Less is More: on the Over-Globalizing Problem in Graph Transformers

Yujie Xing[1], Xiao Wang[2 †], Yibo Li[1], Hai Huang[1], Chuan Shi[1 †]

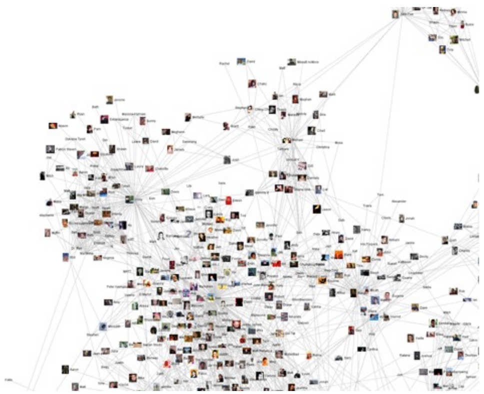
[1] Department of Computer Science, Beijing University of Posts and Telecommunication, Beijing, China.
[2] School of Software, Beihang University, Beijing, China.

ICML 2024 Oral

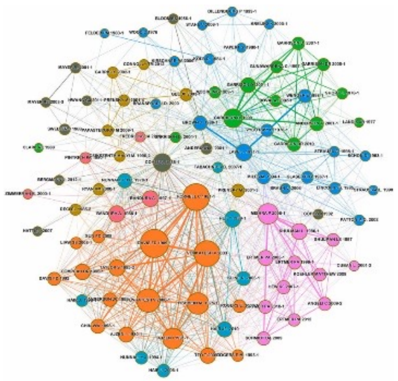
CONTENTS

Graph Data

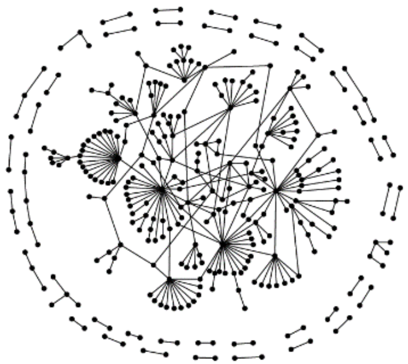
Graph-structured data, an essential and prevalent form in the real world, plays a vital role in modeling object interactions



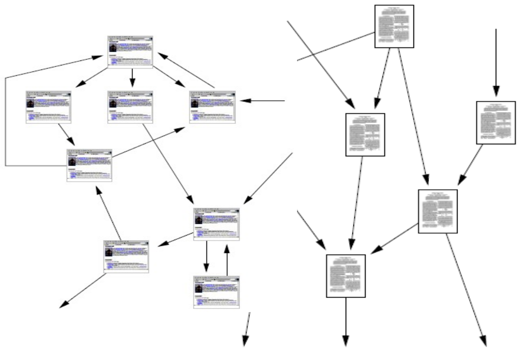
Social networks



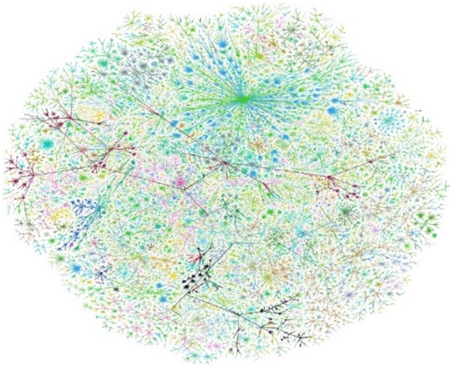
Citation networks



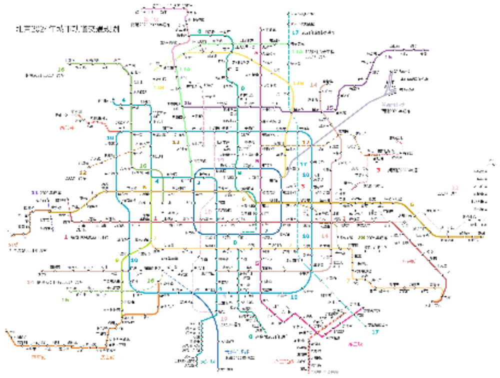
Biomedical networks



Information networks



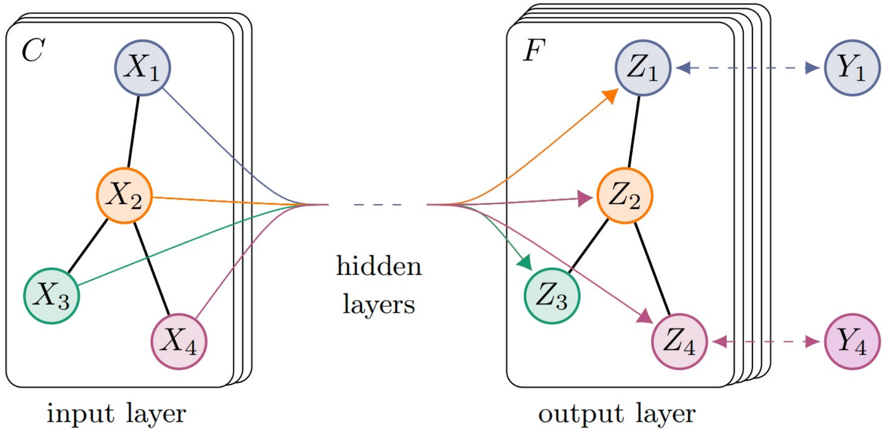
Internet



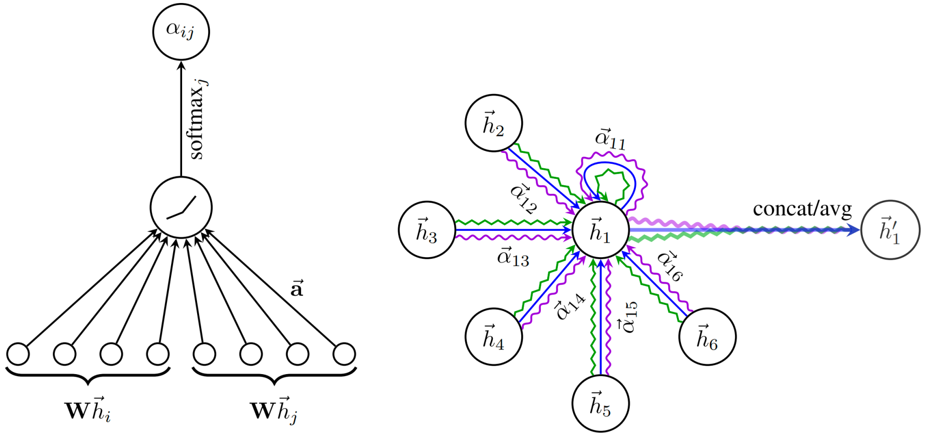
Transport networks

Graph Neural Networks (GNNs)

GNNs effectively utilize their message-passing mechanism to extract useful information and learn high-quality representations from graph data.



Graph Convolutional Networks
(GCN, ICLR 2017)

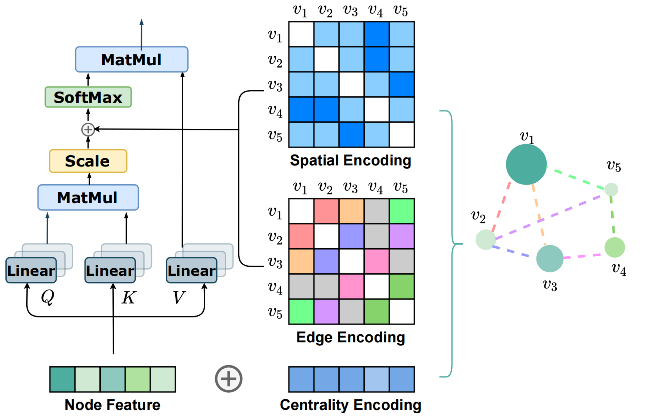


Graph Attention Networks
(GAT, ICLR 2018)

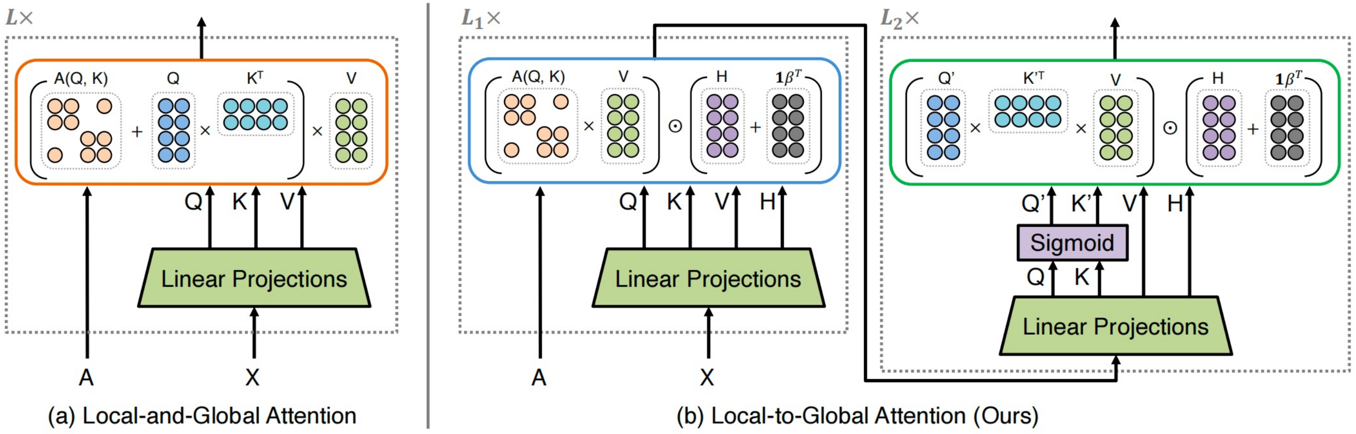
Unable to stack multi-layers due to over-smoothing and over-squashing, resulting in limited receptive fields to near neighbors!!!

Kipf, T. N. and Welling, M.. Semi-supervised classification with graph convolutional networks. ICLR 2017.
Veličković, P., Cucurull, G., et al. Graph attention networks. ICLR 2018.

Graph Transformers construct a fully connected graph and adaptively learn interaction relationships with the powerful global attention mechanism.



Graphormer
(NeurIPS 2021)



Polynormer
(ICLR 2024)

Graph Transformers have achieved remarkable success in graph-level tasks and node-level tasks.

Ying, C., Cai, T., et al. Do transformers really perform badly for graph representation? NeurIPS 2021.
Wu, Q., Zhao, W., et al. Nodeformer: A scalable graph structure learning transformer for node classification. NeurIPS 2022.

Motivation && Contribution

- It is well recognized that the global attention mechanism considers a wider receptive field in a fully connected graph, leading many to believe that useful information can be extracted from all the nodes.

- A key question arises:

Does the globalizing property always benefit Graph Transformers?

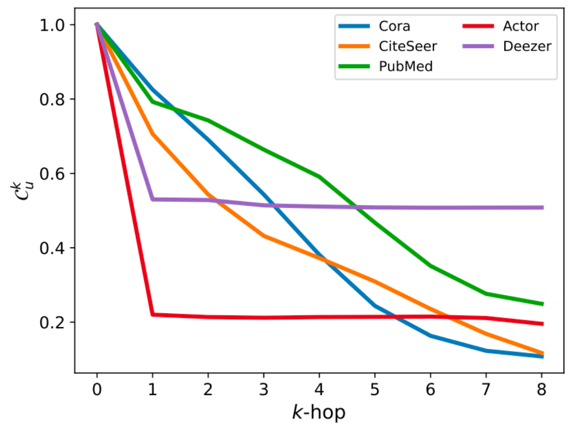
- We reveal the ***over-globalizing problem*** in Graph Transformers by presenting both empirical evidence and theoretical analysis

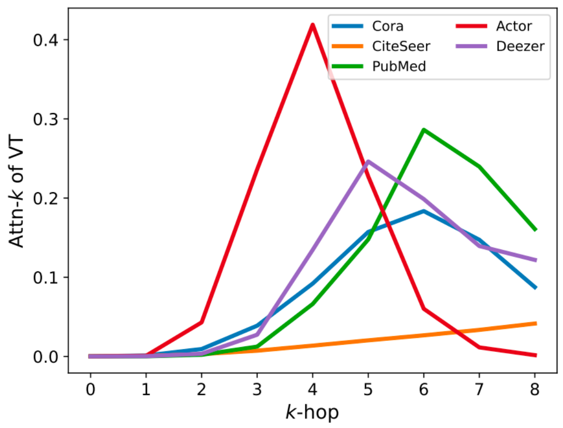
- We propose a novel Bi-Level Global Graph Transformer with Collaborative Training (CoBFormer), to alleviate the over-globalizing problem while keeping the ability to extract valuable information from distant nodes.

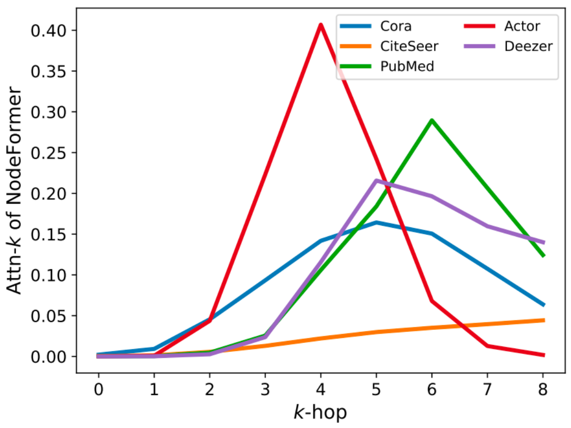
CONTENTS

2 Over-Globalizing Problem Empirical Evidence

We empirically find the ***over-globalizing problem*** *in Graph Transformers.*







$$C_u^k = \frac{|v \in \mathcal{N}^k(u) : \mathbf{y}_u = \mathbf{y}_v|}{|\mathcal{N}^k(u)|},$$

The proportion of the k-th hop neighbors sharing the same label with node u

$$\text{Attn-}k = \mathbb{E}_{u \in \mathcal{V}} \sum_{v \in \mathcal{N}^k(u)} \alpha_{uv}.$$

The average attention scores allocated to the k-th hop neighbors

Near nodes usually contain more useful information

Inconsistent

Transformers overly focuses on those distant nodes

Theorem 3.1. *For a given node u and a well-trained Graph Transformer, let* $\eta_u = \mathbb{E}_{v\in\mathcal{V},\mathbf{y}_u=\mathbf{y}_v}\exp(\frac{\mathbf{q}_u\mathbf{k}_v^T}{\sqrt{d}})$, $\gamma_u = \mathbb{E}_{v\in\mathcal{V},\mathbf{y}_u\neq\mathbf{y}_v}\exp(\frac{\mathbf{q}_u\mathbf{k}_v^T}{\sqrt{d}})$. *Then, we have:*

$$\begin{aligned}\|\mathbf{Z}-\hat{\mathbf{A}}\mathbf{Z}\|_F &\leq \sqrt{2}L\sum_{u\in\mathcal{V}}\sum_{v\in\mathcal{V},\mathbf{y}_u\neq\mathbf{y}_v}\alpha_{uv} \\ &= \sqrt{2}L\sum_{u\in\mathcal{V}}\frac{1}{1+\frac{\mathcal{C}_u}{1-\mathcal{C}_u}\frac{\eta_u}{\gamma_u}}.\end{aligned} \tag{5}$$

where L is a Lipschitz constant.

Theorem 3.2. *To analyze the impact of k on $\mathcal{C}_u^k$, we assume that each node has an equal probability $\frac{1}{|\mathcal{Y}|}$ of belonging to any given class. Given the edge homophily $\rho = \frac{|(u,v)\in\mathcal{E}:\mathbf{y}_u=\mathbf{y}_v|}{|\mathcal{E}|}$, $\mathcal{C}_u^k$ can be recursively defined as:*

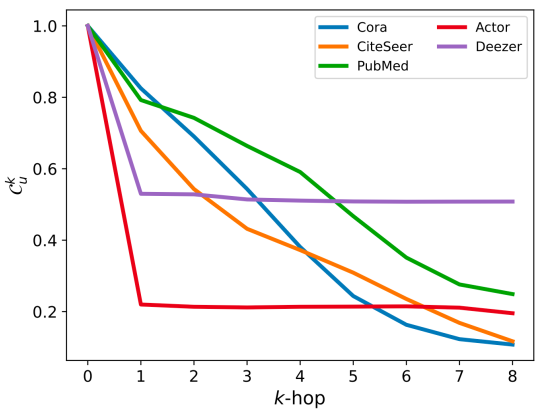
$$\mathcal{C}_u^k = \begin{cases} 1, & if\ k=0 \\ \rho, & if\ k=1 \\ \frac{1+|\mathcal{Y}|\rho\mathcal{C}_u^{k-1}-\rho-\mathcal{C}_u^{k-1}}{|\mathcal{Y}|-1}. & if\ k=2,3,\cdots \end{cases} \tag{6}$$

And $\mathcal{C}_u^k$ possesses the following properties:

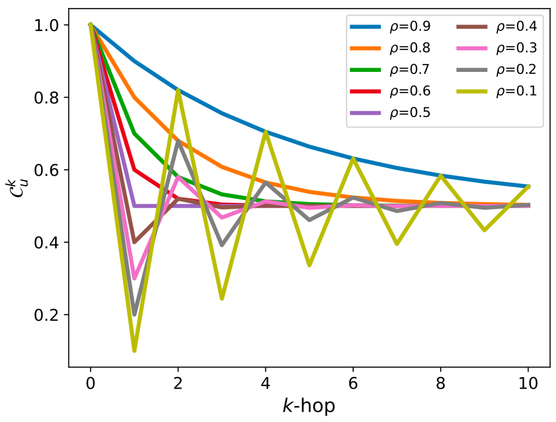
$$\begin{cases} \mathcal{C}_u^\infty = \frac{1}{|\mathcal{Y}|} & \\ \mathcal{C}_u^k \geq \mathcal{C}_u^{k+1}, & if\ \rho\geq\frac{1}{|\mathcal{Y}|},\ k=0,1\cdots \\ \mathcal{C}_u^{2k} > \mathcal{C}_u^{2(k+1)}, & if\ \rho<\frac{1}{|\mathcal{Y}|},\ k=0,1\cdots \\ \mathcal{C}_u^{2k+1} < \mathcal{C}_u^{2(k+1)+1}, & if\ \rho<\frac{1}{|\mathcal{Y}|}.\ k=0,1\cdots \end{cases} \tag{7}$$

Theoretical Analysis: An over-expanded receptive field may adversely affect the global attention due to the ***over-globalizing problem***.

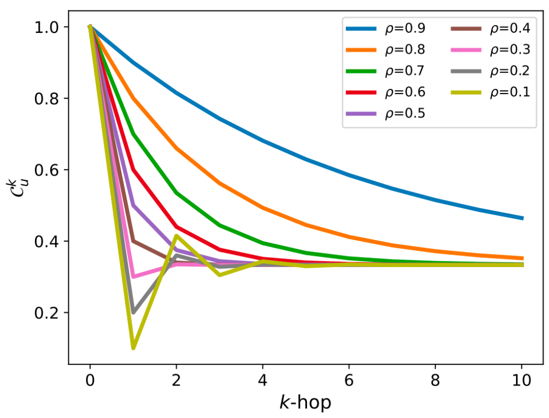
Our theorem aligns well with the real-world scenarios



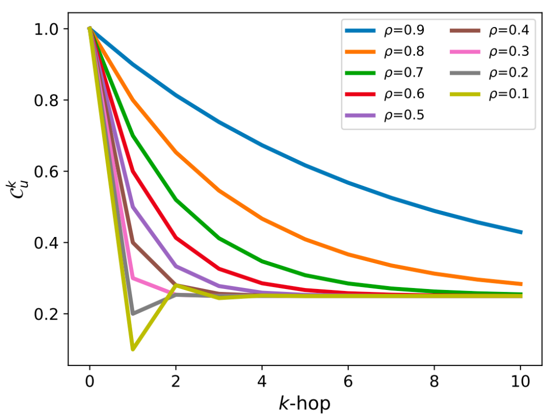
Real-World



$|\mathcal{Y}| = 2$



$|\mathcal{Y}| = 3$



$|\mathcal{Y}| = 4$

Experimental Analysis

Inspired by Theorem 3.1, we define the Attention Signal/Noise Ratio (Attn-SNR) as the metric to quantify the ability of Graph Transformers to distinguish useful nodes as follows:

$$\text{Attn-SNR} = 10 \lg \left(\frac{\sum_{\mathbf{y}_u = \mathbf{y}_v} \alpha_{uv}}{\sum_{\mathbf{y}_u \neq \mathbf{y}_v} \alpha_{uv}} \right).$$

We evaluate the following models using Attn-SNR and Accuracy:

- *VT: Vanilla Transformer*
- *NF: NodeFormer*
- *VT-D: VT but double the attention scores between nodes sharing the same label*

Table 1. The Attn-SNR and testing accuracy of different models.

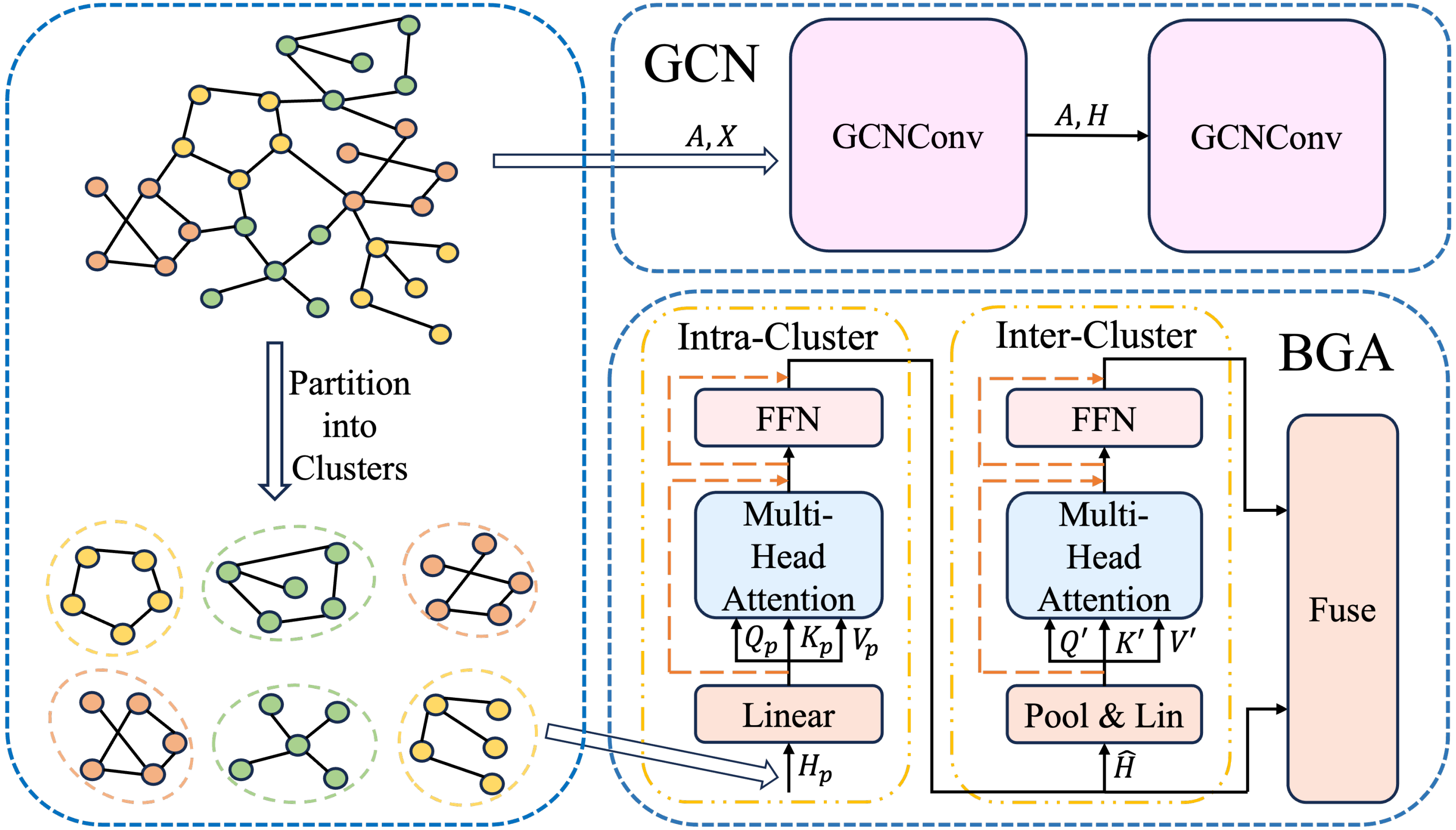
Dataset	Metric	VT	NF	VT-D
Cora	Attn-SNR	-6.97	0.43	12.05
	Accuracy	55.18	80.20	82.12
CiteSeer	Attn-SNR	-7.19	-5.09	8.72
	Accuracy	50.72	71.50	61.80

Experimental analysis:
Solving the ***over-globalizing problem*** can improve the performance of Graph Transformers.

CONTENTS

CoBFormer

We propose a novel Bi-Level Global Graph Transformer with Collaborative Training (CoBFormer).



GCN Module

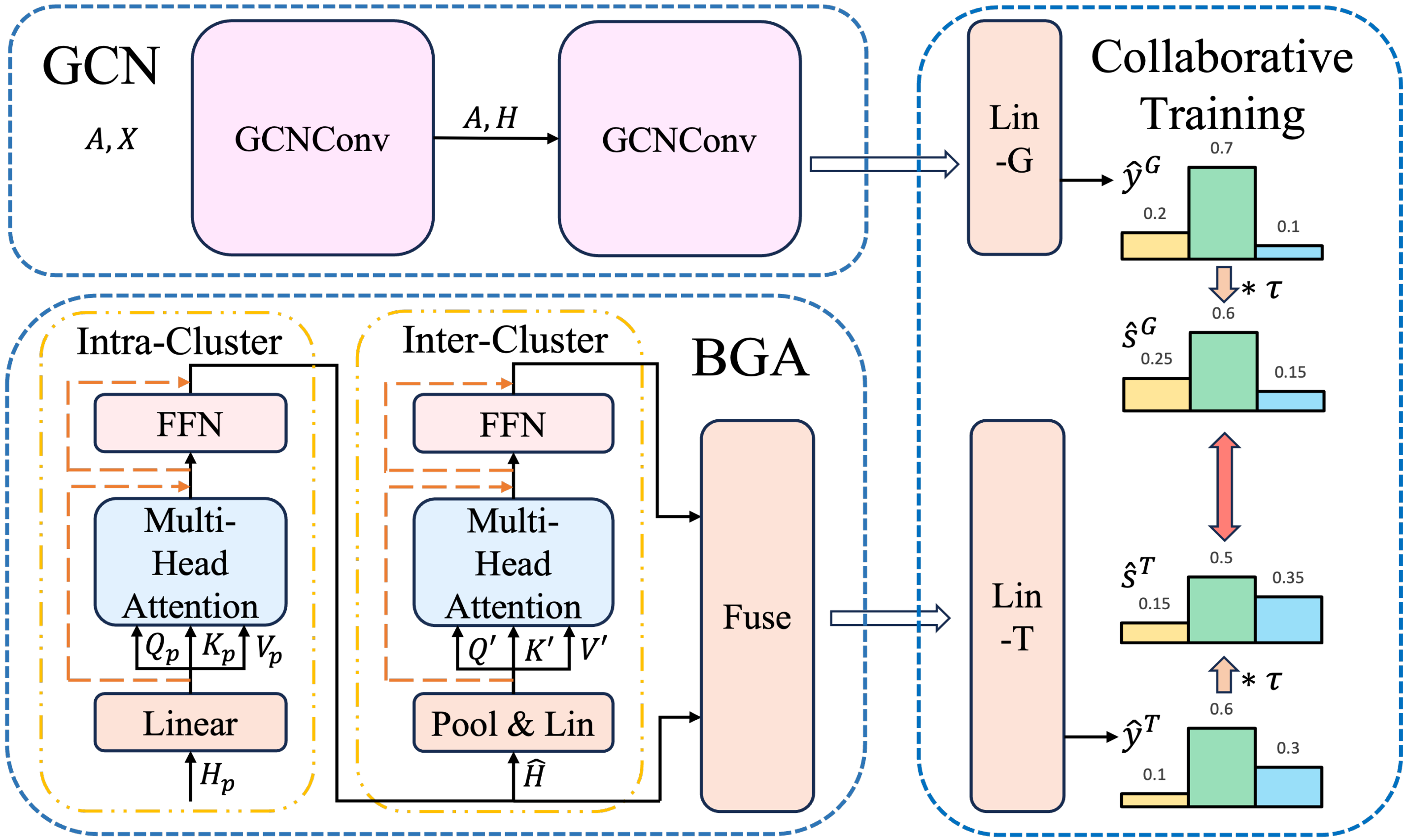
As the local module to supplement the graph structure information ignored by the BGA module

BGA Module

Decouple the information within intra-clusters and between inter-clusters by an intra-cluster Transformer and an inter-cluster Transformer.

3 Method CoBFormer

We propose a novel Bi-Level Global Graph Transformer with Collaborative Training (CoBFormer).



$$
\begin{aligned}
\hat{\mathbf{Z}}^G &= \text{Lin-G}(\text{GCN}(\mathbf{A}, \mathbf{X})),\\
\hat{\mathbf{Z}}^T &= \text{Lin-T}(\text{BGA}(\mathbf{X}, \mathcal{P})).
\end{aligned}
$$

$$\hat{\mathbf{Y}}^G = \text{SoftMax}(\hat{\mathbf{Z}}^G),\ \hat{\mathbf{Y}}^T = \text{SoftMax}(\hat{\mathbf{Z}}^T),$$

$$\hat{\mathbf{S}}^G = \text{SoftMax}(\hat{\mathbf{Z}}^G * \tau),\ \hat{\mathbf{S}}^T = \text{SoftMax}(\hat{\mathbf{Z}}^T * \tau),$$

$$\mathcal{L}_{ce} = -\left(\mathbb{E}_{\mathbf{y}_u, u\in\mathcal{V}_\mathcal{L}} \log(\hat{\mathbf{y}}_u^G) + \mathbb{E}_{\mathbf{y}_u, u\in\mathcal{V}_\mathcal{L}} \log(\hat{\mathbf{y}}_u^T)\right),$$

$$\mathcal{L}_{co} = -\left(\mathbb{E}_{\hat{\mathbf{s}}_u^G, u\in\mathcal{V}_\mathcal{U}} \log(\hat{\mathbf{s}}_u^T) + \mathbb{E}_{\hat{\mathbf{s}}_u^T, u\in\mathcal{V}_\mathcal{U}} \log(\hat{\mathbf{s}}_u^G)\right),$$

$$\mathcal{L} = \alpha * \mathcal{L}_{ce} + (1-\alpha) * \mathcal{L}_{co}.$$

Collaborative Training

Encourage mutual learning between the GCN and BGA module, thus improving their performance.

Theoretical Guarantees

Proposition 4.1. *Given $u \in \mathcal{V}_p, v \in \mathcal{V}_q$, along with a well-trained inter-cluster attention score matrix $\dot{\mathbf{A}} \in \mathbb{R}^{P \times P}$. Let $\dot{\alpha}_{pq}$ represent the attention score between clusters p and q. Then the approximate attention score between node u and v can be expressed as $\hat{\alpha}_{uv} = \frac{\dot{\alpha}_{pq}}{|\mathcal{V}_q|}$.*

Theorem 4.2. *Consider $P(\mathbf{L}, \mathbf{U})$ as the true label distribution, $P_G(\mathbf{L}, \mathbf{U})$ as the predicted label distribution by the GCN, and $P_T(\mathbf{L}, \mathbf{U})$ as the predicted label distribution by the BGA module. The following relations hold:*

$$
\begin{aligned}
\mathbb{E}_{P(\mathbf{L},\mathbf{U})} \log P_G(\mathbf{L}, \mathbf{U}) = & \mathbb{E}_{P(\mathbf{L})} \log P_G(\mathbf{L}) + \\
& \mathbb{E}_{P_T(\mathbf{U}|\mathbf{L})} \log P_G(\mathbf{U}|\mathbf{L}) - \\
& \mathrm{KL}(P_T(\mathbf{U}|\mathbf{L}) \| P(\mathbf{U}|\mathbf{L})), \\
\mathbb{E}_{P(\mathbf{L},\mathbf{U})} \log P_T(\mathbf{L}, \mathbf{U}) = & \mathbb{E}_{P(\mathbf{L})} \log P_T(\mathbf{L}) + \\
& \mathbb{E}_{P_G(\mathbf{U}|\mathbf{L})} \log P_T(\mathbf{U}|\mathbf{L}) - \\
& \mathrm{KL}(P_G(\mathbf{U}|\mathbf{L}) \| P(\mathbf{U}|\mathbf{L})),
\end{aligned}
\tag{15}
$$

where $KL(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

Our BGA module can keep a global receptive ability

Our proposed collaborative training can improve the generalization ability of our GCN module and BGA module.

CONTENTS

4 Experiments Node Classification

We conducted node classification on seven real-world datasets including homophilic graphs, heterophilic graphs and large scale networks.

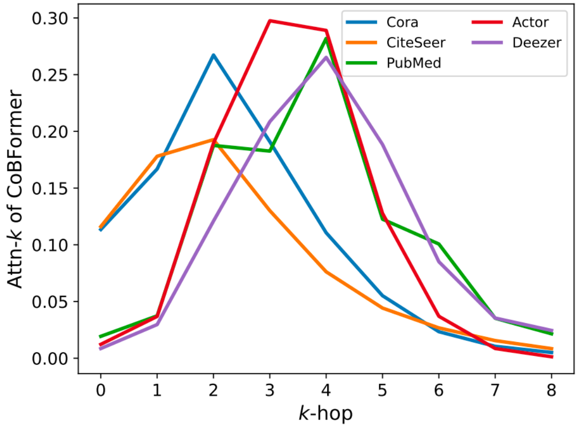
Table 2. Quantitative results ($\% \pm \sigma$) on node classification.

Dataset	Metric	GCN	GAT	NodeFormer	NAGphormer	SGFormer	CoB-G	CoB-T
Cora	Mi-F1	81.44 ± 0.78	81.88 ± 0.99	80.30 ± 0.66	79.62 ± 0.25	81.48 ± 0.94	84.96 ± 0.34	**85.28 ± 0.16**
	Ma-F1	80.65 ± 0.91	80.56 ± 0.55	79.12 ± 0.66	78.78 ± 0.57	79.28 ± 0.49	83.52 ± 0.15	**84.10 ± 0.28**
CiteSeer	Mi-F1	71.84 ± 0.22	72.26 ± 0.97	71.58 ± 1.74	67.46 ± 1.33	71.96 ± 0.13	**74.68 ± 0.33**	74.52 ± 0.48
	Ma-F1	68.69 ± 0.38	65.67 ± 2.28	67.28 ± 1.87	64.47 ± 1.58	68.49 ± 0.65	69.73 ± 0.45	**69.82 ± 0.55**
PubMed	Mi-F1	79.26 ± 0.23	78.46 ± 0.22	78.96 ± 2.71	77.36 ± 0.96	78.04 ± 0.41	80.52 ± 0.25	**81.42 ± 0.53**
	Ma-F1	79.02 ± 0.19	77.82 ± 0.22	78.14 ± 2.51	76.76 ± 0.91	77.86 ± 0.32	80.02 ± 0.28	**81.04 ± 0.49**
Actor	Mi-F1	30.97 ± 1.21	30.63 ± 0.68	35.42 ± 1.37	34.83 ± 0.95	**37.72 ± 1.00**	31.05 ± 1.02	37.41 ± 0.36
	Ma-F1	26.66 ± 0.82	20.73 ± 1.58	32.37 ± 1.38	32.20 ± 1.11	34.11 ± 2.78	27.01 ± 1.77	**34.96 ± 0.68**
Deezer	Mi-F1	63.10 ± 0.40	62.20 ± 0.41	63.59 ± 2.24	63.71 ± 0.58	66.68 ± 0.47	63.76 ± 0.62	**66.96 ± 0.37**
	Ma-F1	62.07 ± 0.31	60.99 ± 0.56	62.70 ± 2.20	62.06 ± 1.28	65.22 ± 0.68	62.32 ± 0.94	**65.63 ± 0.36**
Arxiv	Mi-F1	71.99 ± 0.14	71.30 ± 0.11	67.98 ± 0.60	71.38 ± 0.20	72.50 ± 0.28	**73.17 ± 0.18**	72.76 ± 0.11
	Ma-F1	51.89 ± 0.19	48.84 ± 0.31	46.24 ± 0.20	51.38 ± 0.47	**52.83 ± 0.31**	52.31 ± 0.40	51.64 ± 0.09
Products	Mi-F1	75.49 ± 0.24	76.19 ± 0.40	70.71 ± 0.27	76.41 ± 0.53	72.54 ± 0.80	78.09 ± 0.16	**78.15 ± 0.07**
	Ma-F1	37.02 ± 0.92	35.15 ± 0.20	30.09 ± 0.02	37.48 ± 0.38	33.72 ± 0.42	**38.21 ± 0.22**	37.91 ± 0.44

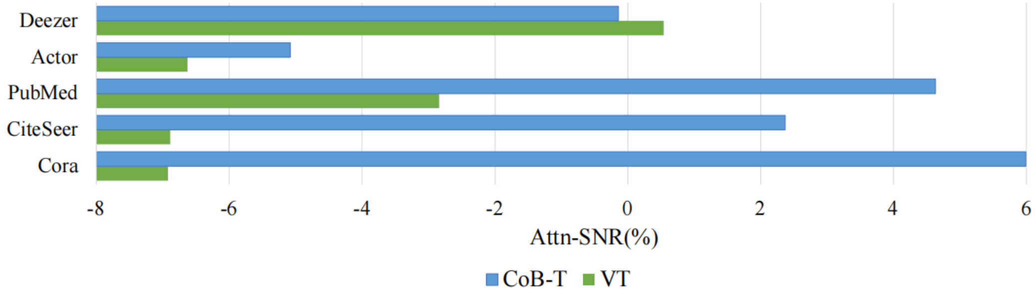
4 Experiments Ablation Studies && Analysis

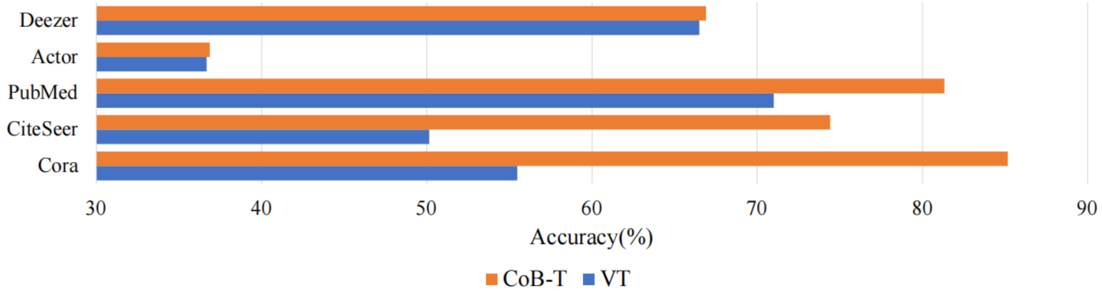
Table 3. Test accuracy and GPU memory of various CoBFormer variants. ‘V-A’ denotes the vanilla global attention. ‘B-A’ represents the BGA module. ‘C-T’ indicates whether collaborative training is applied.

Dataset	V-A	B-A	C-T	CoB-G	CoB-T	MEM
Cora	√	×	×	81.44	54.86	0.85G
	√	×	√	83.78	83.82	0.85G
	×	√	×	81.44	68.72	0.38G
	×	√	√	84.96	85.28	0.38G
PubMed	√	×	×	79.26	71.22	8.42G
	√	×	√	80.38	80.36	8.42G
	×	√	×	79.26	74.52	0.50G
	×	√	√	80.52	81.42	0.50G
Deezer	√	×	×	62.07	66.49	20.23G
	√	×	√	63.67	66.86	20.23G
	×	√	×	62.07	66.56	3.97G
	×	√	√	63.76	66.96	3.97G



Our method can effectively alleviate the over-globalizing problem





4 Experiments Parameter Study

We analyze the key parameters:

- *the collaborative learning strength coefficient α,*
- *the temperature coefficient τ*
- *the number of clusters P.*

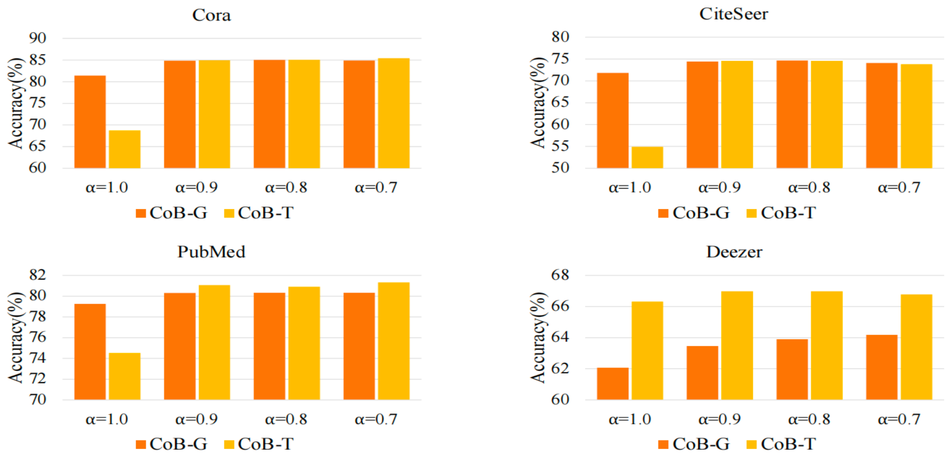


Figure 6. The average test accuracy of CoBFormer for different α.

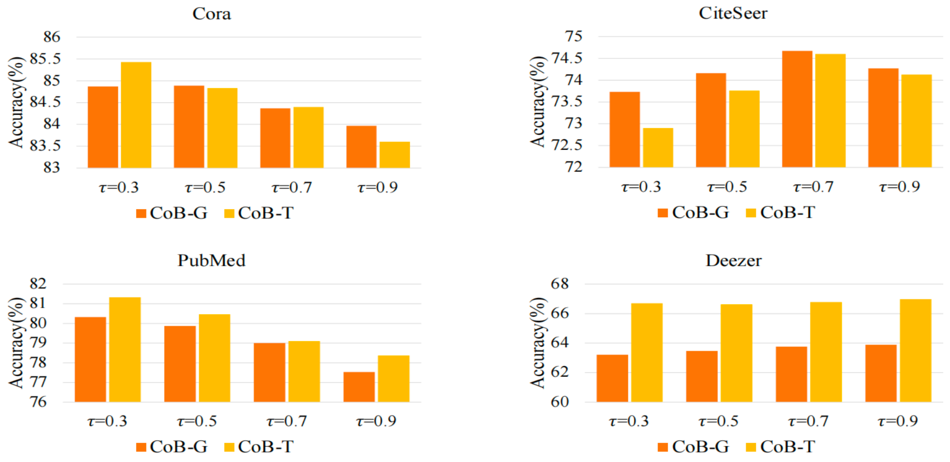


Figure 7. The average test accuracy of CoBFormer for different τ.

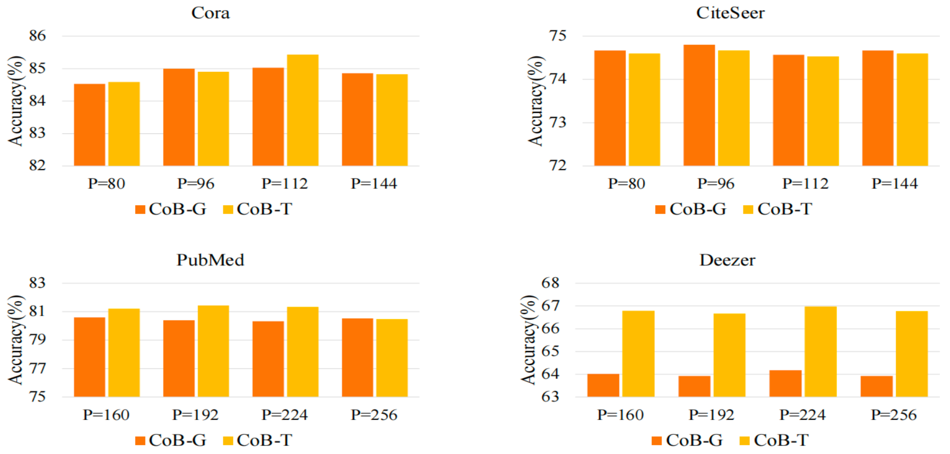


Figure 8. The average test accuracy of CoBFormer for different P.

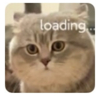
CONTENTS

5 Conclusion

- We discover the ***over-globalizing problem*** in Graph Transformers by presenting the theoretical insights and empirical results.
- We propose ***CoBFormer***, a bi-level global graph transformer with collaborative training, aiming at alleviating the over-globalizing problem and improving the generalization ability.
- Extensive experiments demonstrate that CoBFormer outperforms the state-of-the-art Graph Transformers and effectively solves the over-globalizing problem.
- We believe our work will provide valuable guidelines and insights for the development of advanced Graph Transformers.

Thanks

Q&A



Laner

Scan the QR code to add me as a friend.

Paper: https://arxiv.org/abs/2405.01102
Code: https://github.com/null-xyj/CoBFormer

yujie-xing@bupt.edu.cn