Rectifying Magnitude Neglect in Linear Attention

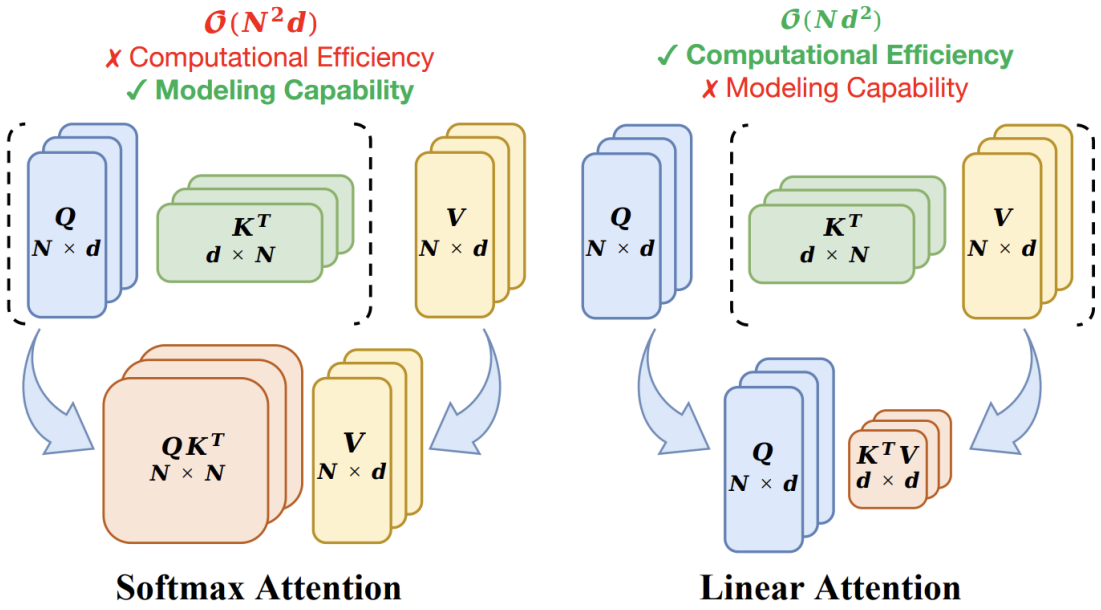
Qihang Fan

Institute of Automation, Chinese Academy of Sciences

Background

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Motivation:**
  - Softmax is too expansive, especially the sequence is very long (high resolution image, video ...).
  - Replace Softmax Attention with **Linear Attention**.

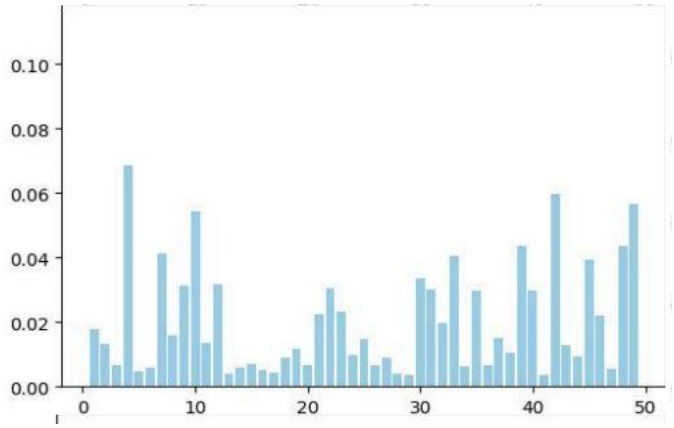


$$Y_i = \sum_{j=1}^{N} \frac{\phi(Q_i)\phi(K_j)^T}{\sum_{m=1}^{N} \phi(Q_i)\phi(K_m)^T} V_j$$

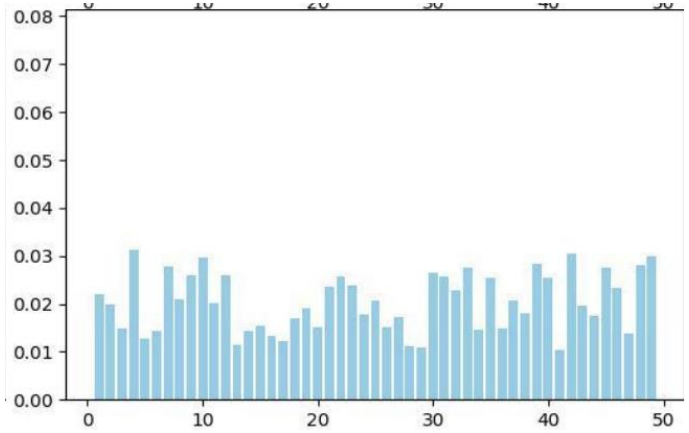
$$= \frac{\phi(Q_i)(\sum_{j=1}^{N} \phi(K_j)^T V_j)}{\phi(Q_i)(\sum_{m=1}^{N} \phi(K_m)^T)};$$

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Motivation:**
  - Linear Attention has poor performance.
  - **Over-smooth**



Softmax



Linear

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Motivation:**
  - Query's Magnitude impact the degree of concentration

$$\frac{\exp(Q_i K_m^T/\sqrt{d})}{\exp(Q_i K_n^T/\sqrt{d})} = p;$$

$$\frac{\exp(aQ_i K_m^T/\sqrt{d})}{\exp(aQ_i K_n^T/\sqrt{d})} = \frac{\exp(Q_i K_m^T/\sqrt{d})^a}{\exp(Q_i K_n^T/\sqrt{d})^a} = p^a$$

Softmax

$$Y_i = \frac{||\phi(Q_i)||\vec{\boldsymbol{\alpha}}_i(\sum_{j=1}^N \phi(K_j)^T V_j)}{||\phi(Q_i)||\vec{\boldsymbol{\alpha}}_i(\sum_{m=1}^N \phi(K_m)^T)} = \frac{\vec{\boldsymbol{\alpha}}_i(\sum_{j=1}^N \phi(K_j)^T V_j)}{\vec{\boldsymbol{\alpha}}_i(\sum_{m=1}^N \phi(K_m)^T)};$$

$$\frac{\phi(Q_i)\phi(K_m)^T}{\phi(Q_i)\phi(K_n)^T} = \frac{||\phi(Q_i)||\vec{\boldsymbol{\alpha}}_i\phi(K_m)^T}{||\phi(Q_i)||\vec{\boldsymbol{\alpha}}_i\phi(K_n)^T} = \frac{\vec{\boldsymbol{\alpha}}_i\phi(K_m)^T}{\vec{\boldsymbol{\alpha}}_i\phi(K_n)^T};$$

Linear

Method

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Method:**
  - Magnitude-Aware Linear Attention.

$$\mathrm{Attn}(Q_i, K_j) = \beta\phi(Q_i)\phi(K_j)^T - \gamma;$$

$$\beta = 1 + \frac{1}{\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T},$$

$$\gamma = \frac{\phi(Q_i)\sum_{m=1}^{N}\phi(K_m)^T}{N},$$

$$\sum_{j=1}^{N}\mathrm{Attn}(Q_i, K_j) = \beta\sum_{j=1}^{N}\phi(Q_i)\phi(K_j)^T - N\gamma = 1;$$

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Method:**
  - Magnitude-Aware Linear Attention.

$$\frac{\beta\phi(Q_i)\phi(K_m)^T - \gamma}{\beta\phi(Q_i)\phi(K_n)^T - \gamma} = p;$$

$$\beta_{new} = \frac{\beta + a - 1}{a},$$

$$\gamma_{new} = a\gamma;$$

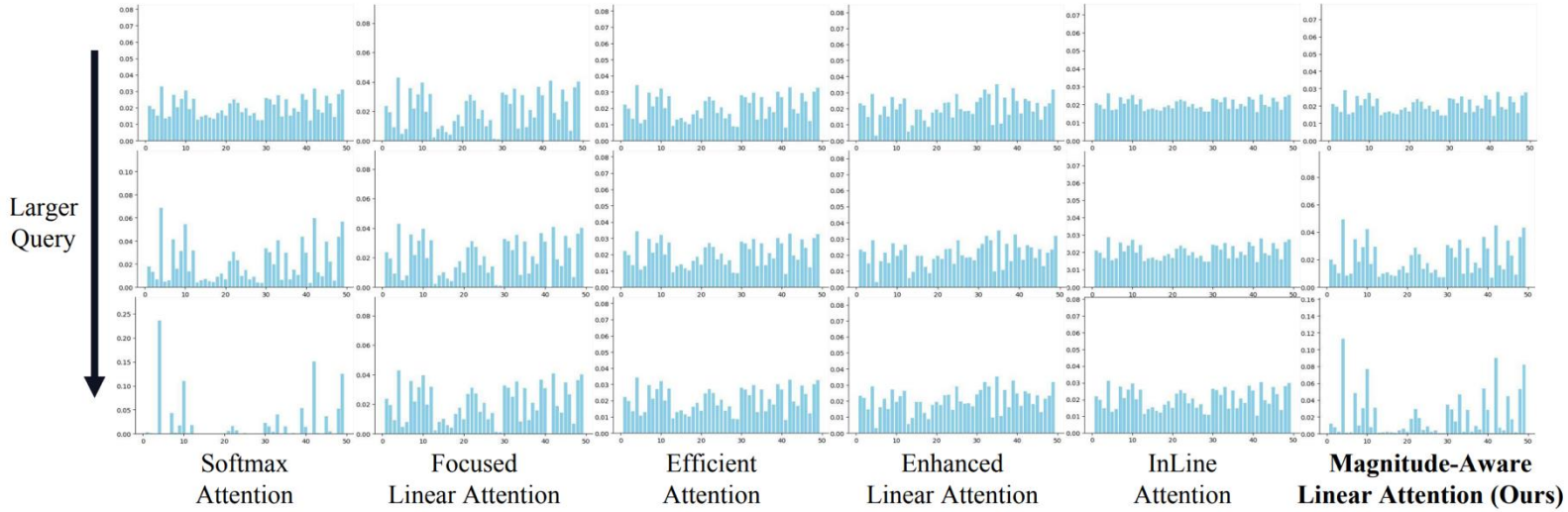
$$\frac{\beta_{new}a\phi(Q_i)\phi(K_m)^T - \gamma_{new}}{\beta_{new}a\phi(Q_i)\phi(K_n)^T - \gamma_{new}}$$

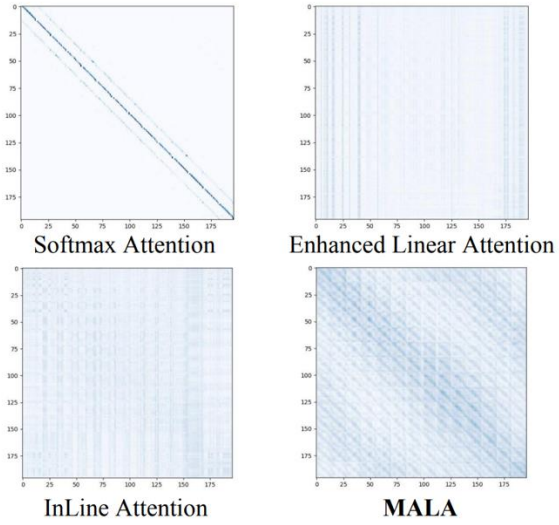
$$= \frac{\beta\phi(Q_i)\phi(K_m)^T - \frac{a\beta}{\beta+a-1}\gamma}{\beta\phi(Q_i)\phi(K_n)^T - \frac{a\beta}{\beta+a-1}\gamma} = p_m;$$

$$p_m > p$$

$$\frac{\exp(aQ_iK_m^T/\sqrt{d})}{\exp(aQ_iK_n^T/\sqrt{d})} = \frac{\exp(Q_iK_m^T/\sqrt{d})^a}{\exp(Q_iK_n^T/\sqrt{d})^a} = p^a$$

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*





Experiments

[ICCV'25] *Rectifying Magnitude Neglect in Linear Attention*

- **Vision & Language & Audio**

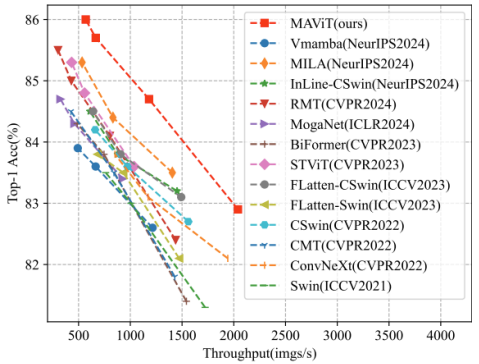


Figure 5. Comparison of general backbones' inference speed on low resolution task (image classification, resolution 224 × 224). The inference speed are measured on A100, batch size 64.

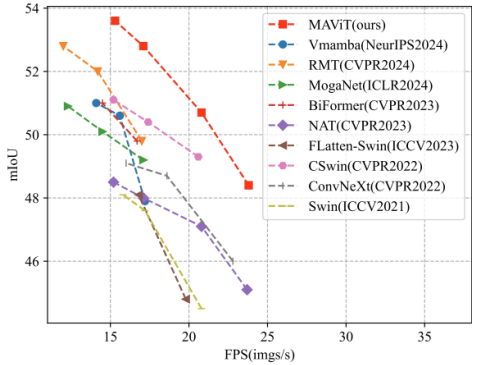


Figure 6. Comparison of general backbones' inference speed on high resolution task (semantic segmentation with UperNet, resolution 512 × 2048). The inference speed are measured on A100, batch size 1.

Model	LMB↑	PIQA↑	Hella↑	Wino↑	ARC-e↑	ARC-c↑	Avg↑
Transformer	31.0	63.3	34.0	50.4	44.5	24.2	41.2
RetNet	28.6	63.5	33.5	52.5	44.5	23.4	41.0
GLA	30.3	64.8	34.5	51.4	45.1	22.7	41.5
MALA	31.0	65.0	34.5	51.9	45.4	23.6	**41.9**

Table 1. MALA in NLP.

Model	Params	WER Without LM		WER With LM	
		testclean↓	testother↓	testclean↓	testother↓
Conformer(S)	10.3	2.7	6.3	2.1	5.0
Linear Attn	10.3	3.4	10.2	2.6	7.3
InLine Attn	10.3	3.1	9.6	2.5	7.3
MALA	10.3	**2.4**	**5.3**	**1.9**	**4.2**

Table 2. MALA in speech recognition.

Model	FLOPs	Throughput↑	FID↓	IS↑
DiT-S/2(400K) [41]	250×6.06G	4.9imgs/s	68.40	–
DiG-S/2(400K) [58]	250×4.30G	3.8imgs/s	62.06	22.81
DiC-S/2(400K) [48]	250×5.90G	–	58.68	25.82
MALA (400K)	**250×4.26G**	**5.6imgs/s**	**49.62**	**32.18**

Table 8. MALA for diffusion.

Thanks!