention in CV & NLP

- Attention is already used as explanations of transformer models in <u>computer vision</u>: Rollout [1], Chefer et al. [2 & 3]
- A long discussion of attention-based interpretation in <u>natural language processing</u> (see [4] for full survey)

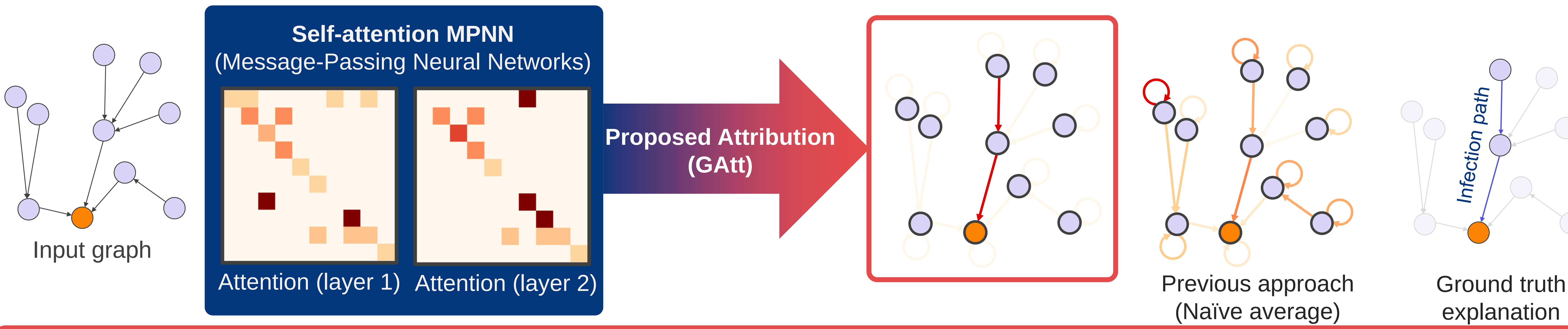
Gap!

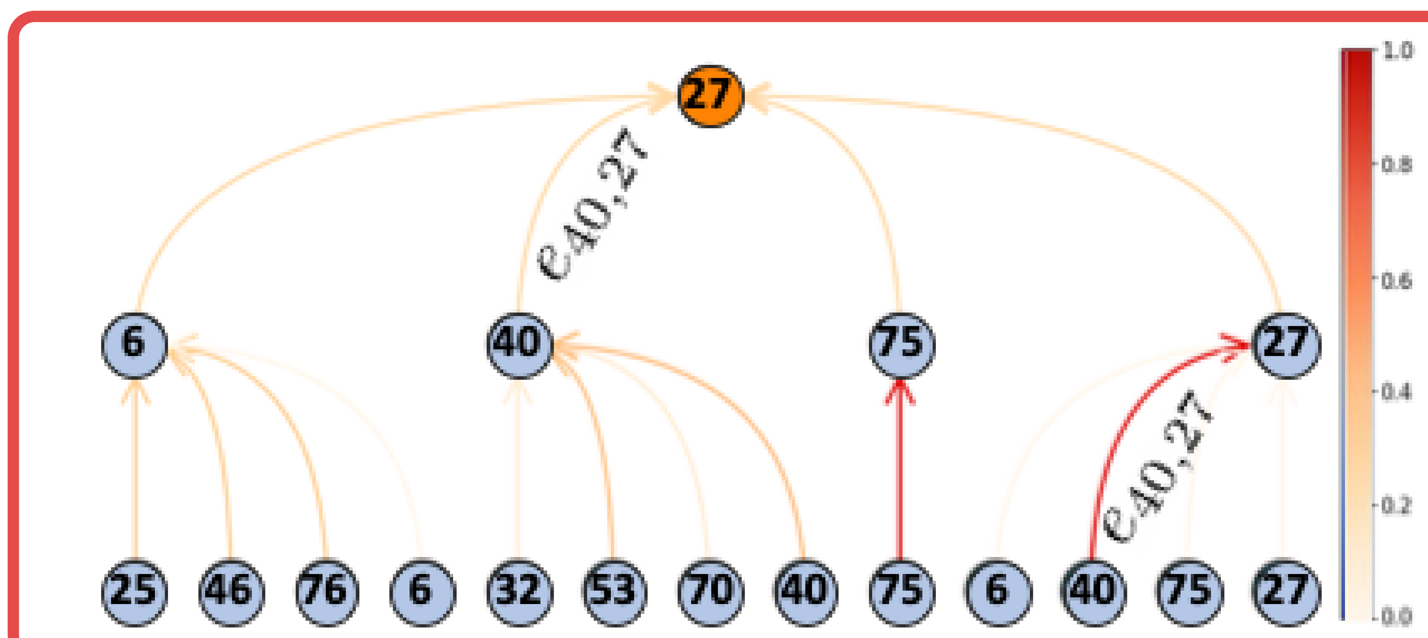
2. Underdeveloped topic in explaining GNNs

- XAI in GNNs is actively studied in recent years. Explanation methods usually aim to find the most relevant part of the input (attribution)
- However, there has been <u>nearly no discussion</u> on the potential of attention as explanation in GNNs



Our work is <u>one of the first</u> to seriously discuss the **potential of attention weights** in **MPNNs** as explanations.





Design Philosophy of GAtt:
<u>Computation Tree Viewpoint</u> <u>of Edge Attribution Calculation from Attention</u>

1) **Proximity effect:** Edges within closer proximity to the target node tend to highly impact the model's prediction compared with distant edges, since they are likely to appear more frequently in the computation tree. **Need to sum all occurances of an edge!**
2) **Contribution adjustment:** The contribution of an edge in the computation tree should be **adjusted by its position (i.e., other edges in the path towards the root)**.

Faithfulness evaluation

Dataset		2-layer GAT/GATv2			3-layer GAT/GATv2		
		GAtt	AvgAtt	Random	GAtt	AvgAtt	Random
Cora	Δ_{PC}	**0.8468/0.1040**	0.1764/0.0121	-0.0056/-0.0036	**0.8642/0.1696**	0.0967/0.0168	0.0045/0.0045
	Δ_{NE}	**0.7112/0.0930**	0.1526/0.0100	-0.0076/0.0019	**0.7690/0.1664**	0.0859/0.0186	0.0040/0.0037
	Δ_{P}	**0.9755/0.9623**	0.7251/0.6226	0.4389/0.4891	**0.9875/0.9966**	0.7075/0.8897	0.5235/0.6107

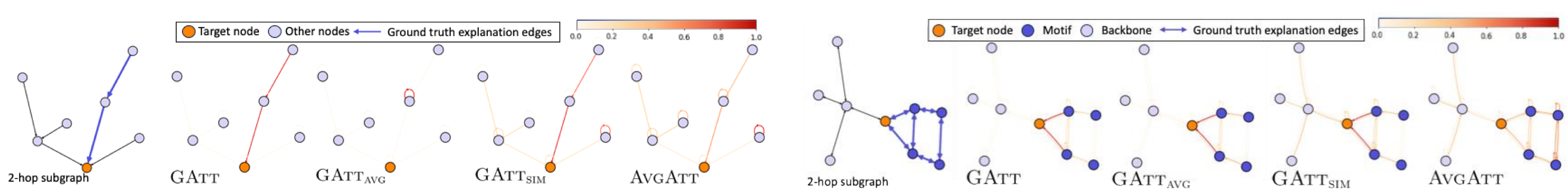
- <u>Faithfulness</u>: How much does the attribution actually reflect the behavior of the underlying model?
- Results (7 datasets, 3 measurements, 2/3 layer **GAT [5] / GATv2 [6] / SuperGAT [7]**) show clear superiority.

Accuracy evaluation

Model	Dataset	GAtt	AvgAtt	SA	GB	IG	GNNEx	PGEx	GM	FDnX	Random
GAT	BA-Shapes	<u>0.9591</u>	0.7977	0.9563	0.6231	0.6231	0.8916	0.8289	0.5316	**0.9917**	0.4975
	Infection	**0.9976**	0.8786	0.8237	0.8949	<u>0.9472</u>	0.9272	0.7173	0.6859	0.6574	0.4811

- <u>Accuracy</u>: How much does the attribution reveal the ground truth explanation?
- Results (2 datasets, 1 attention baseline, **7 post-hoc explanation baselines**) show clear competent performance.

Visualization & Ablation



GAtt shows <u>superior empirical performance</u> on a wide variety of datasets, measures, and attention-based MPNN models.

[1] Abnar and Zuidema, "Quantifying Attention Flow in Transformers", ACL 2020
[2] Chefer et al., "Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers, ICCV 2021
[3] Chefer et al., "Transformer Interpretability Beyond Attention Visualization, CVPR 2021
[4] Bibal et al., "Is Attention Explanation? An Introduction to the Debate", ACL 2022
[5] Velickovic et al., "Graph Attention Networks". ICLR 2018
[6] Brody et al., How Attentive are Graph Attention Networks? ICLR 2022
[7] Kim & Oh, How to Find Your Friendly Neighborhood: Graph Attention Design with Self-Supervision. ICLR 2021

arXiv preprint

Homepage