DiEP: Adaptive Mixture-of-Experts Compression through Differentiable Expert Pruning

Paper ID: 17474

Sikai Bai (Presenter), Haoxi Li, Jie Zhang, Zicong Hong, Song Guo

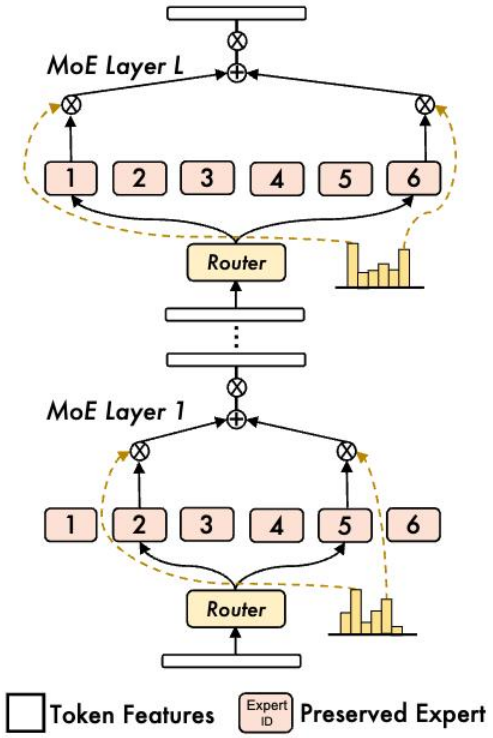
香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

NEURAL INFORMATION PROCESSING SYSTEMS

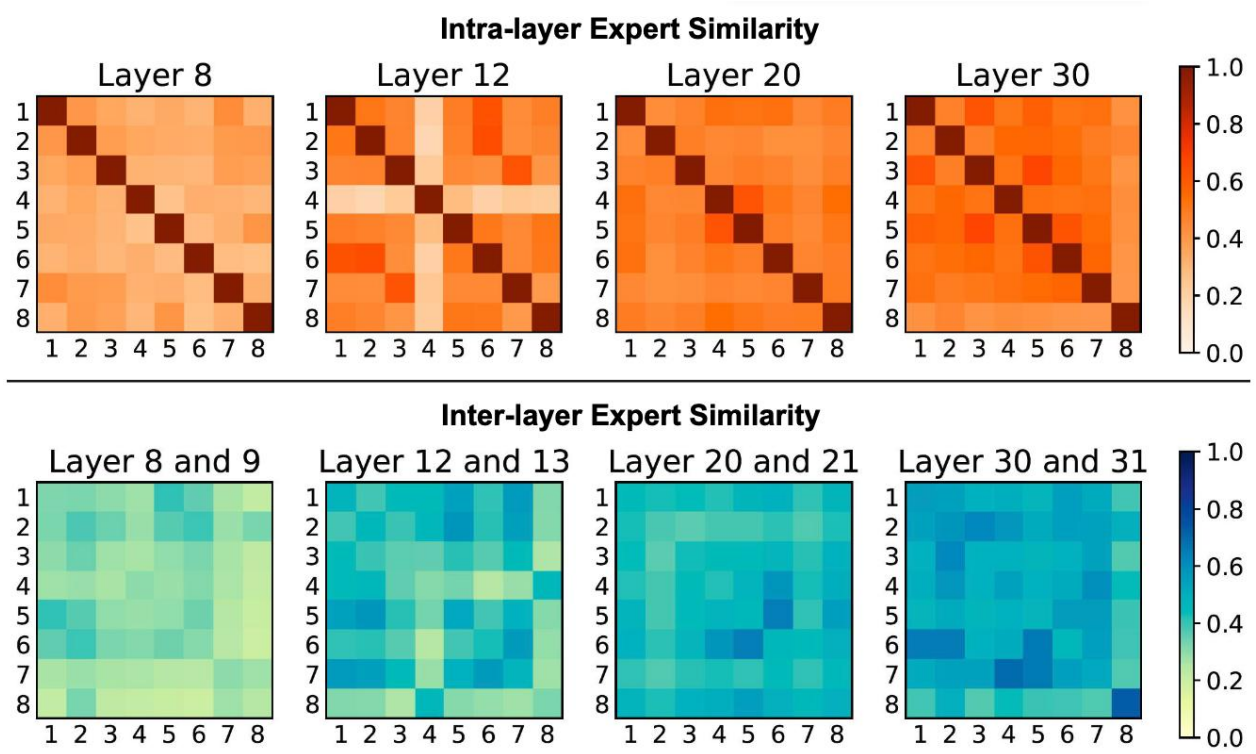
Table of Comments

Table of Comments

Motivation



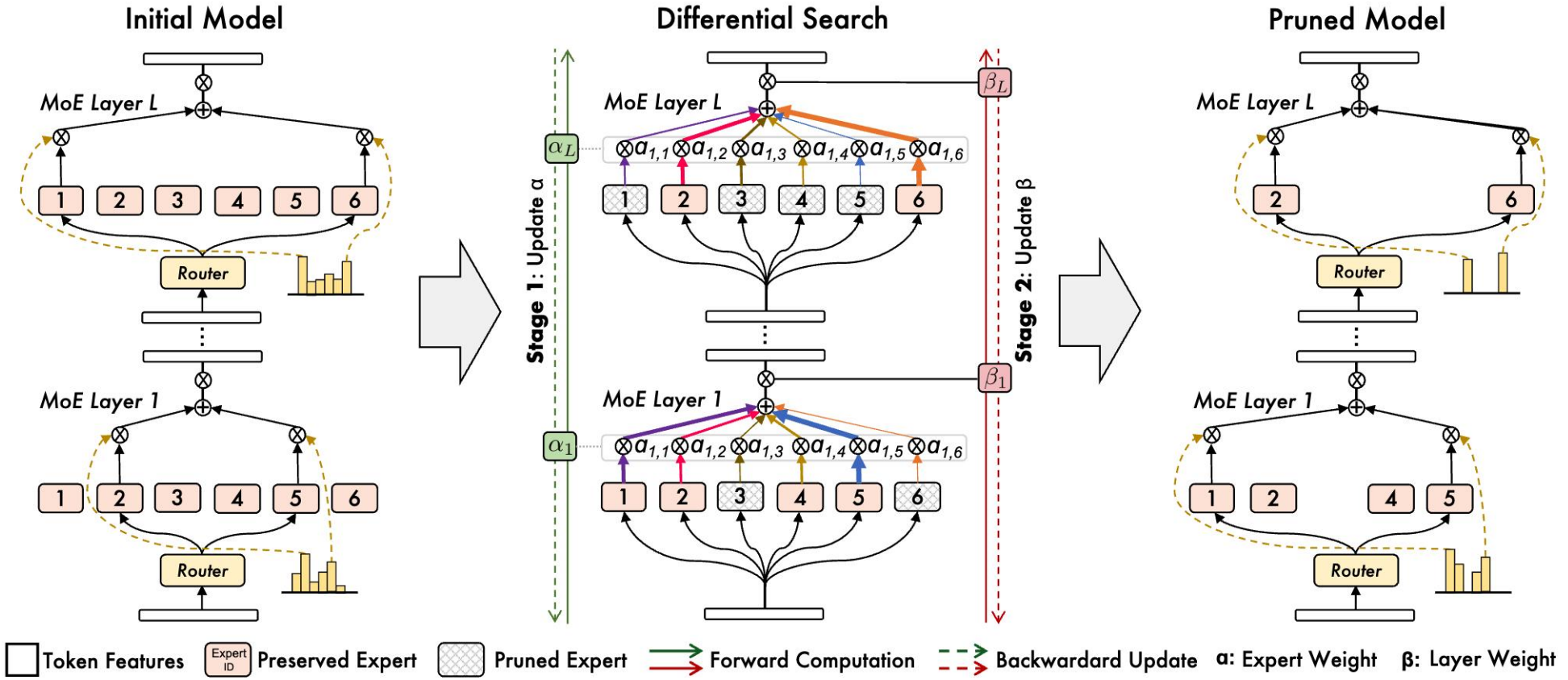
Mixture of Expert



Varying expert similarity across intra-layer and inter-layer.

Table of Comments

Our approach



The schematic illustration of the Differentiable Expert Pruning (DiEP) framework.

Our approach

Sparse Expert Search Space:

MoE layer is modeled as a directed acyclic graph (DAG) consisting of only two nodes.

- An input node representing the token representations entering the expert layer.
- An output node representing the sum of selected expert transformations.

Continuous Relaxation and Optimization:

Alternating Update Strategy.

$$\alpha^t \leftarrow \alpha^t - \eta_\alpha \nabla_\alpha \mathcal{L}(\alpha^t, \beta^t),$$
$$\beta^t \leftarrow \beta^t - \eta_\beta \nabla_\beta \mathcal{L}(\alpha^t, \beta^t).$$

Pruning Strategy.

$$s_i^{(l)} = \alpha_i^{(l)} \cdot \beta^{(l)}.$$

Adaptive Skipping During Inference:

- Assume experts with indices e_0 and e_1 are selected, with $w_{e1} < w_{e0}$.
- if $w_{e1} < \gamma w_{e0}$, expert e_1 is skipped, where $\gamma = \gamma_1 \times \gamma_2$.
- γ_1 is the median value of w_{e1}/w_{e0} across sampled calibration data for each MoE layer.
- γ_2 is the ratio of the CKA similarity $\rho(y_{e0}, y_{e1})$ to the mean CKA similarity $\rho(y_{ei}, y_{ej})$ across all data samples in layer l.

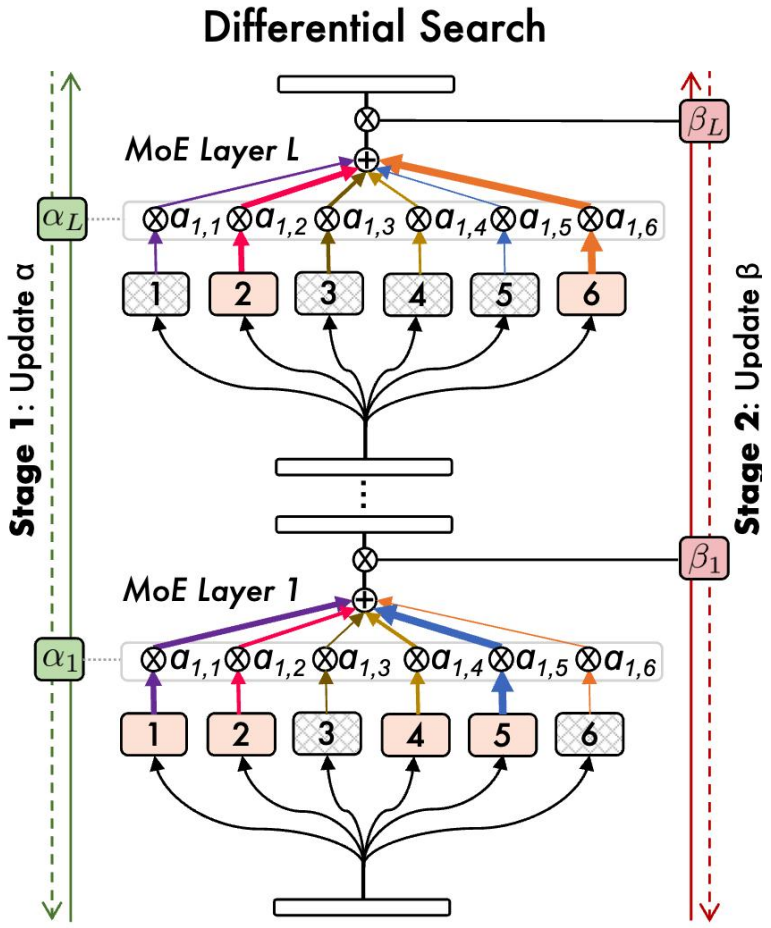
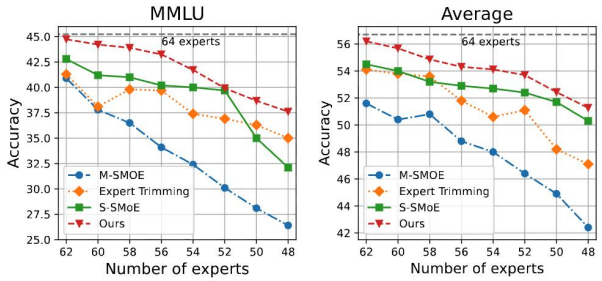


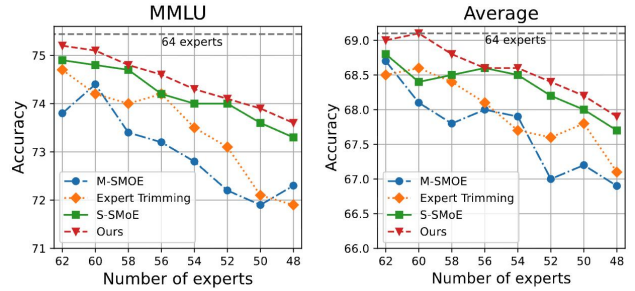
Table of Comments

Experiments

Model	Method $r = 25\%/50\%$	MMLU humanities	MMLU social science	MMLU stem	MMLU other	MMLU avg	BoolQ	OpenBookQA	RTE	Average
Mixtral 8×7B	Full	60.5	77.8	58.9	74.2	67.9	85.3	35.4	71.5	65.1
	M-SMOE	51.8/24.8	60.5/26.5	46.9/24.7	60.5/25.0	54.9/25.3	82.6/39.9	32.0/11.6	70.4/50.9	60.0/31.9
	Expert Trimming	49.2/36.9	59.7/45.6	45.0/35.1	58.2/43.4	54.1/45.7	77.2/76.6	33.0/26.4	56.6/55.9	55.2/51.2
	NAEE	52.4/43.5	66.4/52.7	49.0/40.4	63.7/43.5	58.7/47.3	84.0/80.8	32.6/28.8	67.9/61.4	60.8/54.6
	S-MOE	56.0/48.0	73.1/57.0	52.4/43.3	68.2/54.6	59.9/50.8	86.4/83.3	31.4/26.2	69.3/67.1	61.5/55.9
	DiEP(Ours)	**58.8/52.9**	**75.4/69.3**	**56.8/49.1**	**72.0/63.5**	**64.9/57.9**	**86.6/84.0**	**33.1/29.6**	**70.7/68.2**	**63.8/59.9**
Mixtral 8×7B -Instruct	Full	61.2	79.7	59.6	75.8	68.1	88.5	36.6	72.2	66.4
	M-SMOE	48.5/33.8	62.3/37.5	44.0/33.8	55.3/35.4	52.0/35.0	85.3/77.6	29.0/26.4	67.5/61.8	58.5/50.2
	Expert Trimming	52.9/45.0	74.3/61.1	50.5/39.2	64.8/50.8	58.6/47.3	86.3/83.0	37.0/32.3	63.2/66.8	61.3/57.3
	NAEE	55.9/48.7	69.5/55.6	54.1/42.3	68.7/56.2	62.4/52.8	87.3/84.8	35.6/30.4	70.0/**75.5**	63.8/60.9
	DiEP(Ours)	**61.2/55.1**	**78.1/72.3**	**59.4/53.4**	**73.8/67.8**	**67.3/61.3**	**87.7/85.6**	**35.9/31.0**	**72.2**/74.0	**65.8/63.0**
Mixtral 8×22B	Full	68.6	84.1	67.1	78.7	72.6	87.9	35.8	71.5	67.0
	M-SMOE	27.3/22.7	25.4/25.8	24.4/24.0	27.9/23.4	26.4/23.9	62.8/62.7	12.8/13.0	54.2/49.5	39.1/37.3
	Expert Trimming	58.0/45.7	74.9/57.7	54.1/42.0	70.2/45.7	64.3/47.8	81.5/74.4	35.2/27.0	69.3/57.4	62.6/51.7
	NAEE	60.4/53.9	78.0/67.2	59.5/52.3	73.0/64.2	67.7/59.4	87.4/80.5	35.0/31.1	70.1/67.9	65.1/59.7
	S-MOE	62.3/57.8	78.5/69.7	60.2/51.3	73.4/64.2	68.6/60.8	87.6/83.1	**35.8**/33.2	71.1/68.1	65.7/61.3
	DiEP(Ours)	**65.0/58.9**	**81.8/73.2**	**63.2/54.2**	**76.0/68.7**	**70.7/62.4**	**87.7/84.5**	**35.8/34.4**	**71.3/70.4**	**66.4/62.9**



(a) Deepseek-MoE-16B.



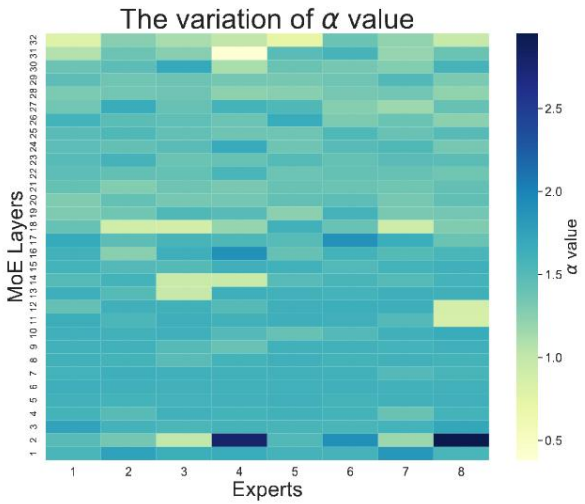
(b) Qwen2-57B-14A.

Zero-shot performance comparison on five advanced MoE models across various NLP tasks.

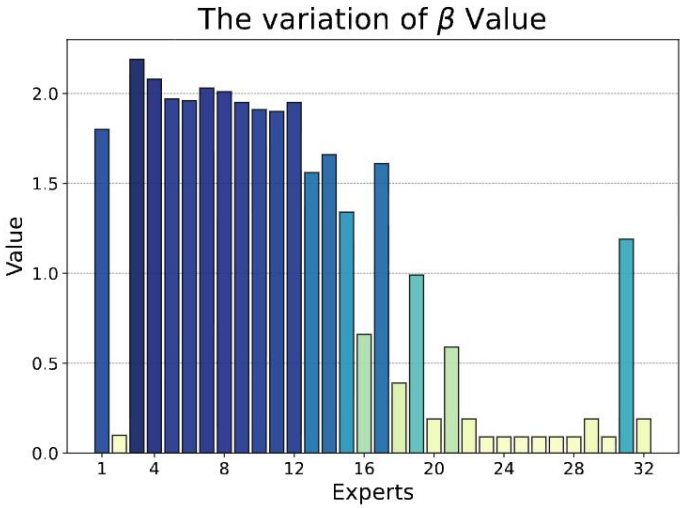
Experiments

Method	MMLU	BoolQ	OpenBookQA	RTE	Avg.
Baseline	58.7/47.3	84.0/80.8	32.6/28.8	67.9/61.4	60.8/54.6
W_α	60.5/51.0	86.0/82.8	32.2/27.8	67.5/65.3	61.6/56.7
W_β	61.0/51.4	85.1/83.3	32.0/29.6	67.3/66.2	61.3/57.6
$W_\alpha + W_\beta(random)$	57.6/49.2	85.6/83.4	32.3/27.2	66.4/62.1	60.47/55.5
$W_\alpha + W_\beta(1:2)$	55.1/46.2	81.5/77.4	30.6/26.8	66.4/64.2	57.8/54.2
$W_\alpha + W_\beta(2:1)$	63.3/54.2	85.4/83.5	32.6/29.8	68.2/67.5	62.4/58.8
$W_\alpha + W_\beta(3:1)$	64.6/55.2	85.9/84.2	32.8/29.6	69.7/67.8	63.3/59.2
DiEP(Ours)	**64.9/57.9**	**86.6/84.0**	**33.1/29.6**	**70.7/68.2**	**63.8/59.9**

Effectiveness of Components



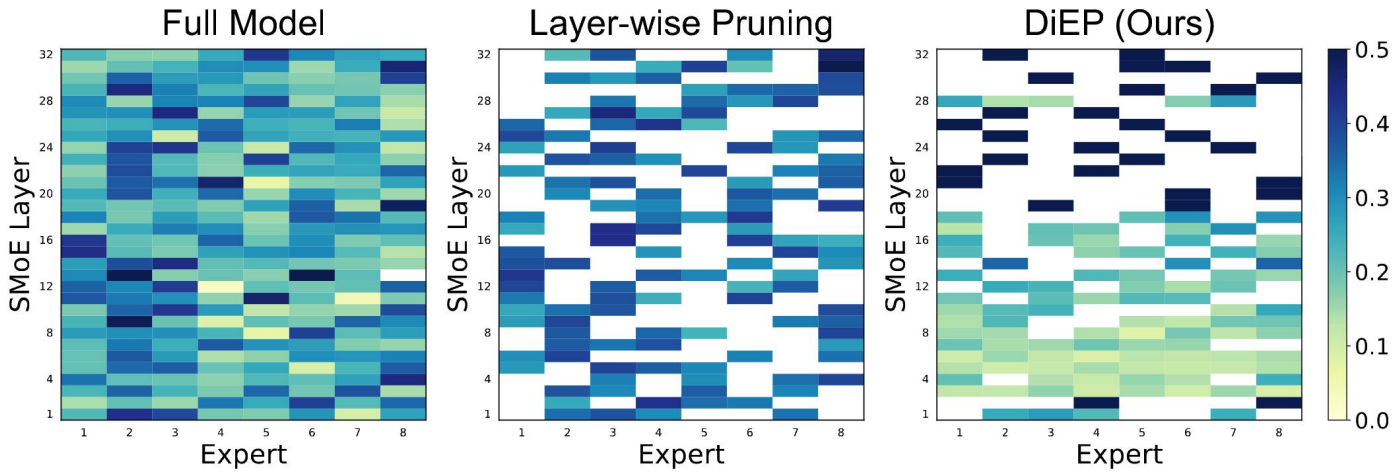
(a) Intra-layer scores α .



(b) Inter-layer scores β

Visualization analysis of values distribution for intra-layer scores α and inter-layer scores β

Experiments



Distribution of expert activation frequencies in the Mixtral 8×7B.

Method	Mixtral 8×7B	Mixtral 8×22B	Deepseek-MoE-16B	Qwen2-57B-14A
NAEE	1.31h	1.57h	≈ 94000d	≈ 113000d
DiEP(Ours)	**0.23h**	**0.31h**	**0.28h**	**0.34h**

Pruning time comparison of our DiEP and NAEE on different models

r	Pruning	Skipping	Avg. Acc	Speedup ↑	GPU ↓
0%			65.1	1.00×	1.00 ×
0%		✓	64.1	1.07 ×	1.00 ×
25%	✓		63.8	1.18×	0.76 ×
25%	✓	✓	63.3	1.21×	0.76 ×
50%	✓		59.9	1.26×	0.52 ×
50%	✓	✓	59.6	1.28 ×	0.52 ×

Model	r	Pruning	Skipping	Avg. Acc	Speedup ↑	GPU ↓
Deepseek-MoE-16B	0%			56.2	1.00×	1.00 ×
	0%		✓	55.7	1.04×	1.00 ×
	6.25%	✓		54.8	1.07×	0.95×
	6.25%	✓	✓	54.4	1.08×	0.95×
	12.5%	✓		54.1	1.11×	0.89×
	12.5%	✓	✓	53.6	1.13×	0.89×

Inference cost analysis on Mixtral 8×7B and Deepseek-MoE-16B after pruning

香港科技大學
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you for listening!

Sikai Bai
Email: whitesk1973@gmail.com