Supplementary Material for “LGI-GT: Graph Transformers with Local and Global Operators Interleaving”

Paper ID: 2470

A Dataset Statistics

In the main paper, we performed comparison experiment on five datasets to demonstrate the superiority of our LGI-GT over the state-of-the-art (SOTA) GNNs and GTs, and further conducted another comparison experiment to validate the effectiveness of our proposed LGI scheme. Statistics of all the datasets concerned are concluded in Table 1.

B Hyperparameters and Runtime

For the experiment comparing with the SOTA methods, the final hyperparameters for our LGI-GT on related datasets are concluded in Table 2, while the runtime hardware and indicators (including number of parameters and time consumed) are shown in Table 3.

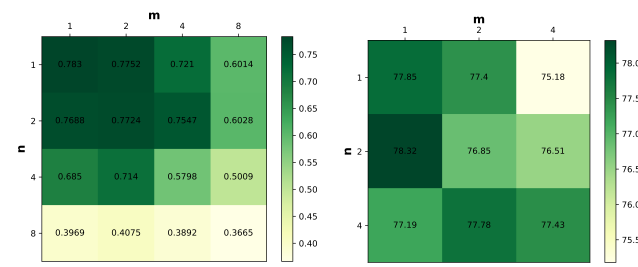
C Different Combinations of n and m

Although we have tried a different configuration for values of n (the number of GConvs) and m (the number of TLayers) when demonstrating the effectiveness of the LGI scheme in the main paper ($n = 2, m = 1$), here we explore more on this.

Figure 1 shows how the values of n and m influence the performance of LGI-GT on CLUSTER and ogbg-moltox21. From Figure 1(a), we can see that the best configuration for CLUSTER is just $n = m = 1$ and the performance is degraded as either n or m becomes larger, while Figure 1(b) shows $n = 2, m = 1$ is the best for ogbg-moltox21. We can conclude that there is no such a fixed combination of n and m optimal across all the datasets, and small values of them are recommended in our experience.

D Effectiveness of the Skip Forward Propagating Method for the [CLS] Token

Particularly for our LGI-GT, the embedding of the [CLS] token is propagated in a skip manner. Here we compare the performance of LGI-GT models with different readout methods for aggregating node representations to get a final graph representation. See Figure 2, our proposed [CLS] skip propagating method is consistently better than the other readout methods, demonstrating its great compatibility with LGI-GT.



(a) CLUSTER (b) ogbg-moltox21

Figure 1: Performance w.r.t. different combinations of n and m.

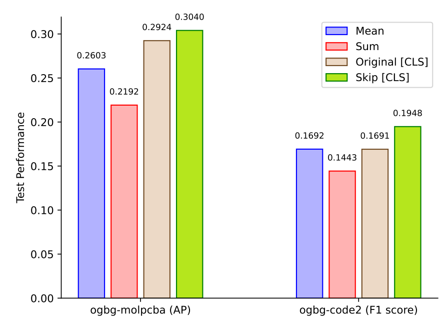


Figure 2: Performance on ogbg-molpcba and ogbg-code2 w.r.t. the readout methods.

E Visualizations

In addition to examples shown in the main paper, we give more visualization results here in Figure 3. Also, the first column displays the original molecules from ogbg-molpcba, and the other columns from left to right are the visualization results of LGI-GT, GNN+Transformer and Parallel GT in turn. We can see LGI-GT made the [CLS] node attend more on join nodes of several motifs and important nodes to better distinguish different graphs or motifs, which demonstrates LGI-GT is good at handling structure information and focuses on the discriminative nodes.

Dataset	**# Graphs**	**Average # nodes**	**Average # edges**	**Directed**	**Task**	**Task level**	**Metric**
ZINC	12,000	23.2	24.9	No	regression	graph	MAE
PATTERN	14,000	118.9	3,039.3	No	binary classification	inductive node	Accuracy
CLUSTER	12,000	117.2	2,150.9	No	6-class classification	inductive node	Accuracy
ogbg-molpcba	437,929	26.0	28.1	No	128-task binary classification	graph	AP
ogbg-code2	452,741	125.2	124.2	Yes	5-token sequence prediction	graph	F1 score
NCI1	4,110	29.9	32.3	No	binary classification	graph	Accuracy
NCI109	4,127	29.7	32.1	No	binary classification	graph	Accuracy
ogbg-molbbbp	2,039	24.1	26.0	No	1-task binary classification	graph	ROC-AUC
ogbg-moltox21	7,831	18.6	19.3	No	12-task binary classification	graph	ROC-AUC

Table 1: Summary of datasets: the top five are used in comparison experiments, while the others are for the ablation study.

Hyperparameter	**ZINC**	**PATTERN**	**CLUSTER**	**ogbg-molpcba**	**ogbg-code2**
# Blocks (Layers)	10	6	16	5	4
# Hidden dim	64	64	48	384	256
GConv	GINEConv	GCNConv	GCNConv	EELA	GCNConv
TLayer	Transformer	Transformer	Transformer	Transformer	Transformer
# Heads	4	4	8	8	4
GConv dropout	0.0	0.0	0.0	0.3	0.0
Attention dropout	0.5	0.3	0.5	0.3	0.0
TLayer FFN dropout	0.0	0.3	0.1	0.3	0.4
Graph pooling	sum	–	–	CLS	CLS
PE/SE	RWSE-20	RWSE-7	RWSE-6	–	–
# PE dim	28	16	16	–	–
# PE encoder	linear	linear	linear	–	–
Batch size	32	32	32	256	32
Learning rate	0.001	0.0003	0.001	0.0002	0.0002
# Epochs	2000	100	100	100	30
# Warmup Epochs	50	5	5	10	5
Weight decay	1e-5	1e-5	1e-5	1e-4	1e-6
Scheduler	linear	none	cosine	linear	linear

Table 2: The final hyperparameters of our LGI-GT in the comparison experiments. For the sake of fair comparison, we set $n = m = 1$ for LGI-GT on all these five datasets (do not tune them as hyperparemeters).

Runtime	**ZINC**	**PATTERN**	**CLUSTER**	**ogbg-molpcba**	**ogbg-code2**
# Parameters	841,701	252,432	381,650	9,738,368	12,846,898
Hardware	GTX 1080 Ti	GTX 1080 Ti	RTX 3090	RTX 3090	RTX 3090
Time (epoch/total)	34s / 18.74h	32s / 0.90h	55s / 1.52h	190s / 5.28h	1463s / 12.19h

Table 3: Runtime hardware and indicators of our LGI-GT in the comparison experiments.

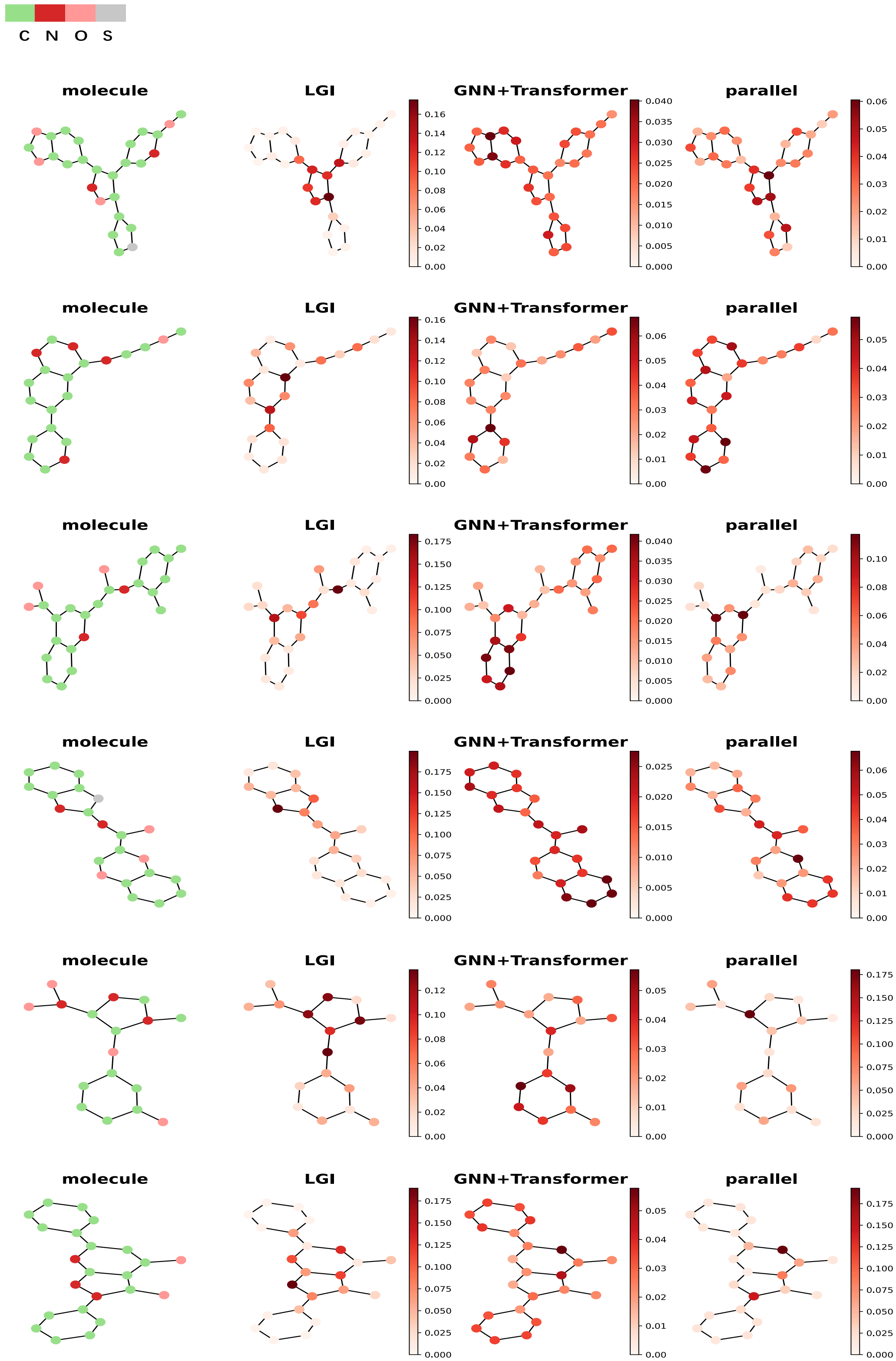


Figure 3: Visualization of the [CLS] node attention to the real nodes.