GraphMAE: Self-Supervised Masked Graph Autoencoders

Zhenyu Hou
Tsinghua University
houzy21@mails.tsinghua.edu.cn

Xiao Liu
Tsinghua University
liuxiao21@mails.tsinghua.edu.cn

Yukuo Cen
Tsinghua University
cyk20@mails.tsinghua.edu.cn

Yuxiao Dong*
Tsinghua University
yuxiaod@tsinghua.edu.cn

Hongxia Yang
DAMO Academy, Alibaba Group
yang.yhx@alibaba-inc.com

Chunjie Wang
BirenTech Research
cjwang@birentech.com

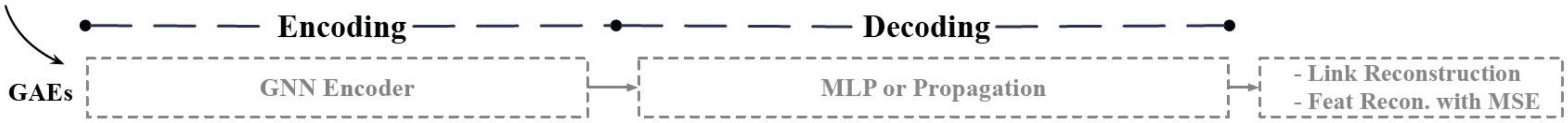
Jie Tang*
Tsinghua University
jietang@tsinghua.edu.cn

KDD 2022
Code: github.com/THUDM/GraphMAE

2022.08.24 • ChongQing

Reported by Chenghong Li

Introduction

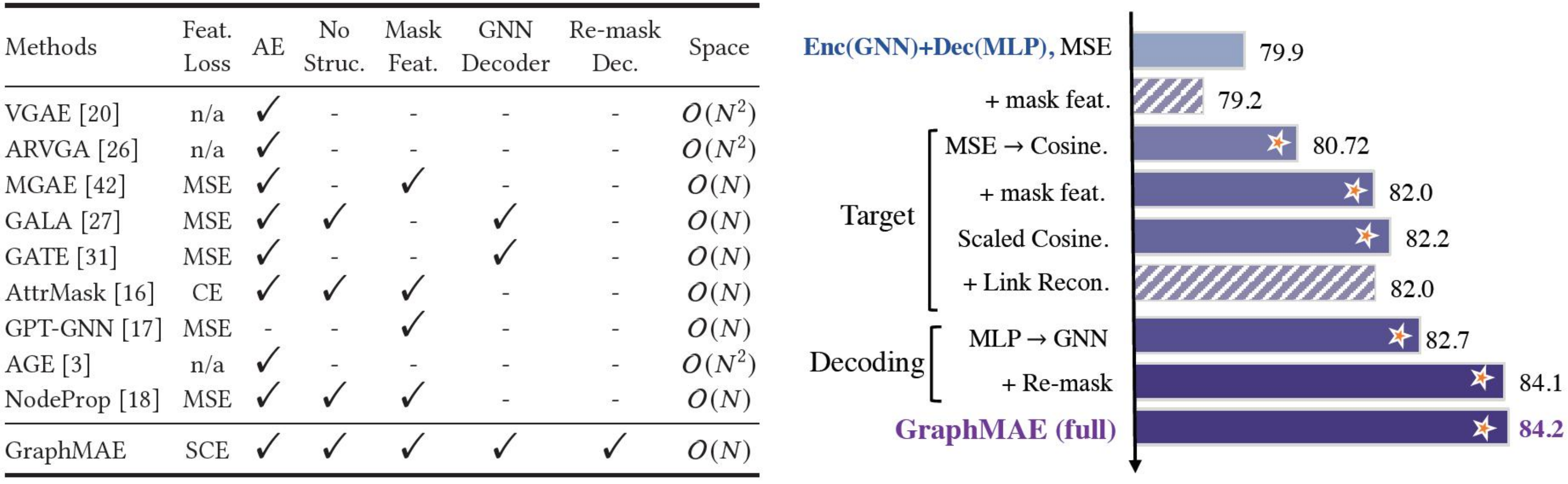


GAEs

Encoding — GNN Encoder

Decoding — MLP or Propagation

- Link Reconstruction
- Feat Recon. with MSE



Methods	Feat. Loss	AE	No Struc.	Mask Feat.	GNN Decoder	Re-mask Dec.	Space
VGAE [20]	n/a	✓	-	-	-	-	$O(N^2)$
ARVGA [26]	n/a	✓	-	-	-	-	$O(N^2)$
MGAE [42]	MSE	✓	-	✓	-	-	$O(N)$
GALA [27]	MSE	✓	✓	-	✓	-	$O(N)$
GATE [31]	MSE	✓	-	-	✓	-	$O(N)$
AttrMask [16]	CE	✓	✓	✓	-	-	$O(N)$
GPT-GNN [17]	MSE	-	-	✓	-	-	$O(N)$
AGE [3]	n/a	✓	-	-	-	-	$O(N^2)$
NodeProp [18]	MSE	✓	✓	✓	-	-	$O(N)$
GraphMAE	SCE	✓	✓	✓	✓	✓	$O(N)$

(a) Technical comparison between generative SSL methods.

(b) The effect of GraphMAE designs on the performance on Cora dataset.

Figure 1: Comparison between generative SSL methods and the effect of GraphMAE design. *AE*: autoencoder methods; *No Struct.*: no structure reconstruction objective; *Mask Feat.*: use masking to corrupt input features; *GNN Decoder*: use GNN as the decoder; *Re-mask Dec.*: re-mask encoder output before fed into decoder; *Space*: run-time memory consumption; *MSE*: Mean Squared Error; *SCE*: Scaled Cosine Error; *CE*: Cross-Entropy Error; *SCE* represents our proposed Scaled Cosine Error.

Method

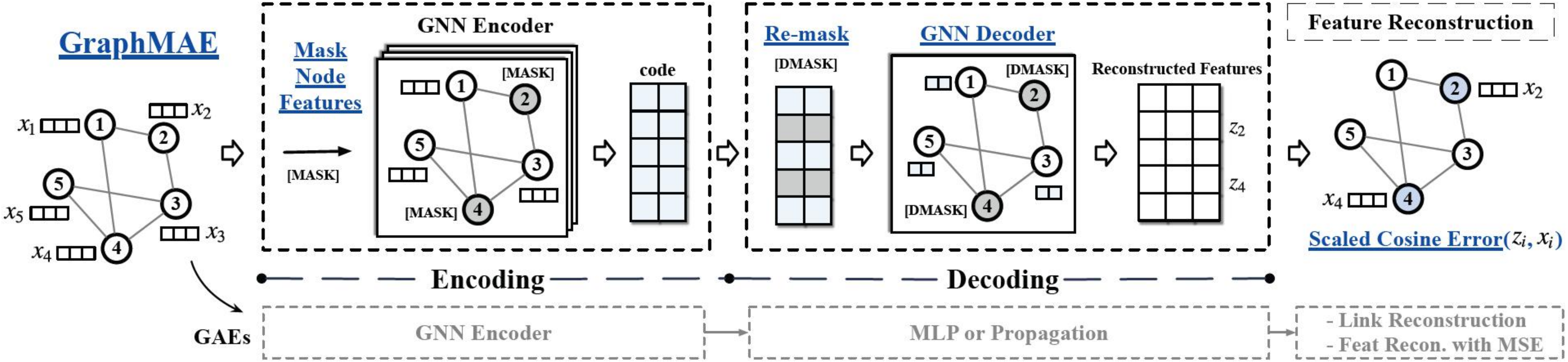


Figure 2: Illustration of GraphMAE and the comparison with GAE. We underline the key operations in GraphMAE. During pre-training, GraphMAE first masks input node features with a mask token [MASK]. The corrupted graph is encoded into code by a GNN encoder. In the decoding, GraphMAE re-masks the code of selected nodes with another token [DMASK], and then employs a GNN, e.g., GAT, GIN, as the decoder. The output of the decoder is used to reconstruct input node features of masked nodes, with the scaled cosine error as the criterion. Previous GAEs usually use a single-layer MLP or Laplacian matrix in the decoding and focus more on restoring graph structure.

Method

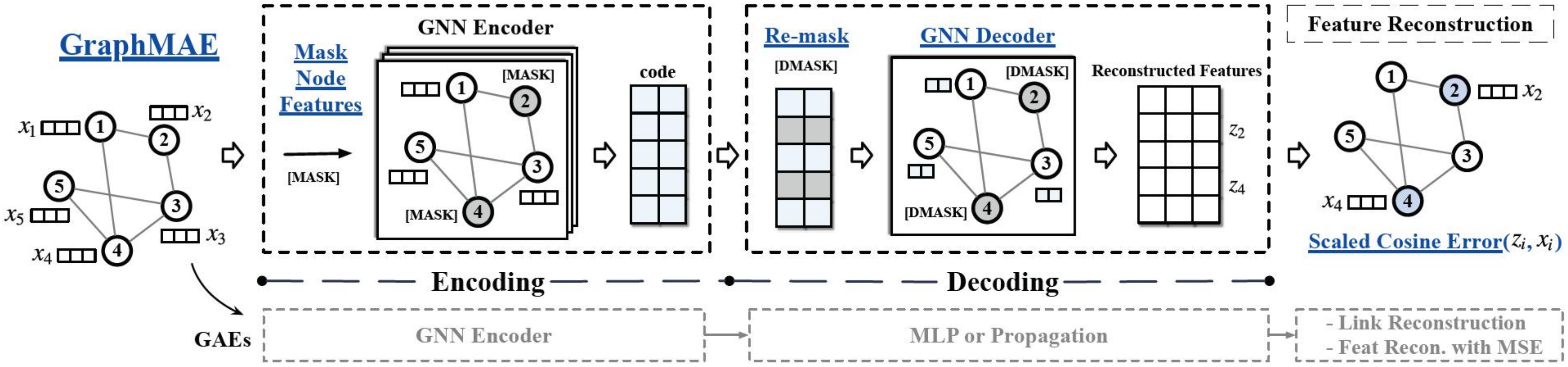


Figure 2: Illustration of GraphMAE and the comparison with GAE. We underline the key operations in GraphMAE. During pre-training, GraphMAE first masks input node features with a mask token [MASK]. The corrupted graph is encoded into code by a GNN encoder. In the decoding, GraphMAE re-masks the code of selected nodes with another token [DMASK], and then employs a GNN, e.g., GAT, GIN, as the decoder. The output of the decoder is used to reconstruct input node features of masked nodes, with the scaled cosine error as the criterion. Previous GAEs usually use a single-layer MLP or Laplacian matrix in the decoding and focus more on restoring graph structure.

$$\widetilde{x}_i = \begin{cases} x_{[M]} & v_i \in \widetilde{\mathcal{V}} \\ x_i & v_i \notin \widetilde{\mathcal{V}} \end{cases}$$

$$\boldsymbol{H} = f_E(\boldsymbol{A}, \boldsymbol{X}),\ \mathcal{G}' = f_D(\boldsymbol{A}, \boldsymbol{H}), \qquad (1)$$

Method

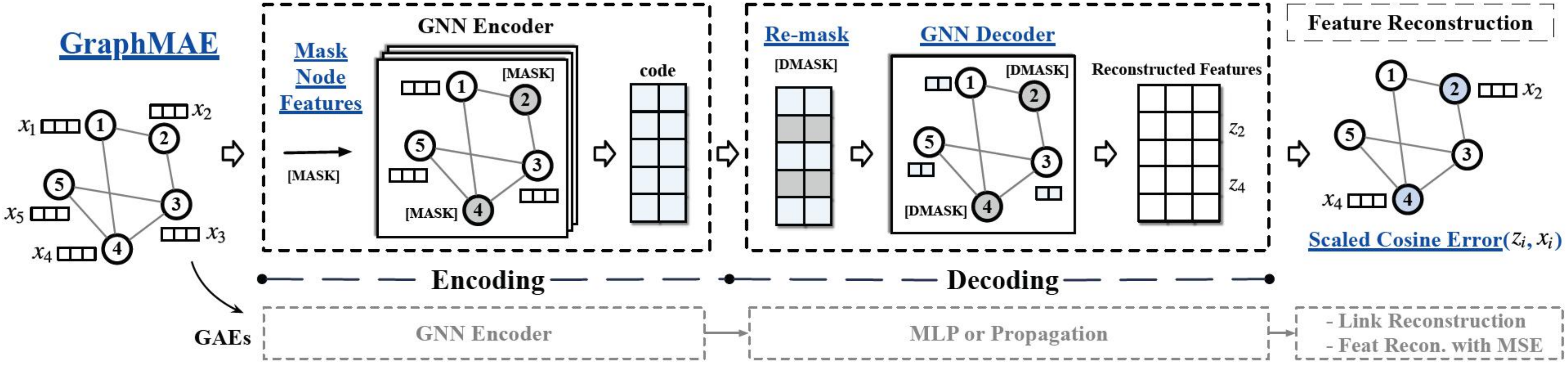


Figure 2: Illustration of GraphMAE and the comparison with GAE. We underline the key operations in GraphMAE. During pre-training, GraphMAE first masks input node features with a mask token [MASK]. The corrupted graph is encoded into code by a GNN encoder. In the decoding, GraphMAE re-masks the code of selected nodes with another token [DMASK], and then employs a GNN, e.g., GAT, GIN, as the decoder. The output of the decoder is used to reconstruct input node features of masked nodes, with the scaled cosine error as the criterion. Previous GAEs usually use a single-layer MLP or Laplacian matrix in the decoding and focus more on restoring graph structure.

$$\widetilde{\boldsymbol{h}}_i = \begin{cases} \boldsymbol{h}_{[M]} & v_i \in \widetilde{\mathcal{V}} \\ \boldsymbol{h}_i & v_i \notin \widetilde{\mathcal{V}} \end{cases}$$

$$\mathcal{L}_{\mathrm{SCE}} = \frac{1}{|\widetilde{\mathcal{V}}|} \sum_{v_i \in \widetilde{\mathcal{V}}} \left(1 - \frac{\boldsymbol{x}_i^T \boldsymbol{z}_i}{\|\boldsymbol{x}_i\| \cdot \|\boldsymbol{z}_i\|}\right)^{\gamma}, \ \gamma \geq 1, \qquad (2)$$

$$Z = f_D(\boldsymbol{A}, \widetilde{\boldsymbol{H}})$$

Experiments

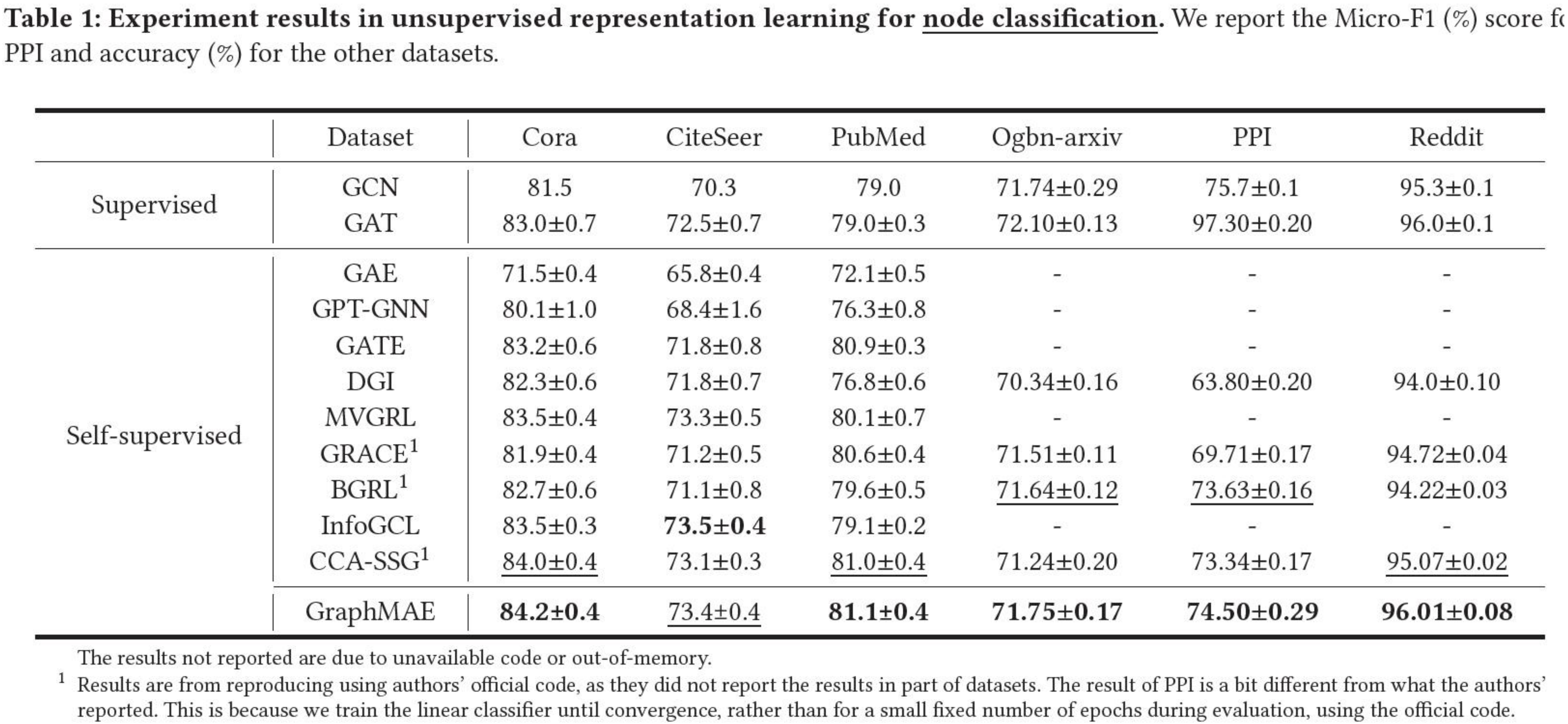


Table 1: Experiment results in unsupervised representation learning for node classification. We report the Micro-F1 (%) score f PPI and accuracy (%) for the other datasets.

	Dataset	Cora	CiteSeer	PubMed	Ogbn-arxiv	PPI	Reddit
Supervised	GCN	81.5	70.3	79.0	71.74±0.29	75.7±0.1	95.3±0.1
	GAT	83.0±0.7	72.5±0.7	79.0±0.3	72.10±0.13	97.30±0.20	96.0±0.1
Self-supervised	GAE	71.5±0.4	65.8±0.4	72.1±0.5	-	-	-
	GPT-GNN	80.1±1.0	68.4±1.6	76.3±0.8	-	-	-
	GATE	83.2±0.6	71.8±0.8	80.9±0.3	-	-	-
	DGI	82.3±0.6	71.8±0.7	76.8±0.6	70.34±0.16	63.80±0.20	94.0±0.10
	MVGRL	83.5±0.4	73.3±0.5	80.1±0.7	-	-	-
	GRACE[1]	81.9±0.4	71.2±0.5	80.6±0.4	71.51±0.11	69.71±0.17	94.72±0.04
	BGRL[1]	82.7±0.6	71.1±0.8	79.6±0.5	71.64±0.12	73.63±0.16	94.22±0.03
	InfoGCL	83.5±0.3	**73.5±0.4**	79.1±0.2	-	-	-
	CCA-SSG[1]	84.0±0.4	73.1±0.3	81.0±0.4	71.24±0.20	73.34±0.17	95.07±0.02
	GraphMAE	**84.2±0.4**	73.4±0.4	**81.1±0.4**	**71.75±0.17**	**74.50±0.29**	**96.01±0.08**

The results not reported are due to unavailable code or out-of-memory.

[1] Results are from reproducing using authors' official code, as they did not report the results in part of datasets. The result of PPI is a bit different from what the authors' reported. This is because we train the linear classifier until convergence, rather than for a small fixed number of epochs during evaluation, using the official code.

Experiments

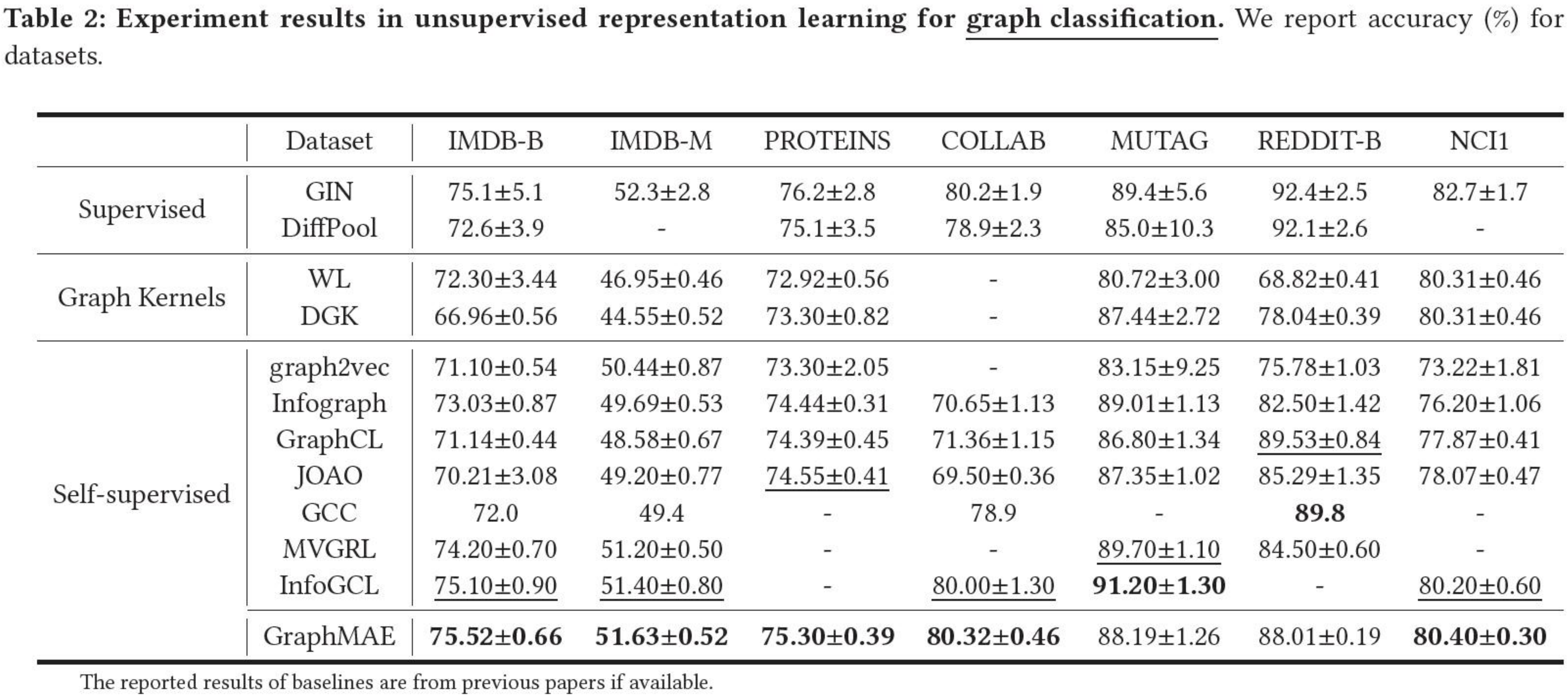


Table 2: Experiment results in unsupervised representation learning for graph classification. We report accuracy (%) for all datasets.

	Dataset	IMDB-B	IMDB-M	PROTEINS	COLLAB	MUTAG	REDDIT-B	NCI1
Supervised	GIN	75.1±5.1	52.3±2.8	76.2±2.8	80.2±1.9	89.4±5.6	92.4±2.5	82.7±1.7
	DiffPool	72.6±3.9	-	75.1±3.5	78.9±2.3	85.0±10.3	92.1±2.6	-
Graph Kernels	WL	72.30±3.44	46.95±0.46	72.92±0.56	-	80.72±3.00	68.82±0.41	80.31±0.46
	DGK	66.96±0.56	44.55±0.52	73.30±0.82	-	87.44±2.72	78.04±0.39	80.31±0.46
Self-supervised	graph2vec	71.10±0.54	50.44±0.87	73.30±2.05	-	83.15±9.25	75.78±1.03	73.22±1.81
	Infograph	73.03±0.87	49.69±0.53	74.44±0.31	70.65±1.13	89.01±1.13	82.50±1.42	76.20±1.06
	GraphCL	71.14±0.44	48.58±0.67	74.39±0.45	71.36±1.15	86.80±1.34	89.53±0.84	77.87±0.41
	JOAO	70.21±3.08	49.20±0.77	74.55±0.41	69.50±0.36	87.35±1.02	85.29±1.35	78.07±0.47
	GCC	72.0	49.4	-	78.9	-	**89.8**	-
	MVGRL	74.20±0.70	51.20±0.50	-	-	89.70±1.10	84.50±0.60	-
	InfoGCL	75.10±0.90	51.40±0.80	-	80.00±1.30	**91.20±1.30**	-	80.20±0.60
	GraphMAE	**75.52±0.66**	**51.63±0.52**	**75.30±0.39**	**80.32±0.46**	88.19±1.26	88.01±0.19	**80.40±0.30**

The reported results of baselines are from previous papers if available.

Experiments

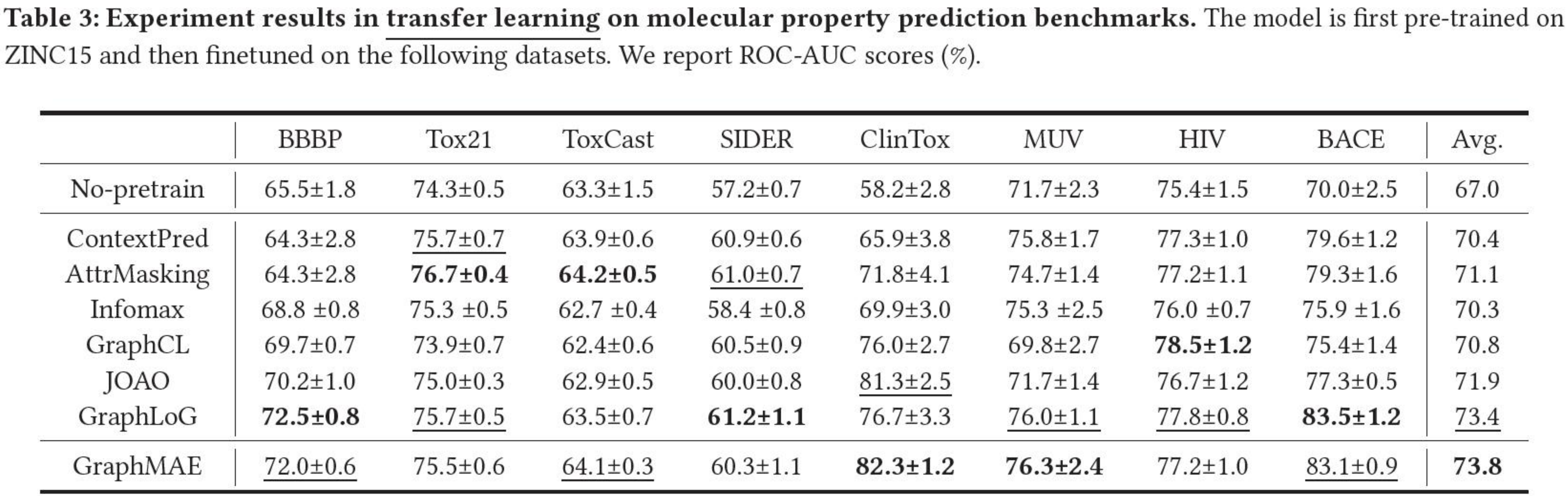


Table 3: Experiment results in transfer learning on molecular property prediction benchmarks. The model is first pre-trained on ZINC15 and then finetuned on the following datasets. We report ROC-AUC scores (%).

	BBBP	Tox21	ToxCast	SIDER	ClinTox	MUV	HIV	BACE	Avg.
No-pretrain	65.5±1.8	74.3±0.5	63.3±1.5	57.2±0.7	58.2±2.8	71.7±2.3	75.4±1.5	70.0±2.5	67.0
ContextPred	64.3±2.8	75.7±0.7	63.9±0.6	60.9±0.6	65.9±3.8	75.8±1.7	77.3±1.0	79.6±1.2	70.4
AttrMasking	64.3±2.8	**76.7±0.4**	**64.2±0.5**	61.0±0.7	71.8±4.1	74.7±1.4	77.2±1.1	79.3±1.6	71.1
Infomax	68.8 ±0.8	75.3 ±0.5	62.7 ±0.4	58.4 ±0.8	69.9±3.0	75.3 ±2.5	76.0 ±0.7	75.9 ±1.6	70.3
GraphCL	69.7±0.7	73.9±0.7	62.4±0.6	60.5±0.9	76.0±2.7	69.8±2.7	**78.5±1.2**	75.4±1.4	70.8
JOAO	70.2±1.0	75.0±0.3	62.9±0.5	60.0±0.8	81.3±2.5	71.7±1.4	76.7±1.2	77.3±0.5	71.9
GraphLoG	**72.5±0.8**	75.7±0.5	63.5±0.7	**61.2±1.1**	76.7±3.3	76.0±1.1	77.8±0.8	**83.5±1.2**	73.4
GraphMAE	72.0±0.6	75.5±0.6	64.1±0.3	60.3±1.1	**82.3±1.2**	**76.3±2.4**	77.2±1.0	83.1±0.9	**73.8**

Experiments

Table 4: Ablation studies of the decoder type, re-mask and reconstruction criterion on node- and graph-level datasets.

	Dataset	Node-Level			Graph-Level	
		Cora	PubMed	Arxiv	MUTAG	IMDB-B
COMP.	GraphMAE	84.2	81.1	71.75	88.19	75.52
	w/o mask	79.7	77.9	70.97	82.58	74.42
	w/o re-mask	82.7	80.0	71.61	86.29	74.42
	w/ MSE	79.1	73.1	67.44	86.30	74.04
Decoder	MLP	82.2	80.4	71.54	87.16	73.94
	GCN	81.3	79.1	71.59	87.78	74.54
	GIN	81.8	80.2	71.41	88.19	75.52
	GAT	84.2	81.1	71.75	86.27	74.04

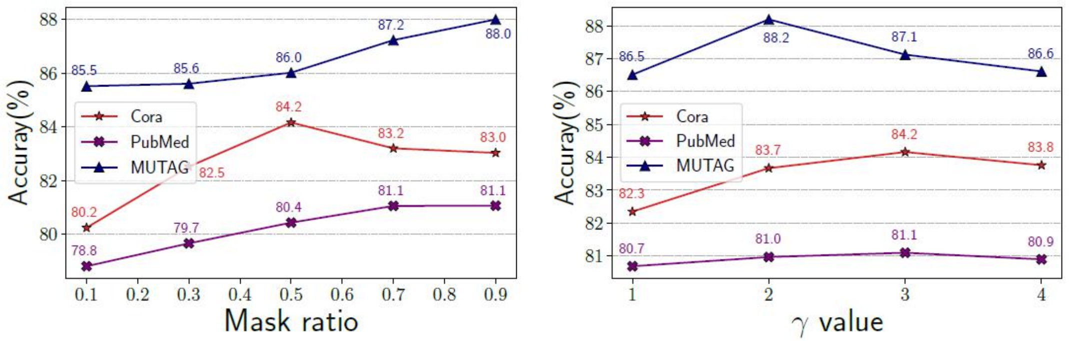


Figure 3: Ablation studies of mask ratio and scaling factor.

Thanks