中国人民大學 RENMIN UNIVERSITY OF CHINA

高瓴人工智能学院 Gaoling School of Artificial Intelligence

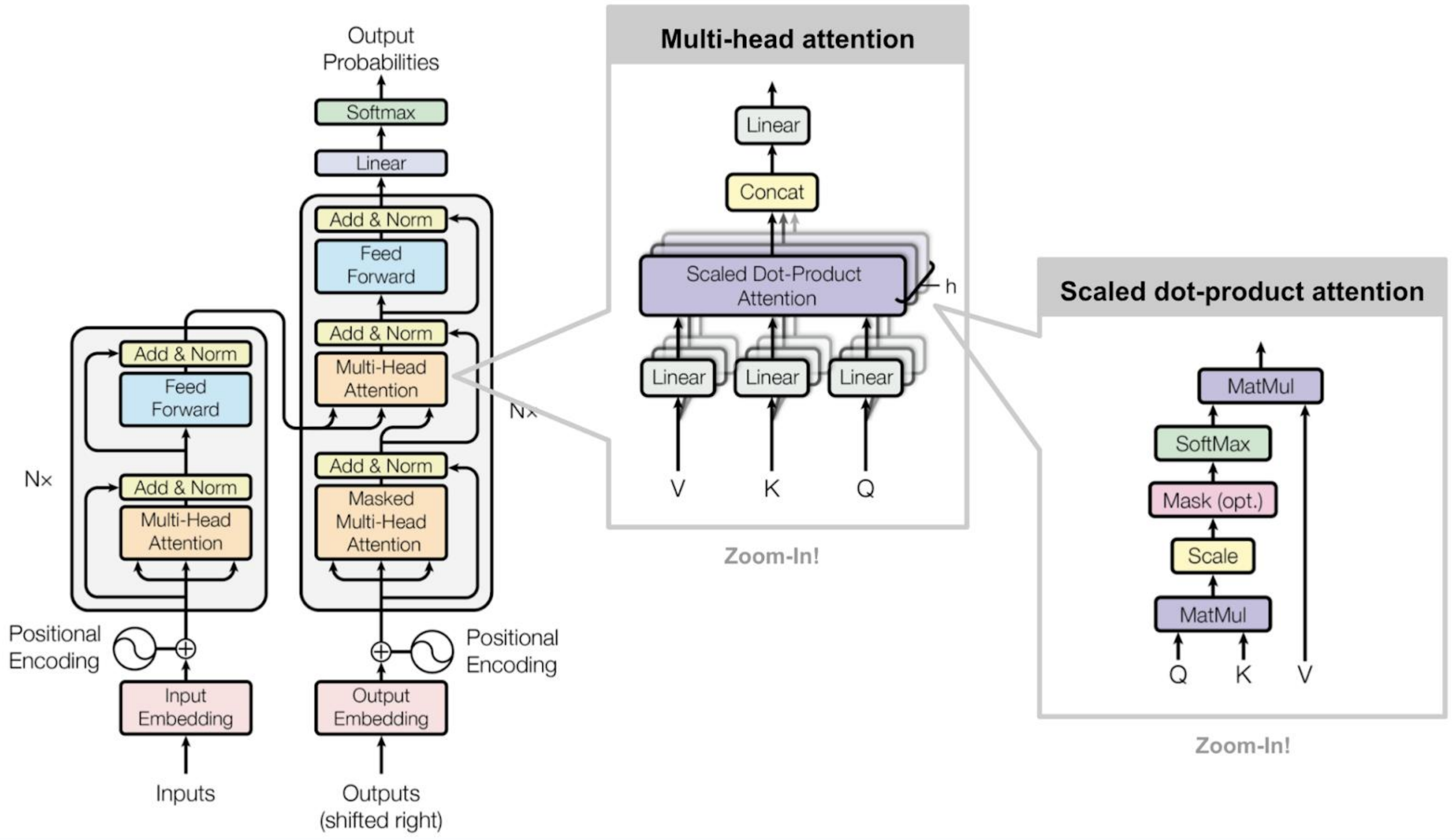
Paper Sharing

MEGA：MOVING AVERAGE EQUIPPED GATED ATTENTION

Lecturer：Yuxin Wu

2023.6.14

Transformer Architecture



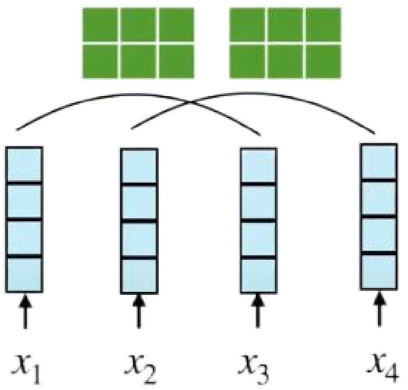
Limitations of Attention Mechanism

- **Weak Inductive Bias**
  - Almost no prior knowledge of dependency patterns
    - Learning directly from data
  - Position information only from absolute/relative positional embeddings

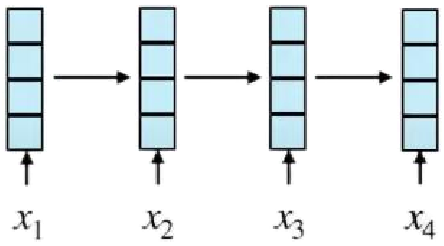
- **Quadratic Complexity**
  - Both time and space
  - $O(hn^2)$: h heads and sequence length of n

Inductive Bias: CNN & RNN

CNN



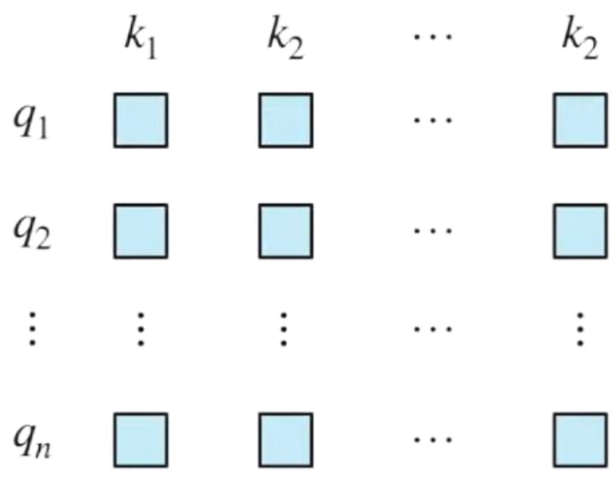
RNN



- ➢ Local dependencies
  - Window size is usually small (e.g. 3 or 5)
- ➢ Time-invariant kernel

- ➢ Sequencial dependencies
- ➢ Time-invariant recurrence

Inductive Bias: Attention

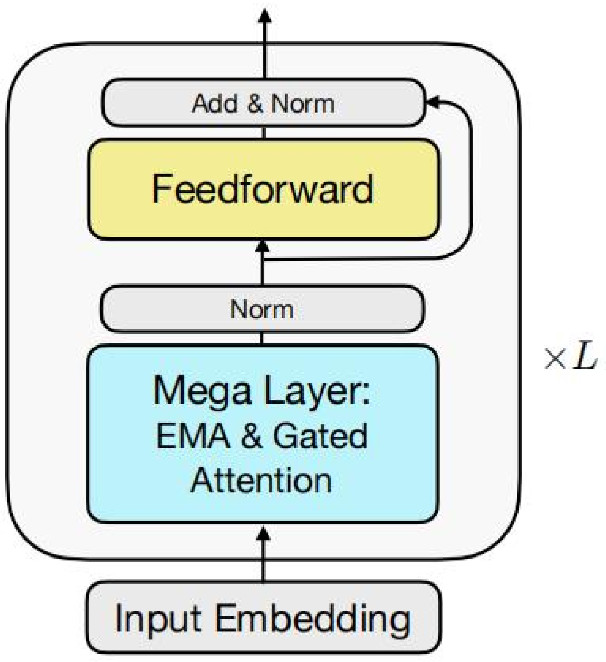


- Pair-wise interaction
- Order-invariant if no positional embeddings

Neither accurate nor efficient for long sequence modeling

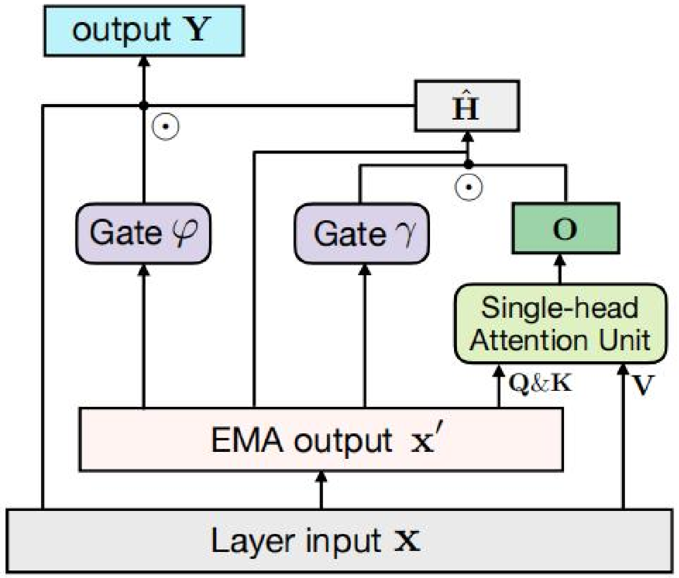
Mega：Overview

- **Effective and efficient drop-in replacement of attention for long sequence modeling**
  - Outstanding results on various types
    - ➢ text, images and audios
  - Exponential Moving Average (EMA)
  - Mega-chunk: linear complexity of time and space



Mega Architecture: Outline

- **Exponential Moving Average (EMA)**
  - Local dependencies that decaying exponentially over time
- **Single-head Gated Attention**
  - Adding a reset gate to the attention output
  - Theoretically proving that single-head gated attention is as expressive as multi-head one
- **Mega-Chunk**
  - Applying attention to local chunks of fixed length
  - Reducing quadratic complexity to linear

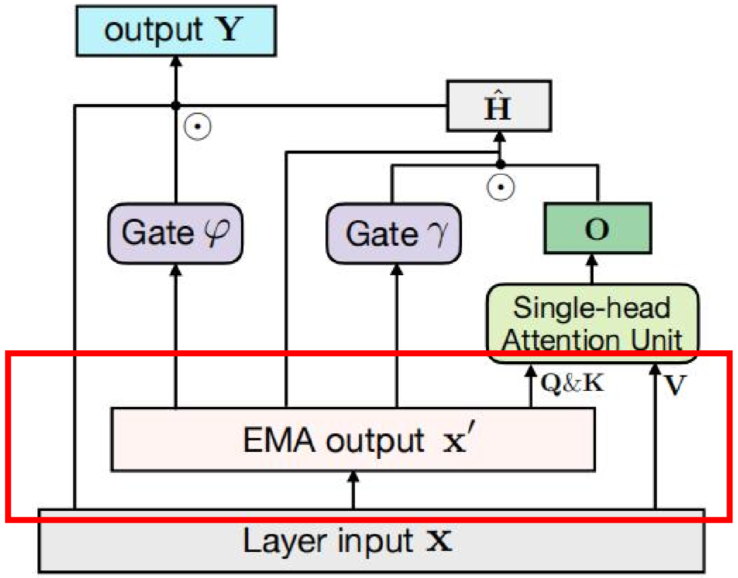


Mega Architecture: Outline

- **Exponential Moving Average (EMA)**
  - Local dependencies that decaying exponentially over time

- **Single-head Gated Attention**
  - Adding a reset gate to the attention output
  - Theoretically proving that single-head gated attention is as expressive as multi-head one

- **Mega-Chunk**
  - Applying attention to local chunks of fixed length
  - Reducing quadratic complexity to linear



Exponential Moving Average (EMA)

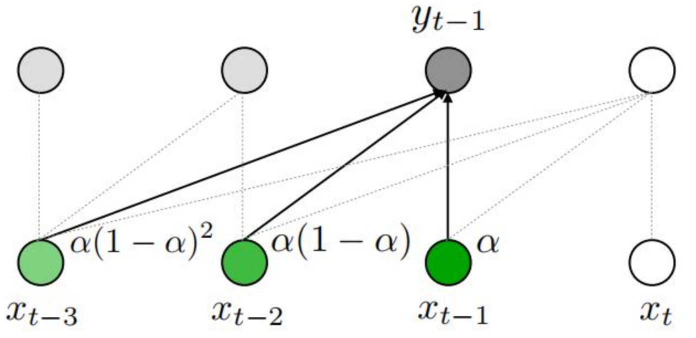
- **Notations:** Assuming 1-dim input sequence X = [X1, X2, ..., Xn], X∈R

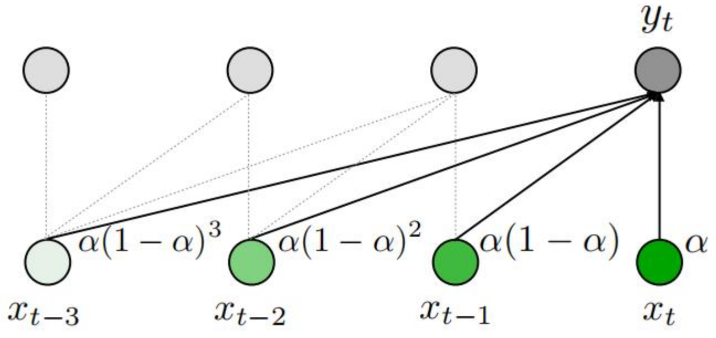
- **EMA:** $\mathbf{y}_t = \boldsymbol{\alpha} \odot \mathbf{x}_t + (1 - \boldsymbol{\alpha}) \odot \mathbf{y}_{t-1} \qquad \boldsymbol{\alpha} \in (0, 1)$

Learnable Parameters

- **Damped EMA:** $\mathbf{y}_t = \boldsymbol{\alpha} \odot \mathbf{x}_t + (1 - \boldsymbol{\alpha} \odot \boldsymbol{\delta}) \odot \mathbf{y}_{t-1} \qquad \boldsymbol{\delta} \in (0, 1)$

Relaxing the coupled weights





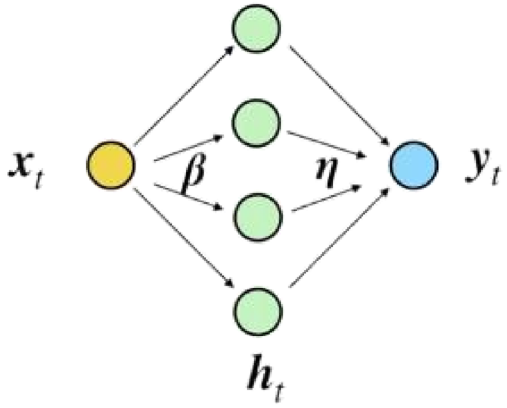
Multi-dimentional Damped EMA

- **Expanding** Xt **to** h **dimensions**

$$\mathbf{u}_t = \beta \mathbf{x}_t \in \mathbb{R}^h$$

- **Applying damped EMA individually to each dimension**

$$\mathbf{h}_t = \boldsymbol{\alpha} \odot \mathbf{u}_t + (1 - \boldsymbol{\alpha} \odot \boldsymbol{\delta}) \odot \mathbf{h}_{t-1} \in \mathbb{R}^h$$



- **Mapping the** h**-dimensional vector back to 1 dimension**

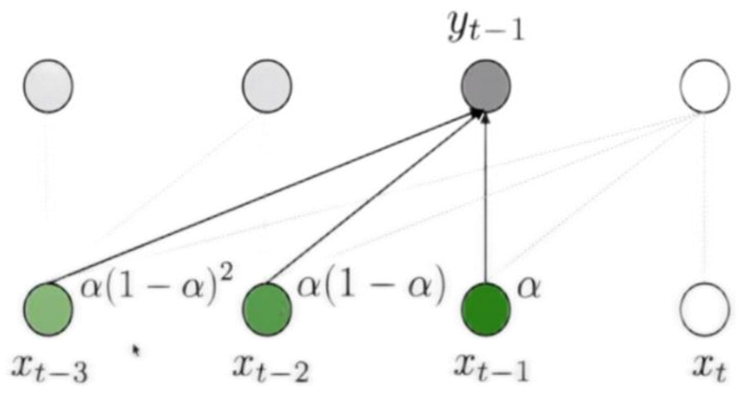
$$\mathbf{y}_t = \eta^T \mathbf{h}_t \in \mathbb{R}$$

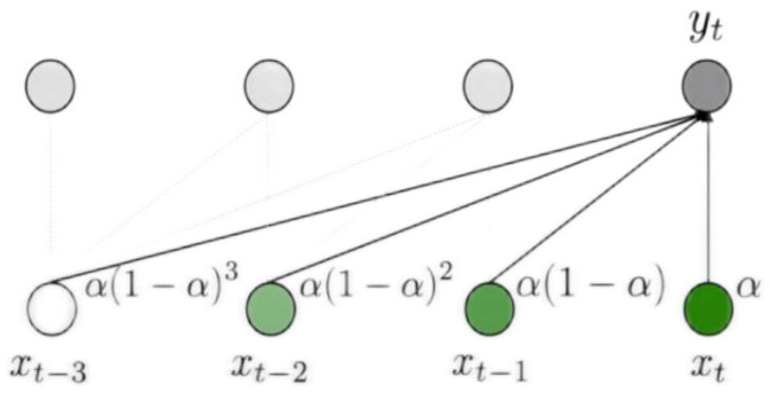
Efficient computation of EMA

- Efficiently compute EMA outputs of all tokens in parallel

$$\mathbf{y}_t = \boxed{\boldsymbol{\alpha}} \odot \mathbf{x}_t + \boxed{(1 - \boldsymbol{\alpha})} \odot \mathbf{y}_{t-1}$$

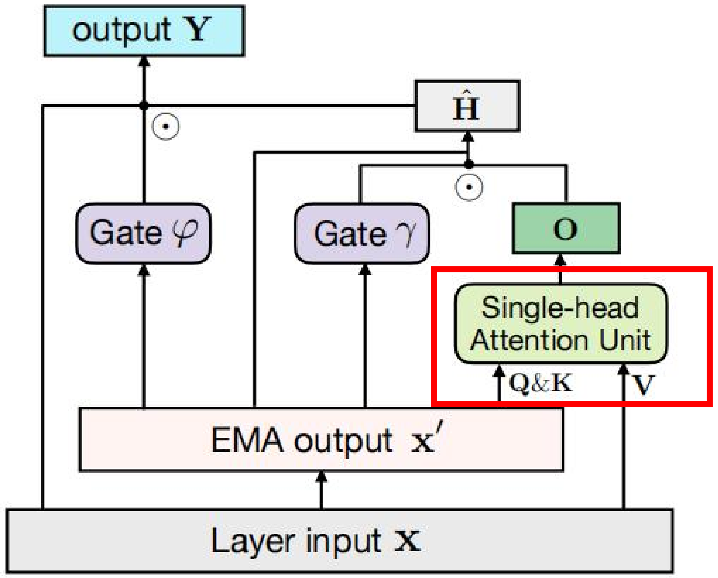
EMA weights are input independent





Mega Architecture: Outline

- Exponential Moving Average (EMA)
  - Local dependencies that decaying exponentially over time
- **Single-head Gated Attention**
  - Adding a reset gate to the attention output
  - Theoretically proving that single-head gated attention is as expressive as multi-head one
- Mega-Chunk
  - Applying attention to local chunks of fixed length
  - Reducing quadratic complexity to linear



Single-head Gated Attention

- Gated Attention Unit (GAU; Hua et al. (2022)) as the backbone

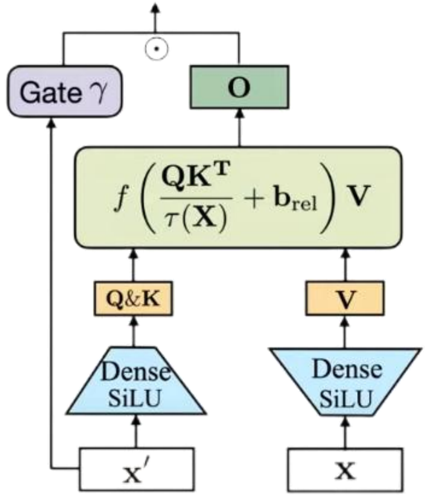
- first use the output from the EMA to compute the shared representation in GAU

$$\boldsymbol{X}' = \mathrm{EMA}(\boldsymbol{X}) \qquad \in \mathbb{R}^{n \times d}$$
$$\boldsymbol{Z} = \phi_{\mathrm{silu}}(\boldsymbol{X}' W_z + b_z) \qquad \in \mathbb{R}^{n \times z}$$

- the query and key sequences are computed by applying per-dimension scalars and offsets to Z, and the value sequence is from the original X:

$$\boldsymbol{Q} = \boldsymbol{\kappa}_q \odot \boldsymbol{Z} + \boldsymbol{\mu}_q$$
$$\boldsymbol{K} = \boldsymbol{\kappa}_k \odot \boldsymbol{Z} + \boldsymbol{\mu}_k$$
$$\boldsymbol{V} = \phi_{\mathrm{silu}}(\boldsymbol{X} W_v + b_v)$$

Single-head Gated Attention



- output of attention:

$$\boldsymbol{O} = f\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\tau(\boldsymbol{X})} + \boldsymbol{b}_{\text{rel}}\right)\boldsymbol{V}$$

Subsequently, MEGA introduces the reset gate γ, the update gate φ, and computes the candidate activation output $\hat{\boldsymbol{H}}$:

$$\gamma = \phi_{\text{silu}}(\boldsymbol{X}'W_\gamma + b_\gamma) \qquad \in \mathbb{R}^{n\times v} \tag{12}$$

$$\varphi = \phi_{\text{sigmoid}}(\boldsymbol{X}'W_\varphi + b_\varphi) \qquad \in \mathbb{R}^{n\times d} \tag{13}$$

$$\hat{\boldsymbol{H}} = \phi_{\text{silu}}(\boldsymbol{X}'W_h + (\gamma \odot \boldsymbol{O})U_h + b_h) \qquad \in \mathbb{R}^{n\times d} \tag{14}$$

The final output $\boldsymbol{Y}$ is computed with the update gate φ:

$$\boldsymbol{Y} = \varphi \odot \hat{\boldsymbol{H}} + (1-\varphi) \odot \boldsymbol{X} \qquad \in \mathbb{R}^{n\times d} \tag{15}$$

Single-head Gated Attention

- Single-head gated attention is as expressive as multi-head one

$$\boldsymbol{O}_{\mathrm{SHA}} = \boldsymbol{a}^T \boldsymbol{V} = \begin{bmatrix} \boldsymbol{a}^T \boldsymbol{V}^{(1)} \\ \vdots \\ \boldsymbol{a}^T \boldsymbol{V}^{(h)} \end{bmatrix}, \quad \boldsymbol{O}_{\mathrm{MHA}} = \begin{bmatrix} \boldsymbol{a}^{(1)T} \boldsymbol{V}^{(1)} \\ \vdots \\ \boldsymbol{a}^{(h)T} \boldsymbol{V}^{(h)} \end{bmatrix} \tag{19}$$

It is straightforward to see that $\boldsymbol{O}_{\mathrm{MHA}}$ is more expressive than $\boldsymbol{O}_{\mathrm{SHA}}$, because $\boldsymbol{O}_{\mathrm{MHA}}$ leverages h sets of attention weights.

In the single-head gated attention, we introduce a gate vector $\boldsymbol{\gamma} = \mathcal{G}(\boldsymbol{X})$ for each $\boldsymbol{q}$, and the output of single-head gated attention is $\boldsymbol{O}_{\mathrm{SHGA}} = \boldsymbol{O}_{\mathrm{SHA}} \odot \boldsymbol{\gamma}$. The following theorem reveals the equivalence of $\boldsymbol{O}_{\mathrm{SHGA}}$ and $\boldsymbol{O}_{\mathrm{MHA}}$ w.r.t expressiveness (proof in Appendix B):

Theorem 1 *Suppose the transformation $\mathcal{G}$ is a universal approximator. Then, $\forall \boldsymbol{X}, \exists \boldsymbol{\gamma} = \mathcal{G}(\boldsymbol{X})$ s.t.*

$$\boldsymbol{O}_{\mathrm{SHGA}} = \boldsymbol{O}_{\mathrm{MHA}} \tag{20}$$

Theorem 1 indicates that by simply introducing the gate vector, $\boldsymbol{O}_{\mathrm{SHGA}}$ is as expressive as $\boldsymbol{O}_{\mathrm{MHA}}$. In practice, $\mathcal{G}$ is commonly modeled by a (shallow) neural network, whose universality of approximation has been extensively studied (Hornik et al., 1989; Yarotsky, 2017; Park et al., 2020).

Single-head Gated Attention

- Single-head gated attention is as expressive as multi-head one

Proof We split γ into h heads in the same way as $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$:

$$\gamma = \begin{bmatrix} \gamma^{(1)} \\ \vdots \\ \gamma^{(h)} \end{bmatrix}$$

Then we have

$$\boldsymbol{O}_{\text{SHGA}} = \boldsymbol{a}^T \boldsymbol{V} \odot \gamma = \begin{bmatrix} \boldsymbol{a}^T \boldsymbol{V}^{(1)} \odot \gamma^{(1)} \\ \vdots \\ \boldsymbol{a}^T \boldsymbol{V}^{(h)} \odot \gamma^{(h)} \end{bmatrix}$$

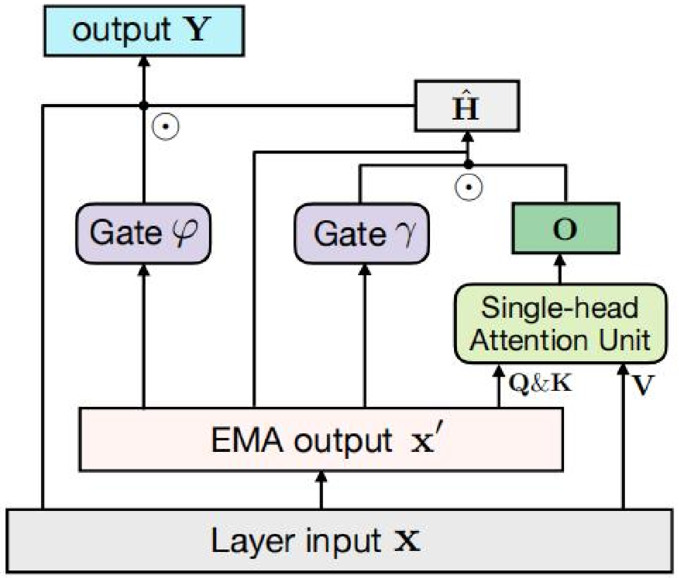
To prove Theorem 1, we need to find γ such that

$$\boldsymbol{a}^T \boldsymbol{V}^{(i)} \odot \gamma^{(i)} = {\boldsymbol{a}^{(i)}}^T \boldsymbol{V}^{(i)} \iff \gamma^{(i)} = {\boldsymbol{a}^{(i)}}^T \boldsymbol{V}^{(i)} \oslash \boldsymbol{a}^T \boldsymbol{V}^{(i)}, \ \forall i \in \{1, \ldots, h\},$$

where $\oslash$ is the element-wise divide operation. Since $\mathcal{G}(\boldsymbol{X})$ is a universal approximator and $\boldsymbol{Q}$, $\boldsymbol{K}$, $\boldsymbol{V}$ and $\boldsymbol{a}$ are all transformed from $\boldsymbol{X}$, γ can theoretically recover ${\boldsymbol{a}^{(i)}}^T \boldsymbol{V}^{(i)} \oslash \boldsymbol{a}^T \boldsymbol{V}^{(i)}$, $\forall \boldsymbol{X}$. ■

Mega Architecture: Outline

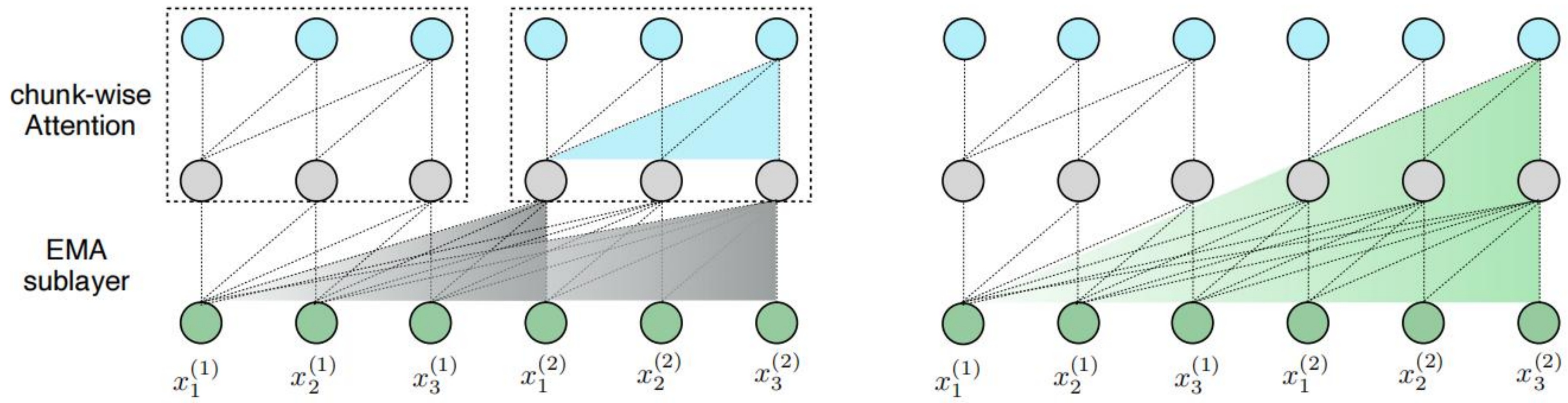
- Exponential Moving Average (EMA)
  - Local dependencies that decaying exponentially over time
- Single-head Gated Attention
  - Adding a reset gate to the attention output
  - Theoretically proving that single-head gated attention is as expressive as multi-head one
- **Mega-Chunk**
  - Applying attention to local chunks of fixed length
  - Reducing quadratic complexity to linear



Mega-Chunk: Effieicent Mega

- Split input sequences into multiple chunks with fixed length

- Applying attention individually to each chunk
  - ➢ Linear complexity & easy implementation
  - ➢ Losing contextual information between chunks
  - ➢ EMA preserves the information from previous chunks



Experiments

- **Long Range Arena (LRA)**
  - 3 tasks on byte-level text classification
  - 3 tasks on pixel-level image classification
- **Language Modeling**
  - Enwiki8 (character-level)
  - WikiText-103(Word-level)
- **Machine Translation**
  - WMT'14 English - German
- **Image Classification**
  - ImageNet-1K
- **Raw Speech Classification**
  - Speech Commands

Experimental Results

Table 2: (**Long Range Arena**) Accuracy on the full suite of long range arena (LRA) tasks, together with training speed and peak memory consumption comparison on the Text task with input length of 4K. ‡ indicates results replicated by us.

Models	ListOps	Text	Retrieval	Image	Pathfinder	Path-X	Avg.	Speed	Mem.
XFM	36.37	64.27	57.46	42.44	71.40	✗	54.39	–	–
XFM‡	37.11	65.21	79.14	42.94	71.83	✗	59.24	1×	1×
Reformer	37.27	56.10	53.40	38.07	68.50	✗	50.67	0.8×	0.24×
Linformer	35.70	53.94	52.27	38.56	76.34	✗	51.36	5.5×	0.10×
BigBird	36.05	64.02	59.29	40.83	74.87	✗	55.01	1.1×	0.30×
Performer	18.01	65.40	53.82	42.77	77.05	✗	51.41	**5.7**×	**0.11**×
Luna-256	37.98	65.78	79.56	47.86	78.55	✗	61.95	4.9×	0.16×
S4-v1	58.35	76.02	87.09	87.26	86.05	88.10	80.48	–	–
S4-v2	59.60	86.82	90.90	88.65	94.20	96.35	86.09	–	–
S4-v2‡	59.10	86.53	90.94	88.48	94.01	96.07	85.86	4.8×	0.14×
MEGA	**63.14**	**90.43**	**91.25**	**90.44**	**96.01**	**97.98**	**88.21**	2.9×	0.31×
MEGA-chunk	58.76	90.19	90.97	85.80	94.41	93.81	85.66	5.5×	0.13×

Experimental Results

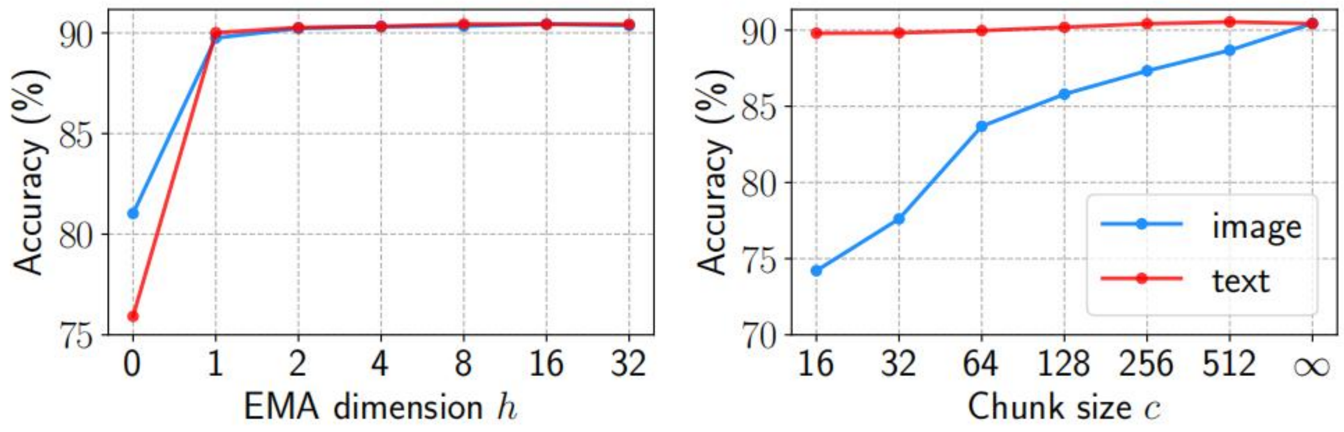


Figure 4: Ablations on EMA dimension and chunk size.

	Text	Image
softmax	**90.43**	89.87
relu^2	90.08	90.22
laplace	90.22	**90.43**

Table 3: Attention functions.

Thank you for listening

主講人：吳雨欣

2023.6.14