中国科学技术大学
University of Science and Technology of China

数据智能实验室
Data Intelligence Lab

NEURAL INFORMATION PROCESSING SYSTEMS

Less but More: Linear Adaptive Graph Learning Empowering Spatiotemporal Forecasting

Jiaming Ma (Presenter), Binwu Wang*, Guanjun Wang, Kuo Yang,

Zhengyang Zhou, Pengkun Wang, Xu Wang, Yang Wang*

University of Science ang Technology of China (USTC)

Brief Introduction

➢ Challenges of Adaptive Graph in Spatiotemporal Learning

Adaptive Graph Learning. Adaptive graph learning is typically formulated through a reparameterization of two learnable node embedding matrices, $\mathbf{E}_1, \mathbf{E}_2 \in \mathbb{R}^{N \times d_G}$, where $d_G \ll N$ is the prescribed dimension of the graph generation embeddings. The adaptive graph is then constructed as [7, 9, 23, 24, 29]:

$$\mathbf{A} = \mathrm{Softmax}\left(\mathrm{ReLU}\left(\mathbf{E}_1 \mathbf{E}_2^\top\right)\right) \in \mathbb{R}^{N \times N}. \tag{1}$$

❖ *Performance*

ReLU function before Softmax makes more Noises.

❖ *Efficiency*

Linear kernel methods meets low-rank dilemma.

NEURAL INFORMATION PROCESSING SYSTEMS

❖ *Performance*

ReLU function before Softmax makes more Noises.

Theorem 1. *Edge Noise Amplification Theory*

Let $\mathbf{E}_1 = [e_{ik}^{(1)}], \mathbf{E}_2 = [e_{jk}^{(2)}] \in \mathbb{R}^{N \times d_G}$ be the graph generating embeddings where all elements belonging to them satisfy an independent normal distribution $\mathcal{N}\left(0, \sigma^2\right)$ with $\sigma > 0$. N corresponds to the number of nodes and $d_G \ll N$ is the given dimensionality of graph generation embeddings. The Adaptive Graph with or without $\mathrm{ReLU}\left(\cdot\right)$ are respectively calculated as follows,

$$\mathbf{A}^R = \mathrm{Softmax}\left(\mathrm{ReLU}\left(\mathbf{E}_1\mathbf{E}_2^\top\right)\right) \in \mathbb{R}^{N\times N}, \quad \mathbf{A} = \mathrm{Softmax}\left(\mathbf{E}_1\mathbf{E}_2^\top\right) \in \mathbb{R}^{N\times N}. \quad (24)$$

Then, the calculation of Adaptive Graph $\mathbf{A}^R$ will lead to more edge noises than $\mathbf{A}$. Specifically, there exits,

(1) If nodes i and j have positive similarity, then $\mathbf{A}_{ij}^R \leq \mathbf{A}_{ij}$;

(2) If nodes i and j have negative similarity, then with high possibility $\mathbf{A}_{ij}^R \geq \mathbf{A}_{ij}$.

**Proof is available in the paper.*

Brief Introduction

❖ *Performance*

ReLU function before Softmax makes more Noises.

Table 7: Performance experiments on evaluating the negativity of ReLU in the adaptive graph convolution. We report the average results in five experiments. $\downarrow$ indicates the relative percentage decreasing regarding each methods itself.

Methods	LargeST-SD			Electricity			KnowAir			Beijing Weibo		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
AGCRN	18.39	33.63	13.78	211.5	1847	16.95	16.34	24.81	63.26	0.8505	1.6998	33.68
w/o ReLU	18.29$_{\downarrow 0.54\%}$	33.18$_{\downarrow 1.34\%}$	13.32$_{\downarrow 3.34\%}$	210.0$_{\downarrow 0.71\%}$	1841$_{\downarrow 0.32\%}$	15.53$_{\downarrow 8.37\%}$	16.05$_{\downarrow 1.77\%}$	24.36$_{\downarrow 1.81\%}$	61.09$_{\downarrow 3.43\%}$	0.8481$_{\downarrow 0.28\%}$	1.6972$_{\downarrow 0.15\%}$	33.40$_{\downarrow 0.83\%}$
GWNet	18.07	29.97	12.70	200.3	1820	13.48	15.49	23.85	56.73	0.8315	1.6777	31.74
w/o ReLU	17.97$_{\downarrow 0.55\%}$	29.33$_{\downarrow 2.14\%}$	12.21$_{\downarrow 3.86\%}$	199.0$_{\downarrow 0.65\%}$	1755$_{\downarrow 3.57\%}$	13.23$_{\downarrow 1.85\%}$	15.49$_{\downarrow 0.00\%}$	23.75$_{\downarrow 0.42\%}$	56.63$_{\downarrow 0.18\%}$	0.8292$_{\downarrow 0.28\%}$	1.6665$_{\downarrow 0.67\%}$	30.88$_{\downarrow 2.71\%}$
D^{2}STGNN	17.13	28.60	12.15	224.8	2110	17.46	15.39	24.31	55.41	0.8489	1.7216	31.89
w/o ReLU	16.99$_{\downarrow 0.82\%}$	28.46$_{\downarrow 0.49\%}$	12.03$_{\downarrow 0.99\%}$	212.6$_{\downarrow 5.43\%}$	2016$_{\downarrow 4.45\%}$	17.33$_{\downarrow 0.74\%}$	15.28$_{\downarrow 0.71\%}$	24.16$_{\downarrow 0.62\%}$	53.24$_{\downarrow 3.92\%}$	0.8346$_{\downarrow 1.68\%}$	1.7208$_{\downarrow 0.05\%}$	31.35$_{\downarrow 1.69\%}$

Table 8: Average convergence epochs on evaluating the negativity of ReLU in the adaptive graph convolution in five experiments. The maximum allowable epochs are 300. $\downarrow$ indicates the relative percentage decreasing regarding each methods.

Datasets	AGCRN	w/o ReLU	GWNet	w/o ReLU	D^{2}STGNN	w/o ReLU
LargeST-SD	216	137$_{\downarrow 57.66\%}$	215	201$_{\downarrow 6.96\%}$	207	186$_{\downarrow 11.29\%}$
Electricity	300	287$_{\downarrow 4.53\%}$	208	185$_{\downarrow 12.43\%}$	56	52$_{\downarrow 7.69\%}$
KnowAir	41	39$_{\downarrow 5.13\%}$	34	34$_{\downarrow 0.00\%}$	36	33$_{\downarrow 9.10\%}$
Beijing Weibo	78	75$_{\downarrow 0.40\%}$	156	73$_{\downarrow 113.70\%}$	47	43$_{\downarrow 9.30\%}$

Brief Introduction

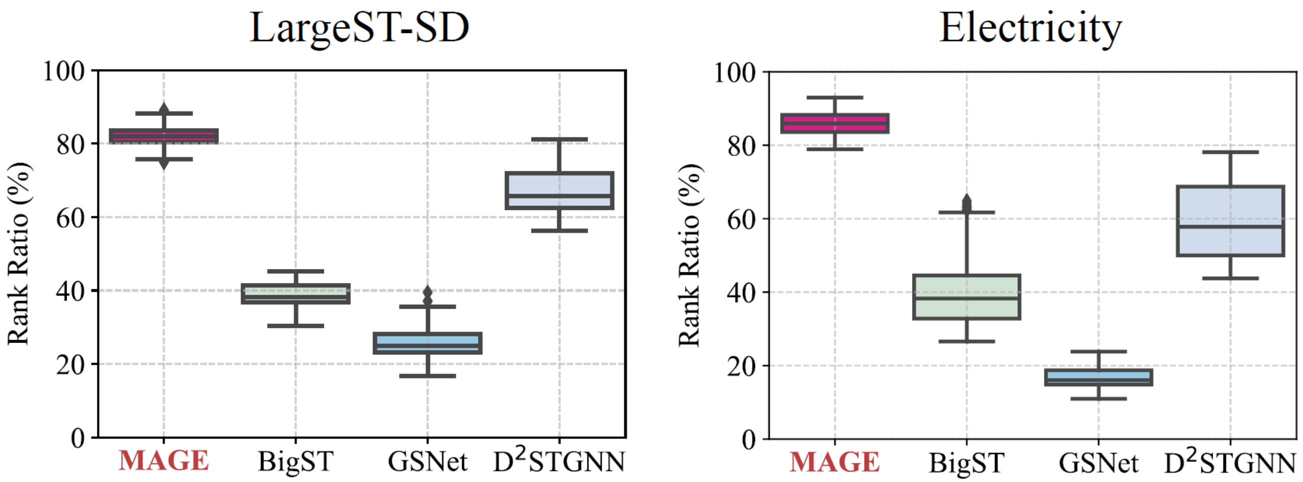
❖ *Efficiency*

Linear kernel methods meets low-rank dilemma.

Low Rank Bottlenecks. However, the approximate method often incurs degradation in representational capacity. To theoretically characterize this trade-off, we leverage matrix theory, where the rank of the learned adjacency matrix can be used as a measure of the high-dimensional information preserved in the feature representations. Specifically, the rank of the adaptive graph satisfies:

$$\mathrm{Rank}\left(\mathbf{A}\right) = \mathrm{Rank}\left(\mathrm{Softmax}\left(\mathbf{E}_1\right)\mathrm{Softmax}\left(\mathbf{E}_2^\top\right)\right) \leq \min\left\{N, d_G\right\} = d_G \ll N. \tag{7}$$

$$\Longrightarrow \mathrm{Rank}(\mathbf{H}^{(c)}) = \mathrm{Rank}(\mathbf{A}\mathbf{H}^{(c-1)}) \leq \min\left\{d_G, N, d\right\} = d_G < d. \tag{8}$$



Methods: MAGE

(1) Discard ReLU

$$\mathbf{A} = \operatorname{Softmax}\left(\mathbf{E}_1\mathbf{E}_2^\top\right) \in \mathbb{R}^{N\times N}.$$

(2) Linear Kernel Methods with Summation (Upper the rank)

$$\mathbf{H} = \sum_{k=1}^{K} \alpha_k \mathbf{A}^{(k)}\mathbf{H} = \sum_{k=1}^{K} \alpha_k \operatorname{Softmax}(\mathbf{E}_1^{(k)}) \operatorname{Softmax}(\mathbf{E}_2^{(k)\top})\mathbf{H} \in \mathbb{R}^{N\times d}.$$

$$\operatorname{Rank}(\mathbf{H}^{(c)}) \leq \min\{d, \sum_{k=1}^{K} \min\{d_G, N, d\}\} = \min\{d, K d_G\}$$

(3) Mixture-of-Adpative Graph Expert (MAGE)

$$\operatorname{MAGE}(\mathbf{H}) = \sum_{k=1}^{K_G} \operatorname{diag}\left(\alpha_{1k}, \alpha_{2k}, \ldots, \alpha_{Nk}\right) \mathbf{A}^{(k)}\mathbf{H},$$

$$\tilde{\alpha}_{ik} = \operatorname{Sigmoid}\left(\mathbf{H}_i^{(c-1)}\boldsymbol{\theta}_k^\top + \gamma_k\right) = \frac{1}{1+\exp\left(-\gamma_k\right)\exp\left(-\mathbf{H}_i^{(c-1)}\boldsymbol{\theta}_k^\top\right)} = \begin{cases} 1, & \gamma_k \to +\infty, \\ 0, & \gamma_k \to -\infty. \end{cases}$$

Methods: MAGE

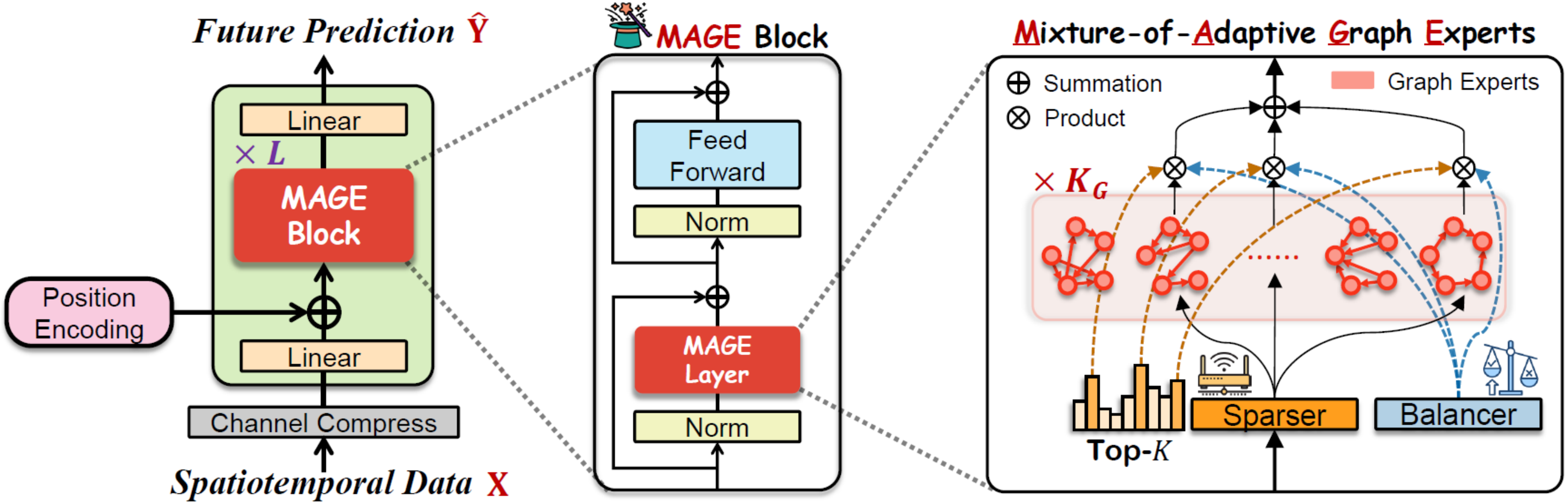


Figure 1: The architecture of MAGE for efficient adaptive graph learning.

Experimental Results

Table 1: Performance comparisons. The **best** and second best mean performance are in corresponding colors. The '-' marker indicates baseline incur out-of-memory issues even on minimum batch size. The '/' marker indicates baseline is not applicable to this dataset due to the absence of key metadata (e.g., latitude and longitude). All experimental results are the average of five independent runs.

Method	SD			GBA			GLA			CA			XTraffic		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
STGCN	19.27	33.57	13.49	23.29	38.15	17.82	22.22	37.98	14.30	20.68	35.68	15.55	13.55	26.58	31.15
DGCRN	17.79	29.31	12.33	20.53	33.40	16.79	-	-	-	-	-	-	-	-	-
AGCRN	18.39	33.63	13.78	20.69	34.30	16.05	20.26	34.86	12.39	-	-	-	-	-	-
GWNet	18.07	29.97	12.70	20.83	33.37	17.30	20.37	32.65	12.71	19.75	31.71	15.84	15.25	28.55	21.94
MTGNN	18.21	30.99	12.36	21.48	34.91	17.17	21.75	35.35	14.88	19.91	32.63	15.11	12.48	23.39	19.50
STNorm	19.36	32.14	12.86	21.99	35.28	17.17	21.84	35.00	12.99	20.37	33.13	15.04	12.03	22.91	18.21
STID	18.03	30.85	12.18	20.65	34.29	16.92	20.40	33.90	12.97	19.04	31.86	14.69	11.62	22.41	19.84
RPMixer	26.01	43.64	18.32	28.84	52.59	26.88	28.55	51.95	19.00	25.44	47.93	20.64	16.68	43.64	32.74
BigST	17.68	29.61	11.66	21.15	34.38	17.80	20.98	34.40	13.30	19.32	32.01	14.93	12.13	23.01	21.42
GSNet	18.75	31.30	12.67	21.88	35.38	18.04	21.31	34.75	13.46	19.60	32.24	15.30	13.35	24.87	27.09
STWave	17.64	29.61	11.83	20.56	33.58	15.14	20.22	33.03	12.38	20.67	33.12	15.76	-	-	-
STAEformer	19.02	31.78	12.65	21.30	34.56	17.63	-	-	-	-	-	-	-	-	-
D^{2}STGNN	17.13	28.60	12.15	21.13	34.09	16.08	-	-	-	-	-	-	-	-	-
PatchSTG	17.46	30.13	11.74	19.75	33.17	14.98	19.30	32.28	11.38	17.68	29.72	12.86	10.63	20.86	19.41
Ours	**16.29**	**28.04**	**10.87**	**19.58**	**32.79**	**14.24**	**18.90**	**31.58**	**11.25**	**17.37**	**29.37**	**12.47**	**10.24**	**20.48**	**17.92**

Method	Electricity			UrbanEV			KnowAir			China City Air Quality			Beijing Weibo		
	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE
STGCN	240.2	2210	14.14	5.91	12.34	19.17	15.77	24.25	57.44	19.56	33.34	28.48	0.8549	1.6861	34.81
DGCRN	250.3	2353	18.14	5.22	11.47	18.70	21.11	30.62	65.89	21.87	35.18	35.05	0.8637	1.7842	31.55
AGCRN	211.5	1847	16.95	5.36	12.20	18.21	16.34	24.81	63.26	19.57	32.65	31.41	0.8505	1.6998	33.68
GWNet	200.3	1820	13.48	5.27	11.37	18.86	15.49	23.85	56.73	18.74	31.72	29.11	0.8315	1.6777	31.74
MTGNN	194.8	1583	16.53	5.27	11.31	18.40	15.74	24.21	58.70	19.62	32.58	30.70	0.8380	1.6653	32.59
STNorm	230.3	1983	14.92	5.43	11.54	19.24	16.00	24.32	59.46	19.72	33.13	30.04	0.8721	1.7228	32.15
STID	174.9	1532	12.48	5.23	11.39	18.24	16.16	24.88	61.41	20.54	34.13	32.86	0.8380	1.6730	32.40
RPMixer	188.6	1574	13.19	6.52	12.62	24.80	16.73	25.96	54.07	19.05	32.46	28.91	1.0190	1.8696	45.58
BigST	190.3	1632	13.85	5.43	11.23	19.79	15.68	24.15	56.52	18.67	31.02	29.37	0.8351	1.6806	31.32
GSNet	191.8	1617	14.98	5.55	11.39	20.26	16.30	24.68	60.37	19.50	32.04	31.29	0.8388	1.6762	32.39
STWave	188.2	1772	11.69	5.04	11.15	17.81	16.35	24.93	61.93	20.26	33.95	32.07	0.8308	1.6849	31.28
STAEformer	200.5	1650	13.75	5.01	11.16	17.64	15.82	24.56	53.28	19.01	31.57	30.34	0.8352	1.6810	32.12
D^{2}STGNN	224.8	2110	17.46	5.07	11.46	17.95	15.39	24.31	55.41	18.82	32.29	26.30	0.8489	1.7216	31.89
PatchSTG	/	/	/	5.16	11.53	17.89	16.08	24.70	56.78	18.98	32.17	29.13	0.8638	1.7561	32.16
Ours	**172.1**	**1499**	**11.57**	**4.95**	**11.00**	**17.43**	**15.36**	**23.42**	**52.77**	**18.52**	**30.88**	**26.13**	**0.7988**	**1.6477**	**29.85**

Experimental Results

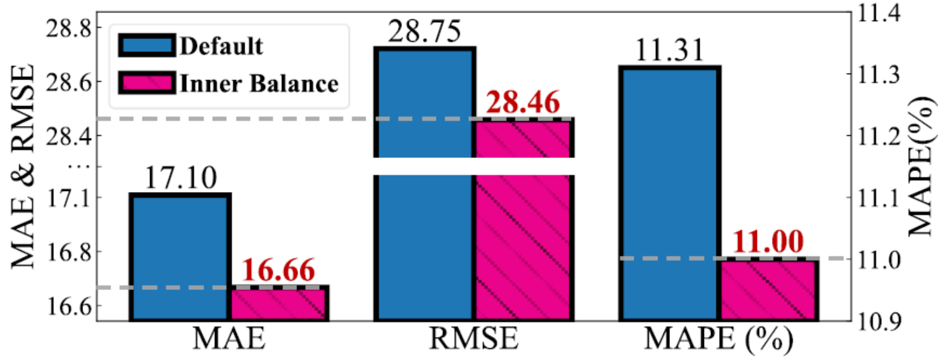
Table 2: Efficiency comparison with SOTA STGNNs. Memory: The maximum memory usage (MB) during training. BS: The maximum allowable batch size during training (up to 64). Train: Average Training Speed (s/epoch). ↑ indicates the relative percentage increasing regarding MAGE.

Method	SD (716)				GBA (2352)				UrbanEV (1682)			
	MAE	Memory	BS	Train	MAE	Memory	BS	Train	MAE	Memory	BS	Train
STAEformer	19.02↑16.75%	39,112↑968.05%	36↑43.75%	384↑1645%	21.30↑8.78%	39,518↑286.67%	5↑92.19%	2529↑4336.84%	5.09↑1.21%	33,680↑502.07%	4↑93.75%	745↑3625%
STWave	17.64↑8.28%	26,524↑624.30%	64↑0.00%	411↑1768%	20.56↑5.01%	40,564↑296.91%	26↑59.38%	1034↑1714.04%	5.04↑1.82%	38862↑594.70%	18↑71.88%	210↑950%
D^2STGNN	17.13↑5.15%	40,270↑999.67%	31↑51.56%	442↑1909%	21.13↑7.91%	39,102↑282.60%	3↑95.31%	5527↑9596.49%	5.12↑2.42%	39006↑597.28%	2↑96.875%	2257↑11185%
PatchSTG	17.46↑7.18%	7,612↑107.86%	64↑0.00%	101↑359%	19.75↑0.87%	27,852↑172.52%	64↑0.00%	326↑471.93%	5.16↑4.24%	12,106↑116.41%	64↑0.00%	25↑25%
Ours	**16.29**	**3,662**	**64**	**22**	**19.58**	**10,220**	**64**	**57**	**4.95**	**5594**	**64**	**20**

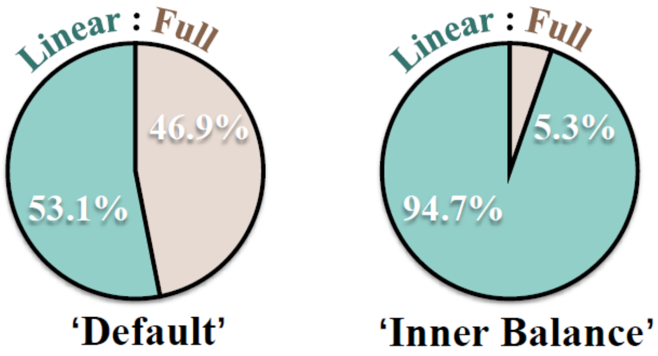
Experimental Results

Table 3: Pareto-optimal study of performance–efficiency trade-offs of adaptive graph type.

Linear : Full		MAE	RMSE	MAPE	Memory	Training
Naïve	0:16	16.52	28.35	10.91	4,308 MB	46 s/epoch
⇓	4:12	16.53	28.29	11.01	3,998 MB	35 s/epoch
	8:8	17.10	28.75	11.31	3,860 MB	30 s/epoch
	12:4	16.29	28.22	10.88	3,696 MB	24 s/epoch
Ours	16:0	**16.29**	**28.04**	**10.87**	**3,662** MB	**22** s/epoch



(a) Performance comparison



(b) True-ratio comparison

Figure 5: The comparison between 'default' load-balancing approach in MAGE and 'Inner balance' approach on the equal ratio setting 8:8.

中国科学技术大学
University of Science and Technology of China

数据智能实验室
Data Intelligence Lab

NEURAL INFORMATION PROCESSING SYSTEMS

◊ Connection & Cooperation

- ❖ Available Code: https://github.com/PoorOtterBob/MAGE.
- ❖ Contact Emails: JiamingMa@mail.ustc.edu.cn.
- ❖ Personal Websit: https://poorotterbob.github.io/.
- ❖ WeChat:

See you San Diego!!!