The 2024 Conference on Empirical Methods in Natural Language Processing
November 12-16, Miami, Florida

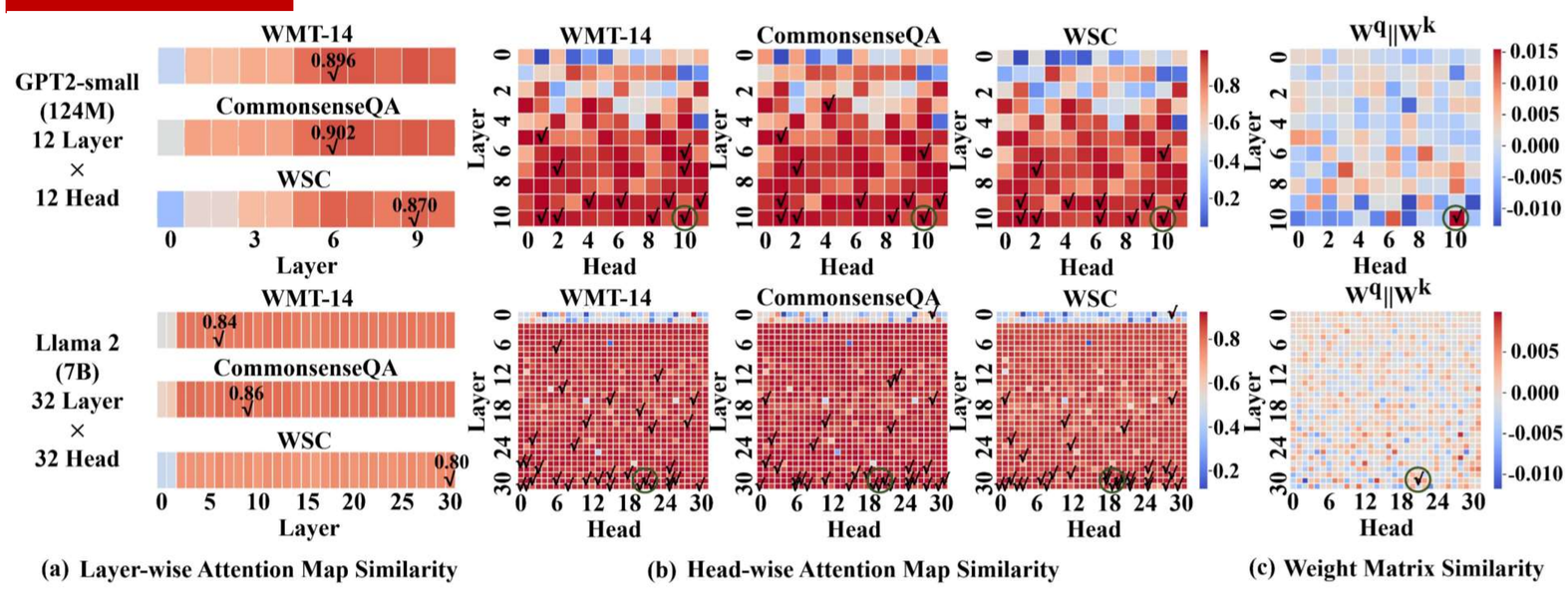
Head-wise Shareable Attention for Large Language Models

Zouying Cao[1,2,3], Yifei Yang[1,2,3], Hai Zhao[1,2,3,*]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
[3]Shanghai Key Laboratory of Trusted Data Circulation and Governance in Web3

Introduction

- We investigate the feasibility of **head-wise weight sharing** for large language models and propose two corresponding methods named **DirectShare** and **PostShare**.
- DirectShare is time-efficient and retain a large portion of the performance when sharing ratio is below 30%. Complementarily, PostShare yields satisfactory performance via post-training, especially under large ratios.

Motivation



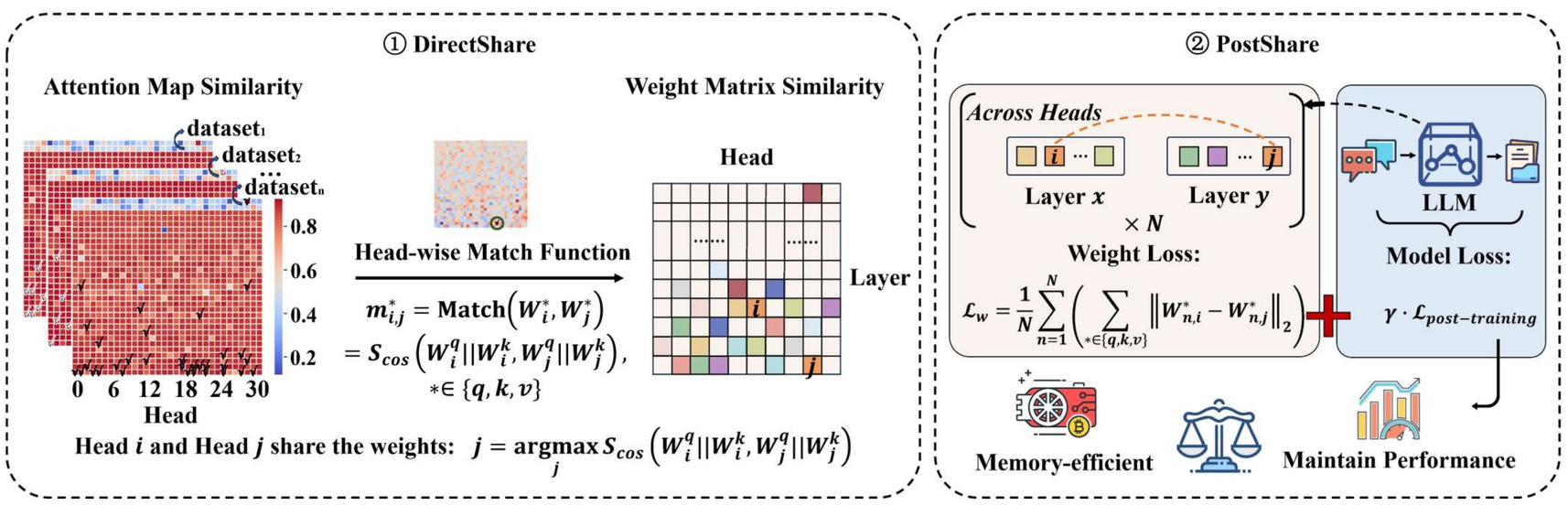
Methodology

① **Head-wise Weight Sharing Strategy**
choose the cosine similarity between the concatenation matrix of W_i^q and W_i^k

$$m_{i,j}^q = m_{i,j}^k = m_{i,j}^v = \mathcal{S}_{cos}(W_i^q||W_i^k, W_j^q||W_j^k) \quad (3)$$

② **DirectShare:**
directly share the weight matrices together between each selected attention head pairs

③ **PostShare:**
softly align model weights via post-training, pushing selected weight matrices more similar



Algorithm 1: DirectShare using Head-wise Weight Sharing Strategy

Input: Sharing ratio α, Original LLM $\mathcal{M}$, Number of layers $\mathcal{L}$, Number of heads per MHA block $\mathcal{H}$
Output: The LLM $\mathcal{M}^*$ after weight sharing

Initialize candidate buffer $\mathcal{D}_\tau$;
for $layer_i \leftarrow 2$ **to** $\mathcal{L}$ **do**
  for $i \leftarrow 1$ **to** $\mathcal{H}$ **do**
    $index_i \leftarrow (layer_i, i)$;
    $index_m \leftarrow$ None;
    $s_m \leftarrow$ -1;
    for $layer_j \leftarrow 1$ **to** $layer_i - 1$ **do**
      for $j \leftarrow 1$ **to** $\mathcal{H}$ **do**
        $index_m \leftarrow (layer_j, j)$;
        Compute $\mathcal{S}_{cos}$ using Eq. 3;
        if $\mathcal{S}_{cos} > s_m$ **then**
          $s_m \leftarrow \mathcal{S}_{cos}$
    Store candidate shareable head pair $< index_i, index_m, s_m >$ in $\mathcal{D}_\tau$;
Sort $\mathcal{D}_\tau$ by descending matching scores s_m;
$\mathcal{N} \leftarrow$ Top_N($\mathcal{D}_\tau$, $\mathcal{L}$, $\mathcal{H}$, α);
$\mathcal{M}^* \leftarrow$ Weight_Share($\mathcal{M}$, $\mathcal{N}$).

Experiments-Main Results

① DirectShare achieves comparable performance to the competitive model pruning methods.

Ratio	Method	CMNLI	OCNLI	AX-b	AX-g	RTE	RACE-middle	RACE-high	OBQA	CSL	TNEWS	Wino-Grande	BoolQ	C-Eval	MMLU
		Reasoning					NLU					Knowledge			
0%	**Llama2-7B**	32.98	33.12	53.53	55.34	49.82	33.15	35.51	31.80	55.62	20.22	54.04	70.67	32.20	46.69
10%	Magnitude	32.99	30.63	56.70	49.44	47.29	25.42	26.47	**28.20**	49.38	14.85	51.58	60.80	22.16	28.20
	LLM-Pruner	32.99	**33.75**	**57.61**	50.00	48.38	28.20	**30.73**	27.20	53.12	19.76	**52.98**	66.09	22.31	38.11
	DirectShare	**33.00**	32.50	54.17	**51.97**	**50.90**	**28.34**	28.96	**28.20**	**54.37**	**20.86**	52.63	**67.74**	**28.75**	**43.43**
30%	Magnitude	33.16	**35.00**	54.71	50.56	46.93	**21.80**	21.53	25.00	45.62	7.01	**50.88**	44.59	24.38	23.15
	LLM-Pruner	32.99	31.25	56.34	**52.53**	48.74	21.52	**22.21**	**26.80**	50.00	10.20	**50.88**	**54.77**	22.82	25.16
	DirectShare	**33.33**	32.50	**57.07**	51.69	**49.10**	21.45	21.53	26.00	**51.25**	**20.22**	50.18	54.43	**26.24**	**26.53**
0%	**Llama2-13B**	32.99	35.00	58.81	50.56	47.29	60.24	58.03	42.40	58.75	22.13	55.44	71.50	40.17	55.81
10%	Magnitude	32.82	33.12	51.99	**50.56**	**48.38**	22.42	21.78	27.40	51.25	15.39	49.82	62.32	22.52	27.54
	LLM-Pruner	**32.99**	**36.25**	**58.70**	50.00	46.93	51.46	50.80	**47.00**	56.25	**20.95**	**55.44**	68.07	30.25	51.45
	DirectShare	**32.99**	**36.25**	57.61	50.00	47.29	**54.04**	**55.63**	39.40	**56.88**	17.94	54.39	**69.45**	**37.17**	**52.81**
30%	Magnitude	**33.78**	33.75	46.65	50.00	**51.99**	21.80	22.01	**28.80**	46.25	4.19	49.12	56.45	**23.99**	22.86
	LLM-Pruner	32.99	34.38	57.16	**54.21**	45.85	23.96	25.33	26.40	53.75	**16.76**	**51.58**	**63.21**	22.17	27.22
	DirectShare	32.99	**35.00**	**58.33**	50.00	46.57	**26.53**	**27.53**	27.40	**59.38**	16.12	50.18	59.36	22.30	**30.79**

Table 1: Evaluation results of DirectShare based on the Llama2-7B and Llama2-13B models. **Bold** and underline indicate the best and the second best results.

② PostShare achieves more stable and satisfactory performance when reducing memory usage, with the cost of time increase.

Ratio	Method	WinoGrande	BoolQ	C-Eval	MMLU	RACE-middle	RACE-high	OBQA	OBQA-fact
0%	**Llama2-7B**	54.04	70.67	32.20	46.69	33.15	35.51	31.8	42.2
30%	DirectShare	50.18	54.43	26.24	26.53	21.45	21.53	26.00	27.60
	PostShare	52.98 ↑ 2.80	66.57 ↑ 12.14	26.38 ↑ 0.14	33.36 ↑ 6.83	29.81 ↑ 8.36	29.45 ↑ 7.92	27.60 ↑ 1.60	33.60 ↑ 6.00

Table 2: Overall Performance of PostShare based on Llama2-7B model.

Ablation Study

① Statistics of Memory Reduction

Sharing Ratio	#Params	GPU Memory
	Llama2-7B	
0%	6.74B/100%	17826M/100%
30% MHA	6.09B/90.36%	15512M/87.02%
30% MHA+FFN	4.74B/70.33%	12932M/72.55%
	Llama2-13B	
0%	13.02B/100%	30800M/100%
30% MHA	11.76B/90.32%	27898M/90.58%
30% MHA+FFN	9.21B/70.74%	23002M/74.68%

- When sharing 30% parameter sharing in the MHA block, our method achieves 10-13% memory.
- When we share 30% of parameter sharing in both MHA and FFN block, the model can save 26-28% GPU memory.

② Ablation on Head-wise Matching Functions

Sharing Ratio	5%		10%		15%		20%		25%		30%		35%	40%
Dataset	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	PIQA	OBQA	OBQA	OBQA
W^q	74.92	29.2	74.97	27.5	73.29	27.8	70.89	27.7	64.64	27.5	58.43	25.5	24.4	25.7
W^k	74.92	28.7	74.27	27.6	71.71	27.7	70.35	27.6	68.77	**27.6**	64.36	27.2	27.6	26.9
W^v	74.92	28.1	74.48	27.7	73.29	26.7	70.46	**28.5**	68.39	25.6	60.17	23.1	23.9	22.5
W^q, W^k, W^v	71.71	27.6	63.55	27.8	54.03	26.8	50.16	24.5	51.41	25.5	51.09	25.5	**29.0**	25.3
$W^q\|\|W^k\|\|W^v$	74.59	**34.7**	74.59	**30.0**	73.45	**30.3**	70.73	28.2	66.59	**27.6**	63.33	27.6	27.1	25.0
$W^q\|\|W^k$(Ours)	**75.84**	33.9	**75.30**	28.2	**74.54**	27.5	**73.01**	27.3	**69.37**	27.5	**65.56**	**28.0**	27.6	**28.6**

Conclusion

- **Head-wise Weight Sharing!**
- **Training-free, Effective!**
  —— **DirectShare**
- **Post-training-based, More Stable Performance!**
  —— **PostShare**

- **Contact us:**
zouyingcao@sjtu.edu.cn
yifeiyang@sjtu.edu.cn
zhaohai@cs.sjtu.edu.cn

- **Full paper**