

Transformer 的競爭者們



Is Attention All You Need?



Current Status: Yes

Time Remaining: 656d 19h 39m 37s

Proposition:

On January 1, 2027, a Transformer-like model will continue to hold the state-of-the-art position in most benchmarked tasks in natural language processing.

For the Motion

Jonathan Frankle
@jefrankle
Harvard Professor
Chief Scientist Mosaic ML



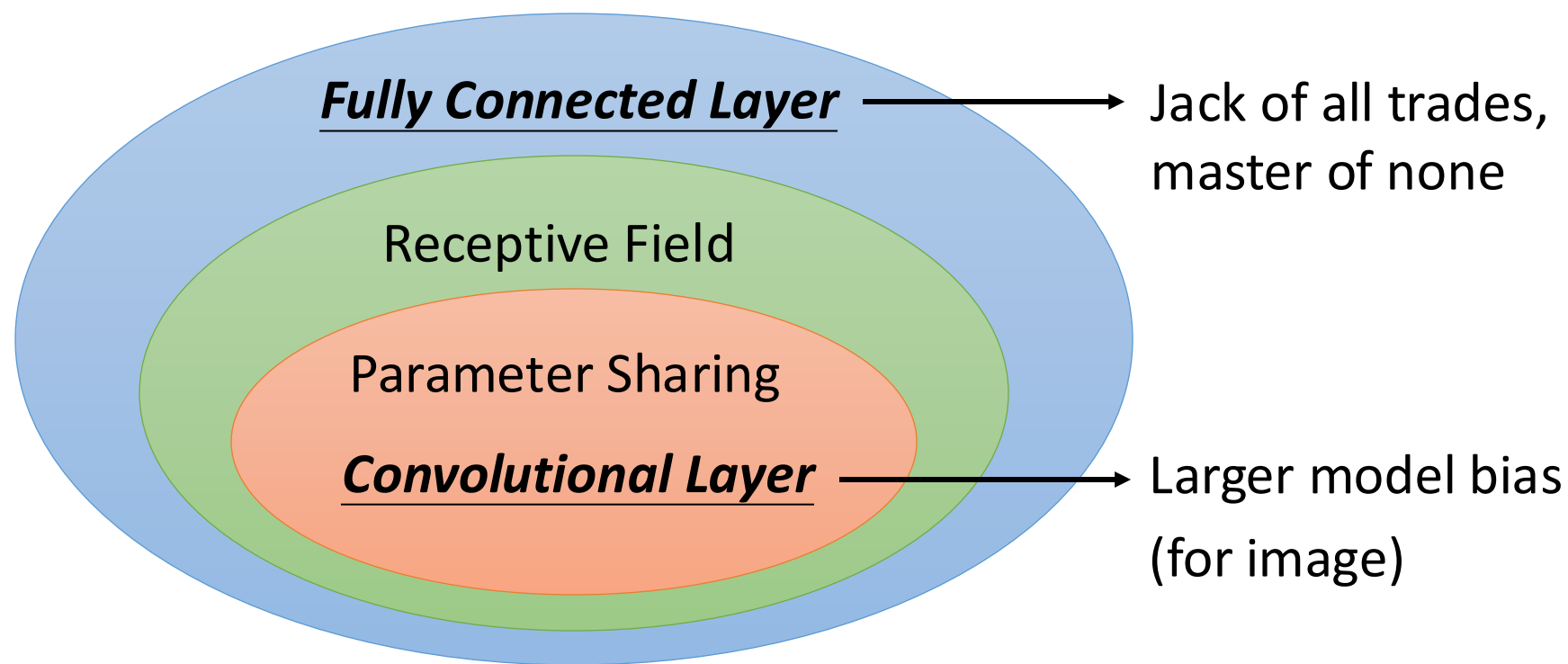
Against the Motion

Sasha Rush
@srush_nlp
Cornell Professor
Research Scientist Hugging Face 🤗



每一種架構的存在都有一個理由

- CNN 存在的理由是什麼？



根據影像的特性，減少需要的參數，避免 **Overfitting**

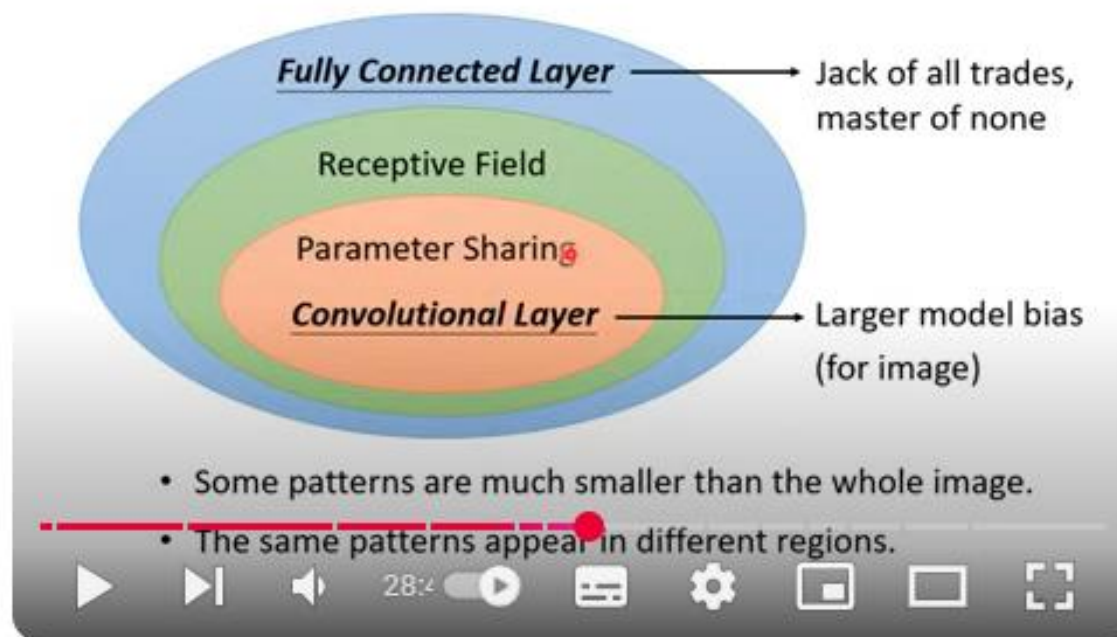
每一種架構的存在都有一個理由

- CNN 存在的理由是什麼？



<https://youtu.be/OP5HcXJg2Aw?si=RPfmHhsrMtuN0QS6>

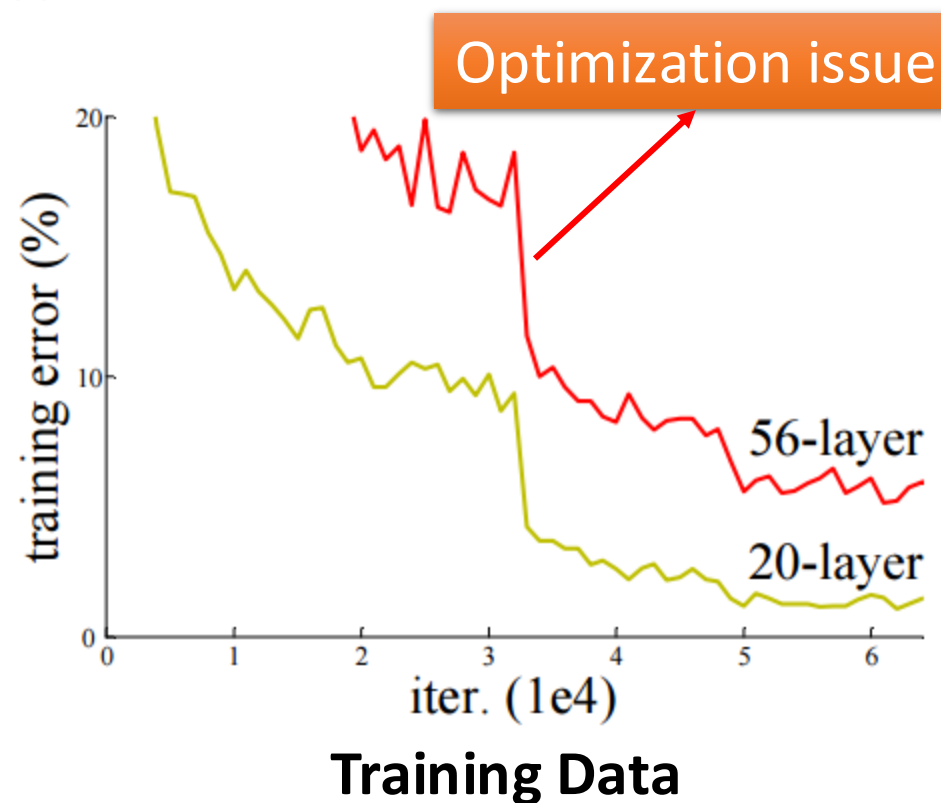
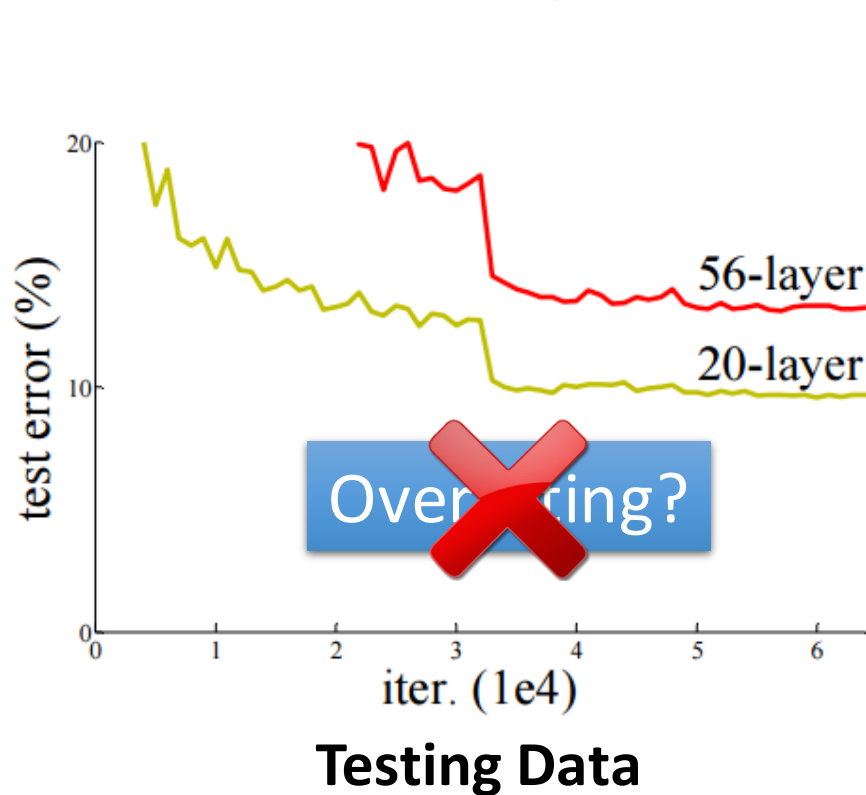
Benefit of Convolutional Layer



【機器學習2021】卷積神經網路 (Convolutional Neural Networks, CNN)

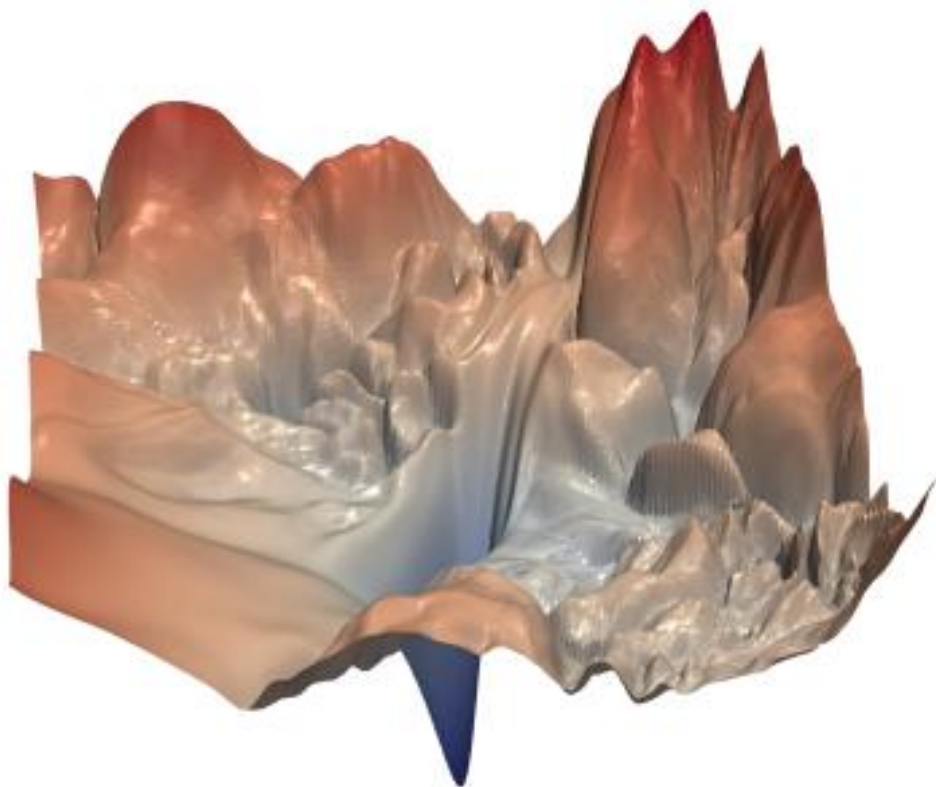
每一種架構的存在都有一個理由

- Residual Connection 存在的理由是什麼？

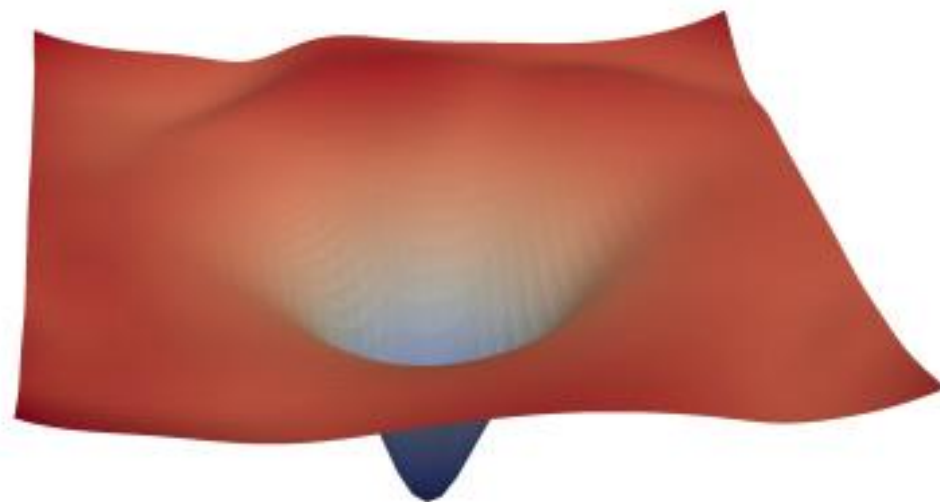


每一種架構的存在都有一個理由

- Residual Connection 存在的理由是什麼？為了讓 Optimization 可以做得更好



(a) without skip connections

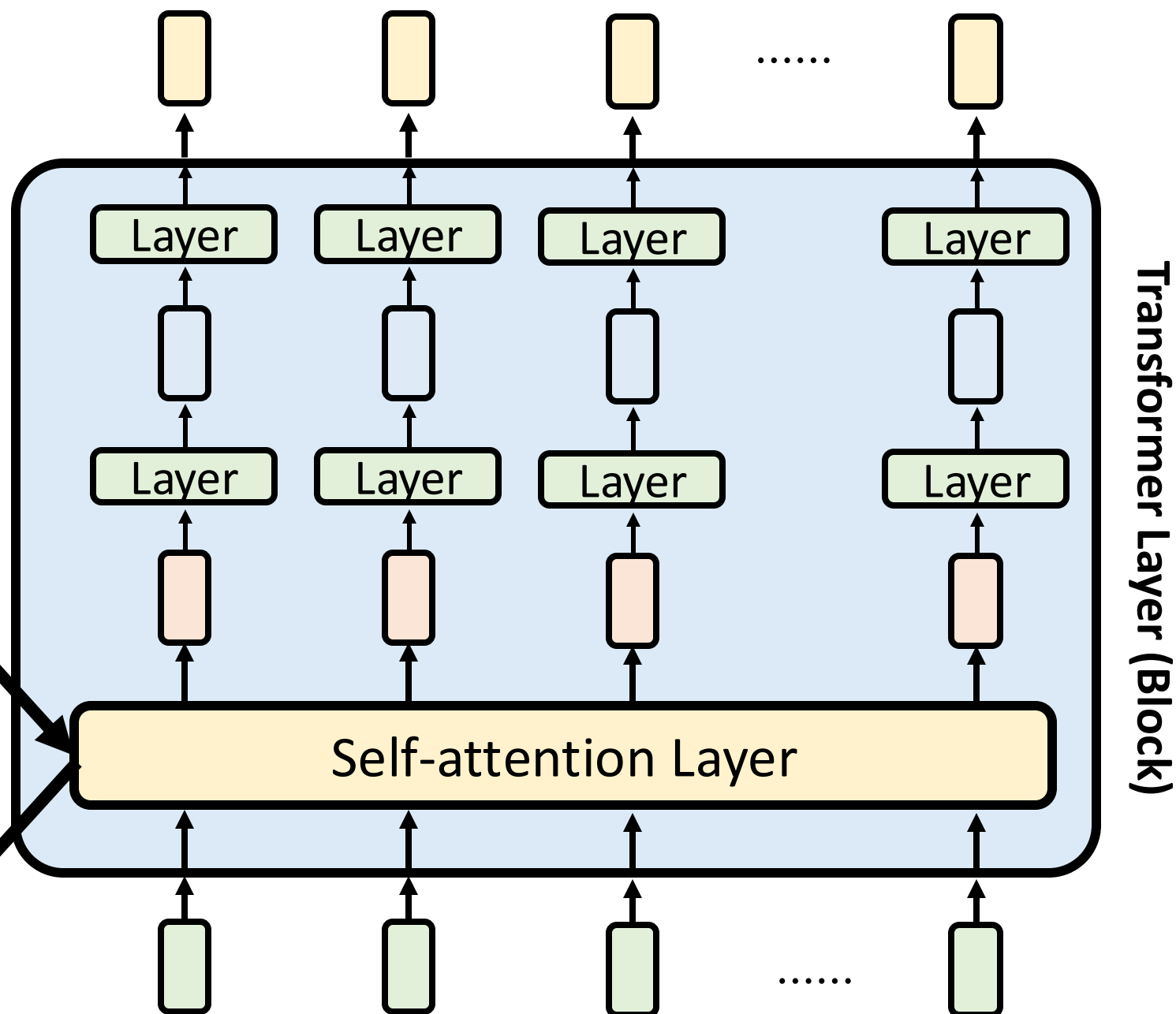


(b) with skip connections

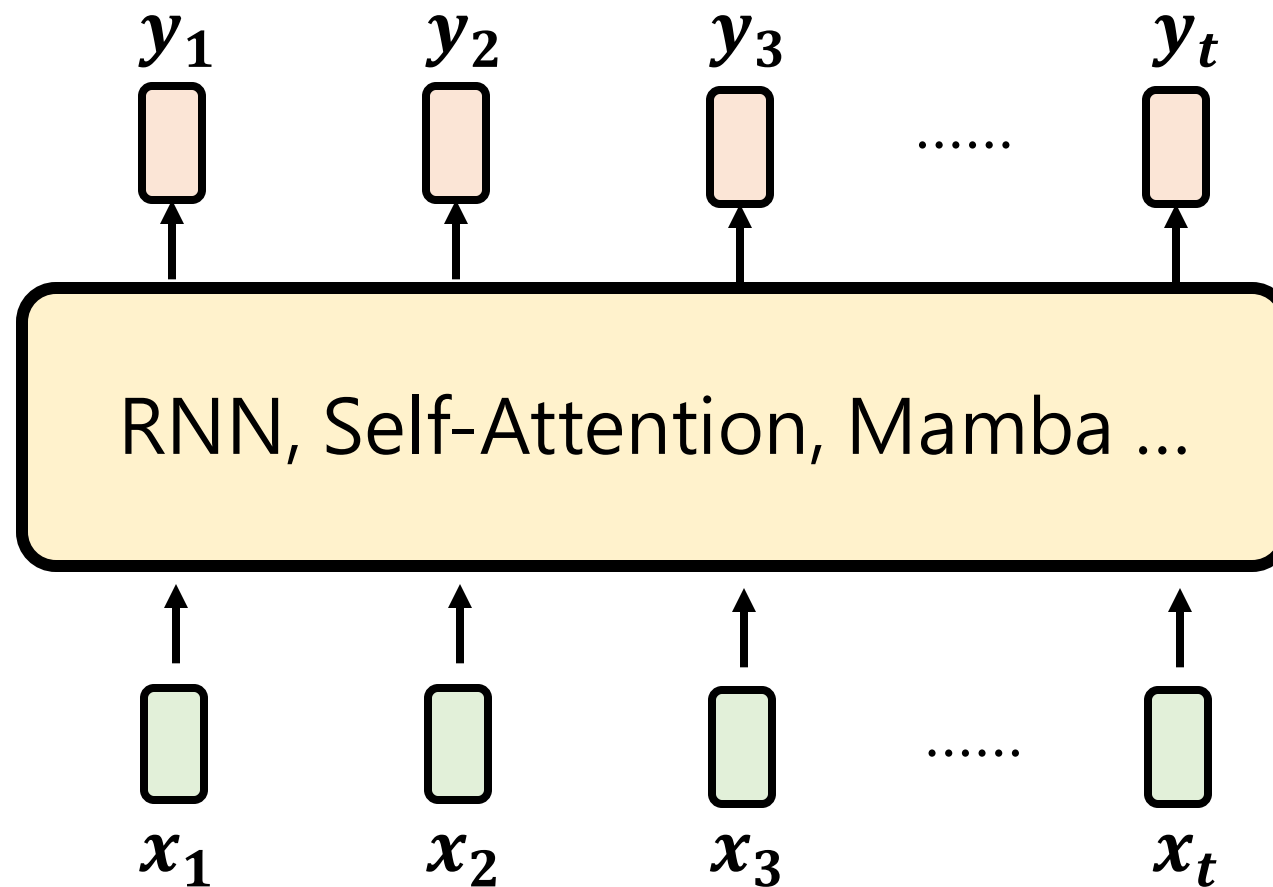
Transformer 出現
的理由是甚麼呢？

RNN (LSTM)

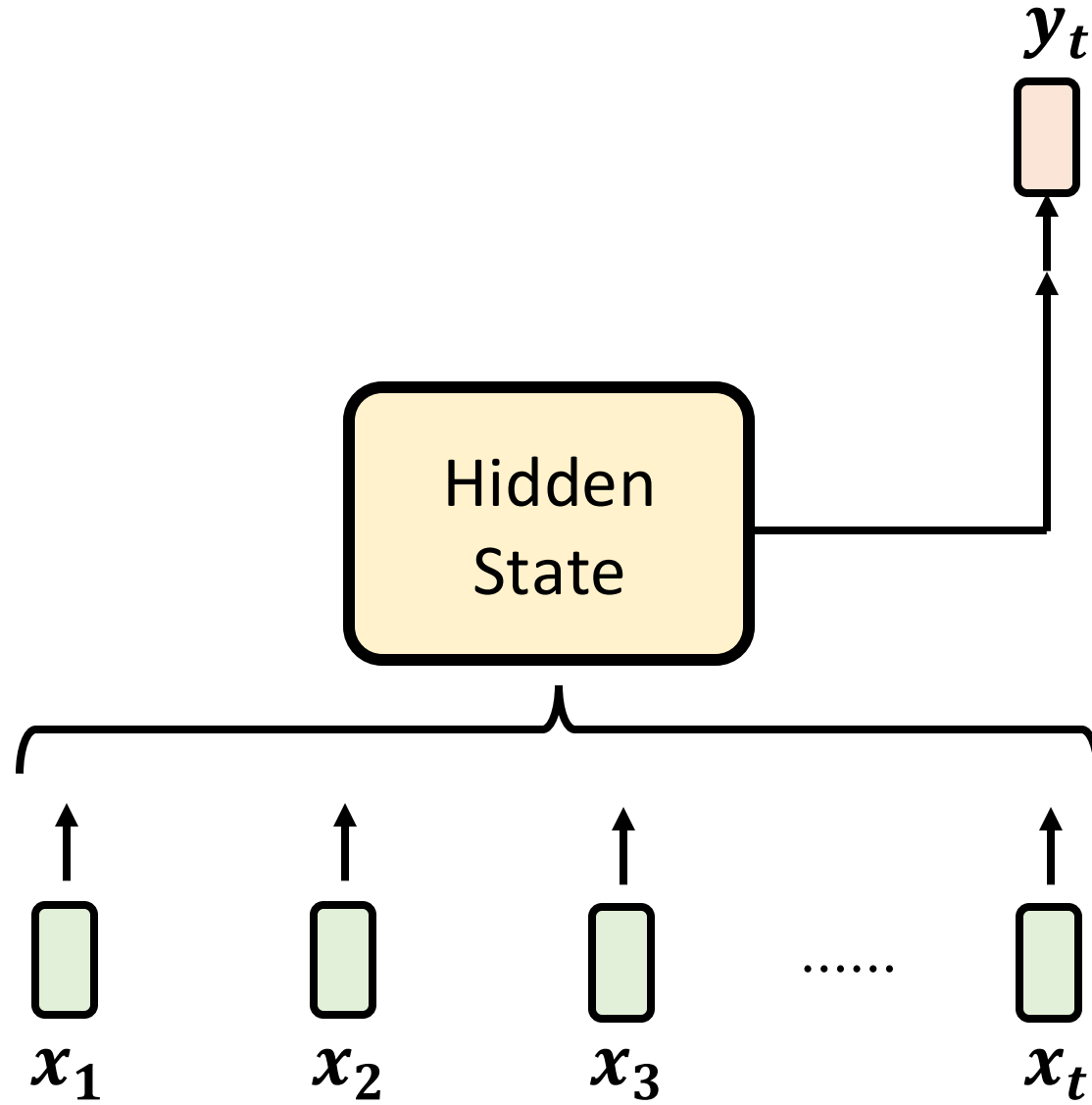
Mamba (and its friends)



要解的問題



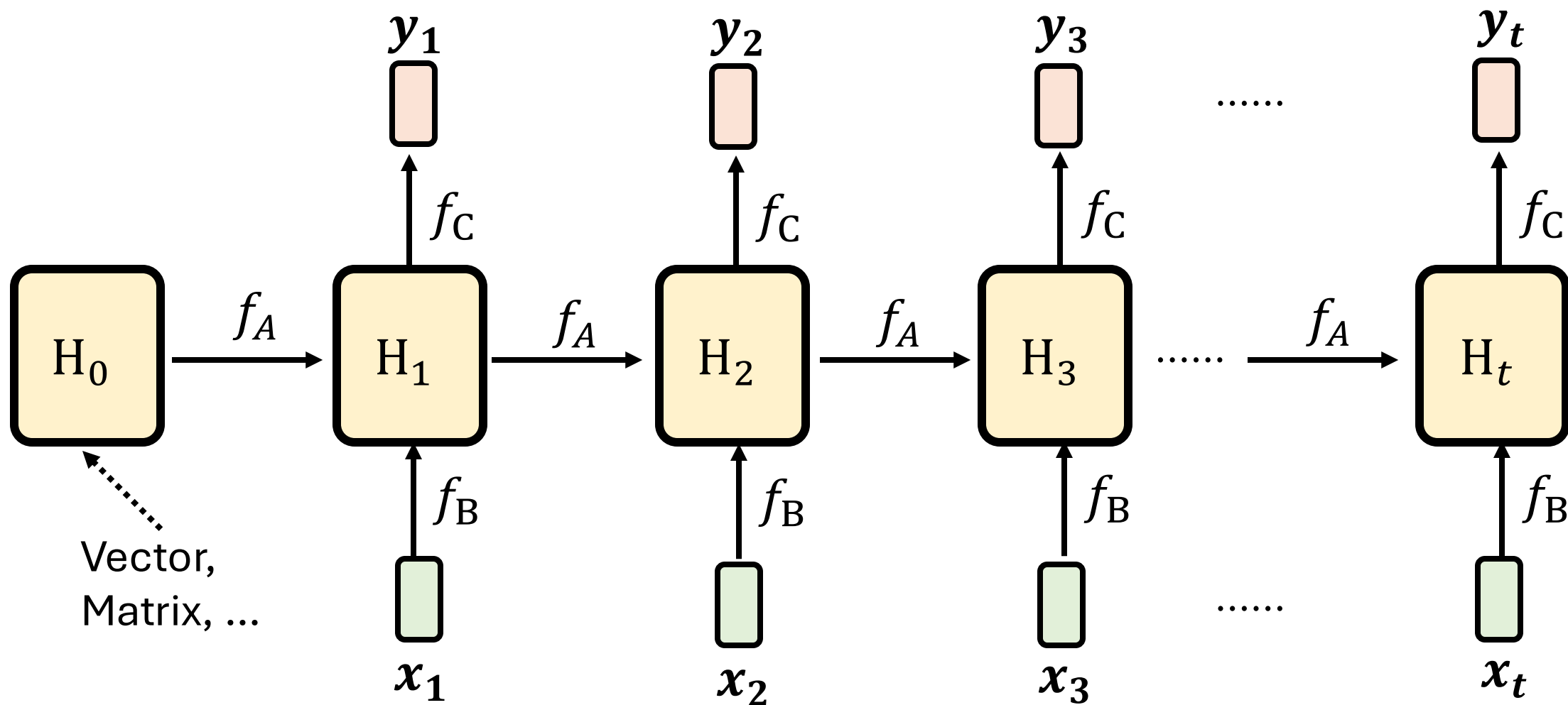
RNN-Style



RNN-Style

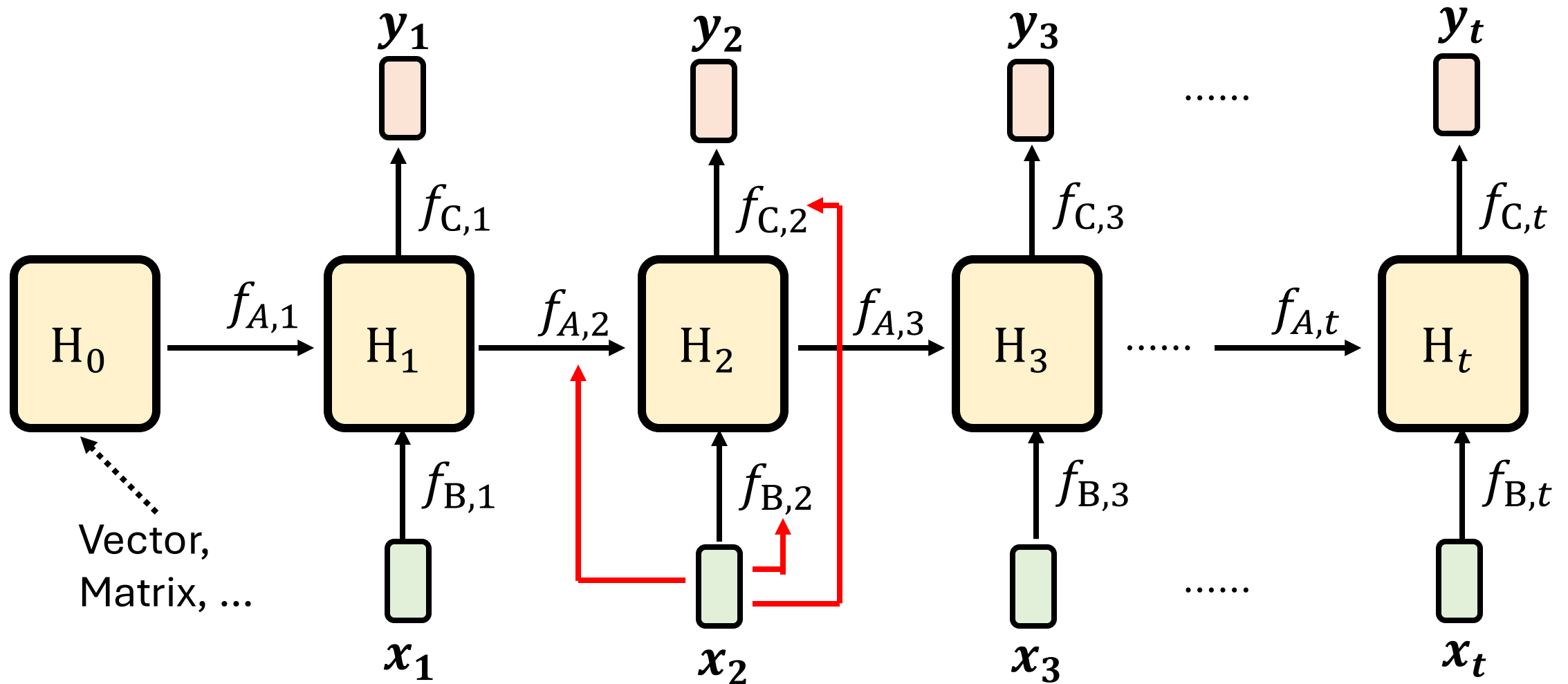
$$H_t = f_A(H_{t-1}) + f_B(x_t)$$

$$y_t = f_C(H_t)$$



RNN-Style

$$H_t = f_{A,t}(H_{t-1}) + f_{B,t}(x_t)$$
$$y_t = f_{C,t}(H_t)$$

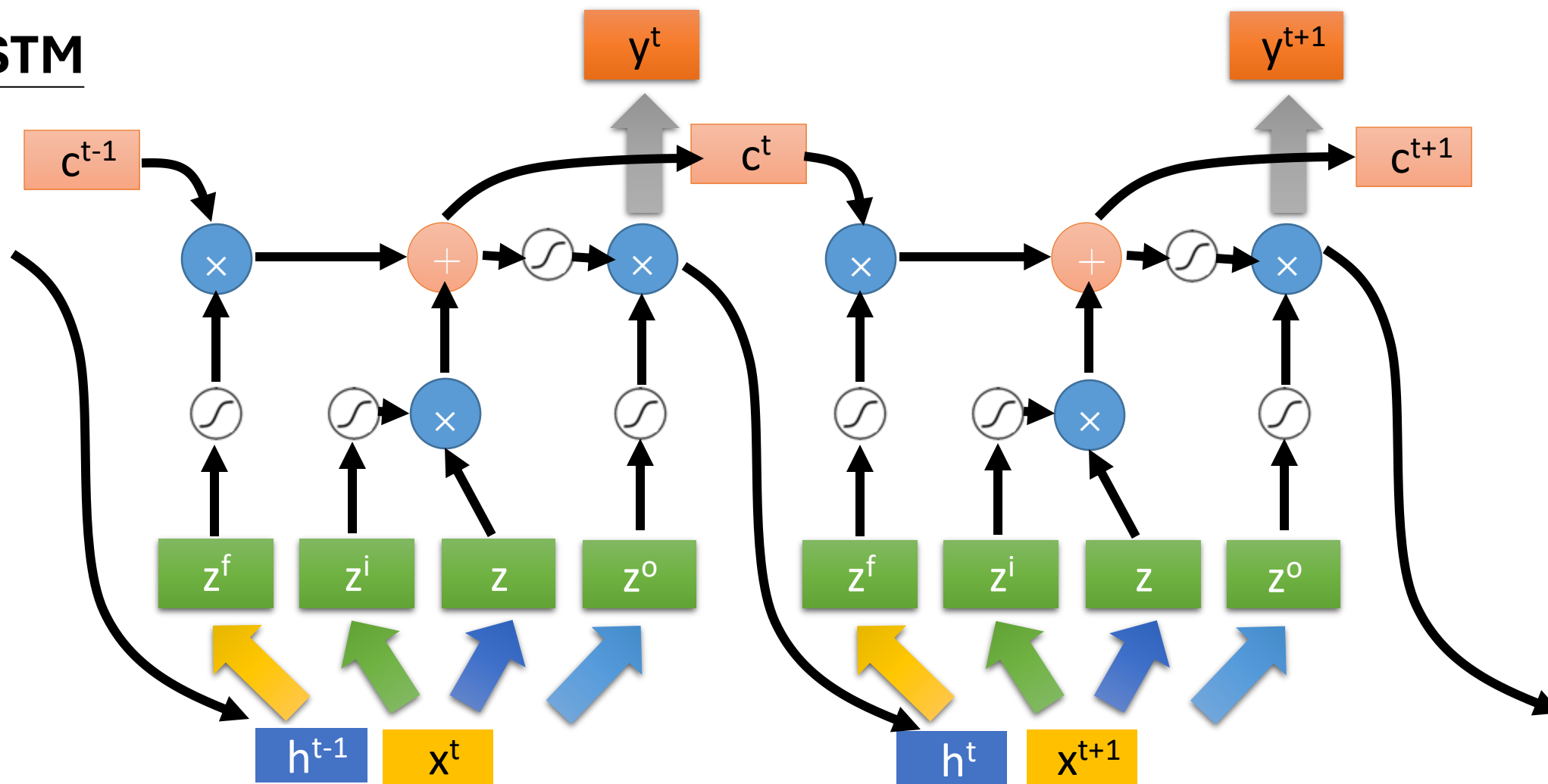


RNN-Style

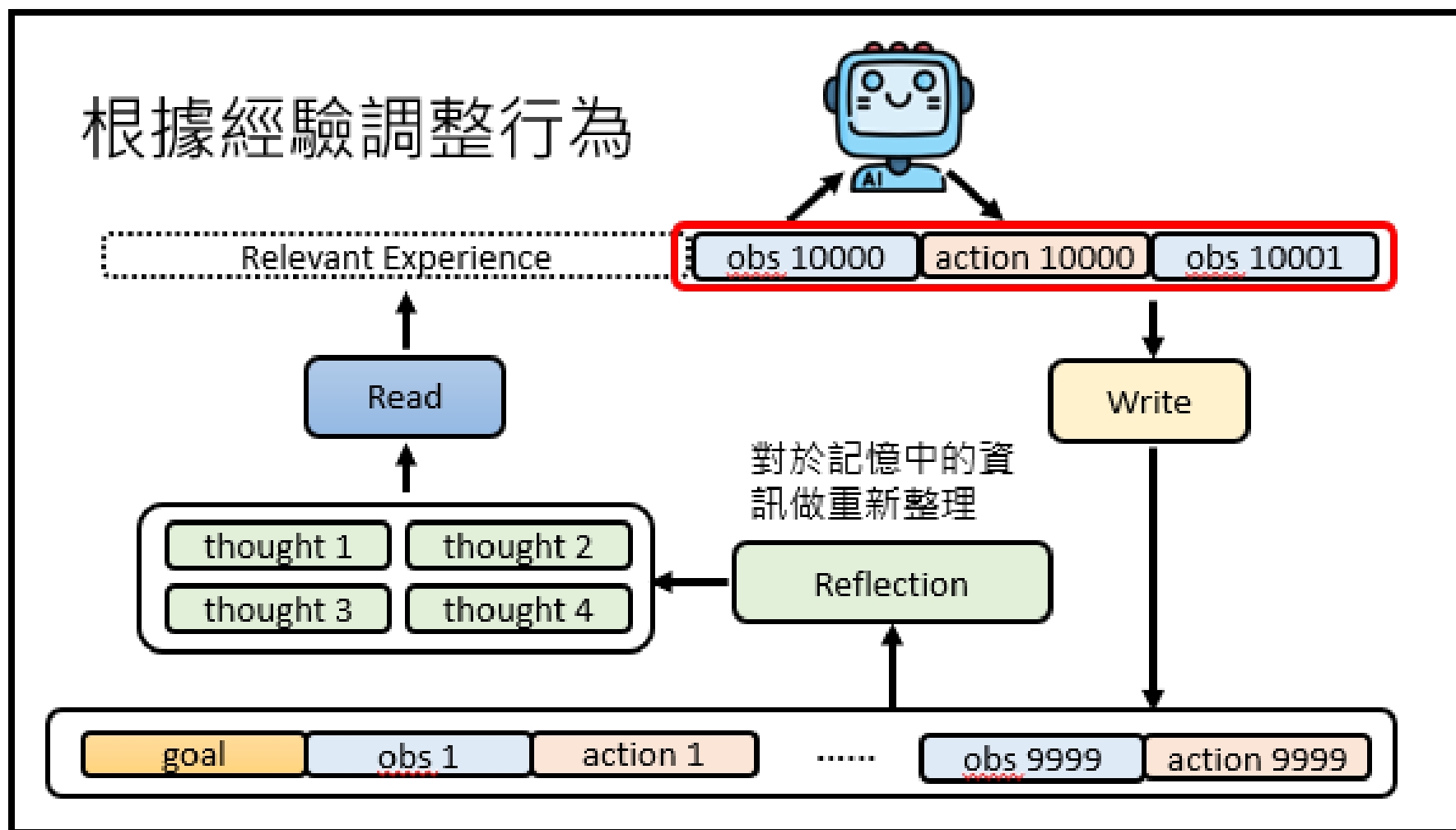
$$H_t = f_{A,t}(H_{t-1}) + f_{B,t}(x_t)$$

$$y_t = f_{C,t}(H_t)$$

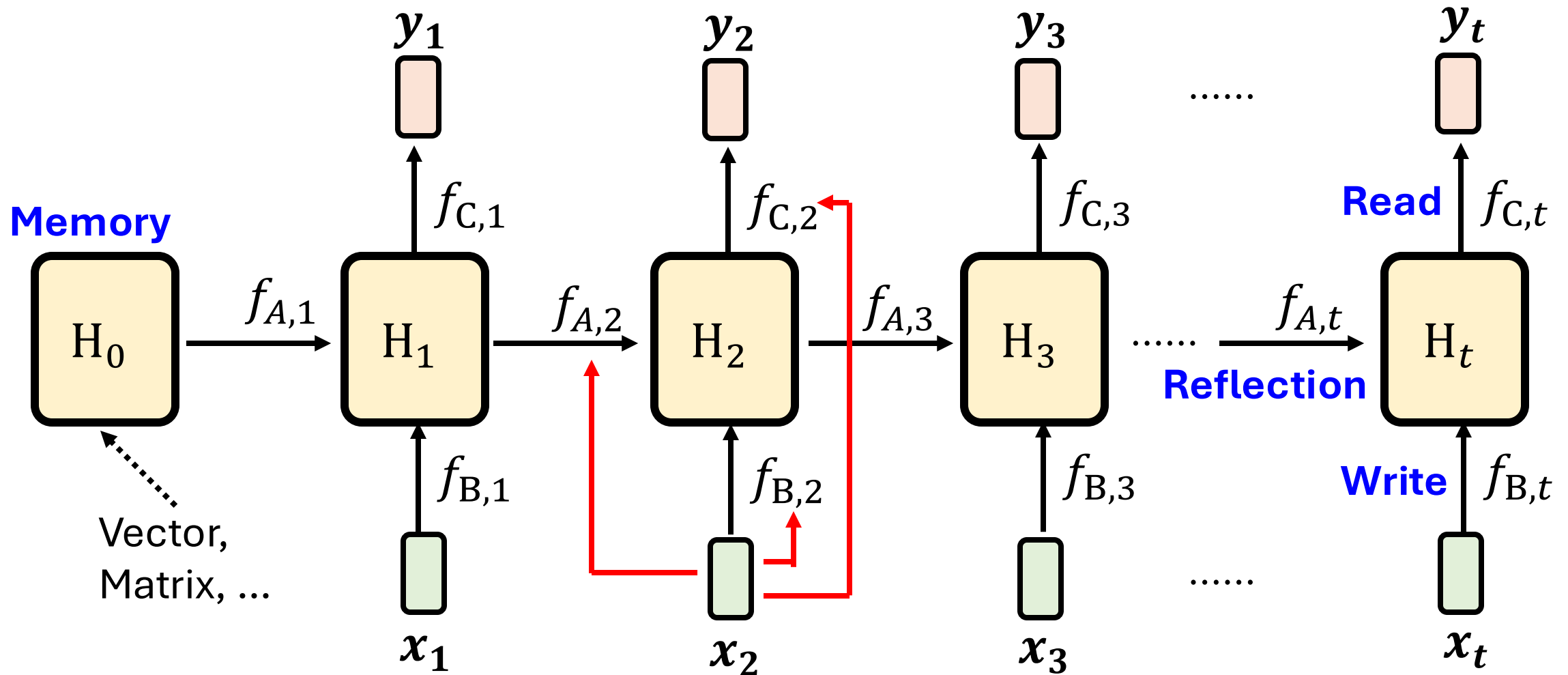
LSTM

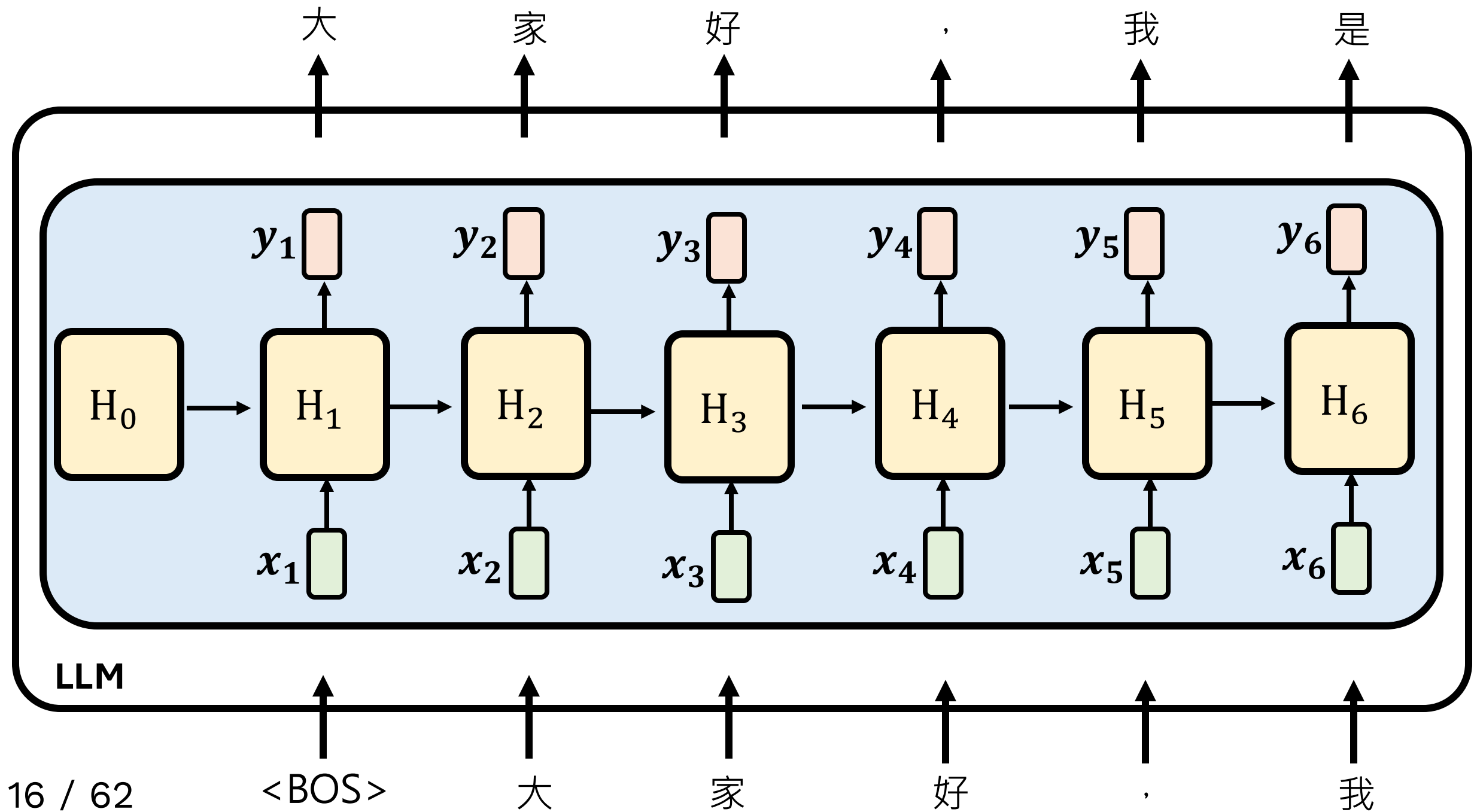


RNN-Style vs. AI Agent's Memory

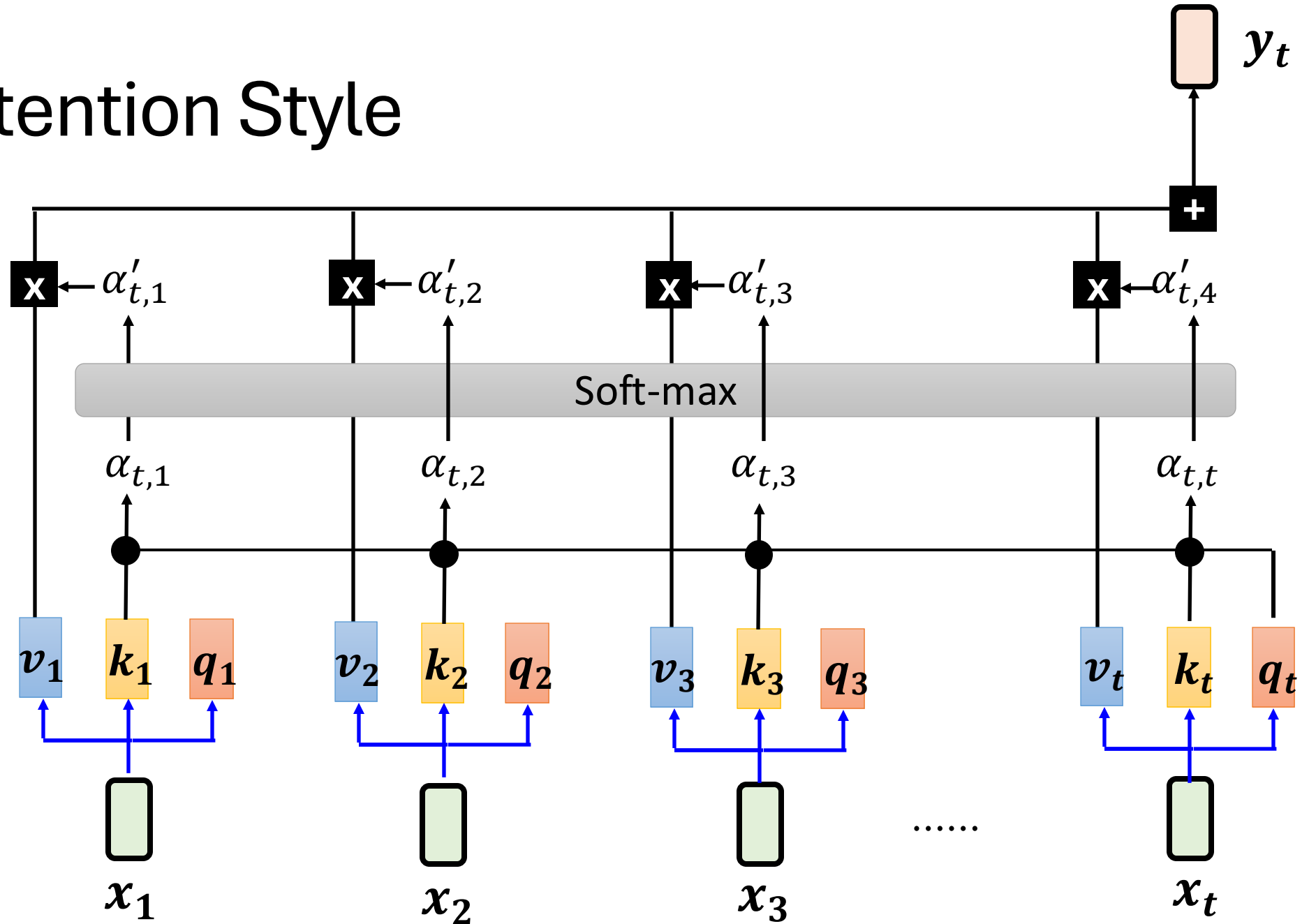


RNN-Style vs. AI Agent's Memory

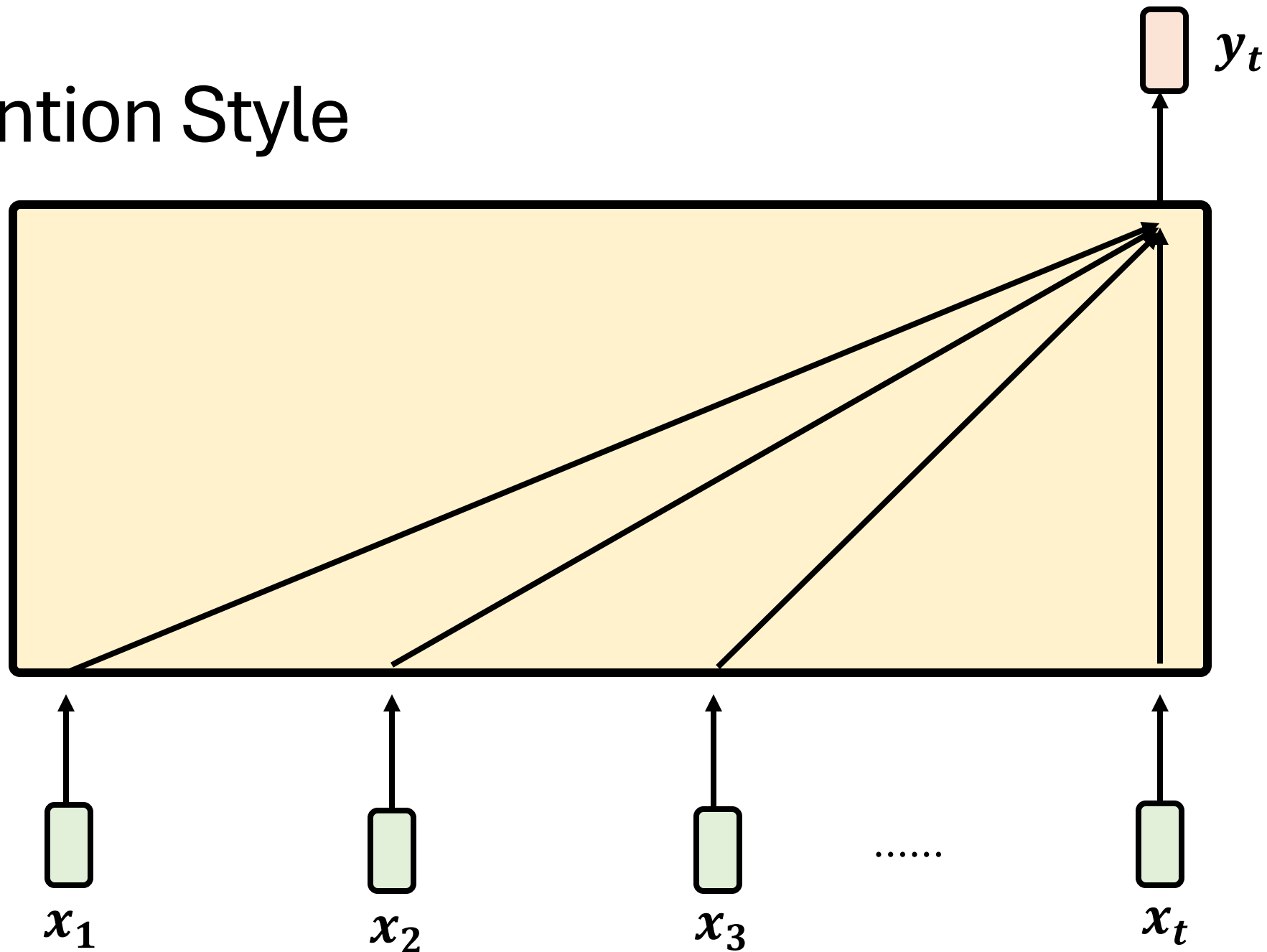




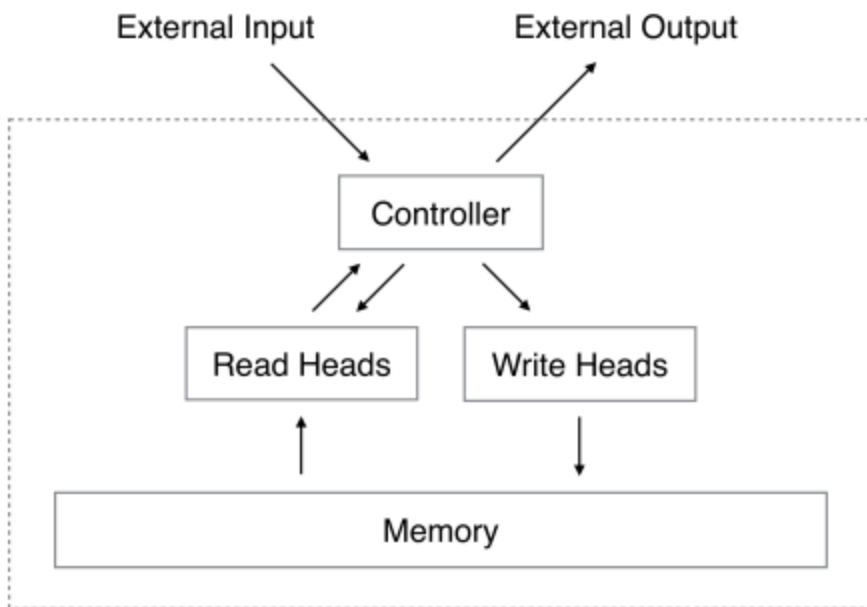
Self-Attention Style



Self-Attention Style



Attention 的概念很早就有了



Neural Turing Machine

<https://arxiv.org/abs/1410.5401>

Joe went to the garden then Fred picked up the milk; Joe moved to the bathroom and Fred dropped the milk, and then Dan moved to the living_room.
Where is Dan? A: living room I believe
Where is Joe? A: the bathroom
Fred moved to the bedroom and Joe went to the kitchen then Joe took the milk there and Dan journeyed to the bedroom; Joe discarded the milk.
Where is the milk now ? A: the milk is in the kitchen
Where is Dan now? A: I think he is in the bedroom
Joe took the milk there, after that Mike travelled to the office, then Joe went to the living_room, next Dan went back to the kitchen and Joe travelled to the office.
Where is Joe now? A: I think Joe is in the office

Memory Networks

<https://arxiv.org/pdf/1410.3916>

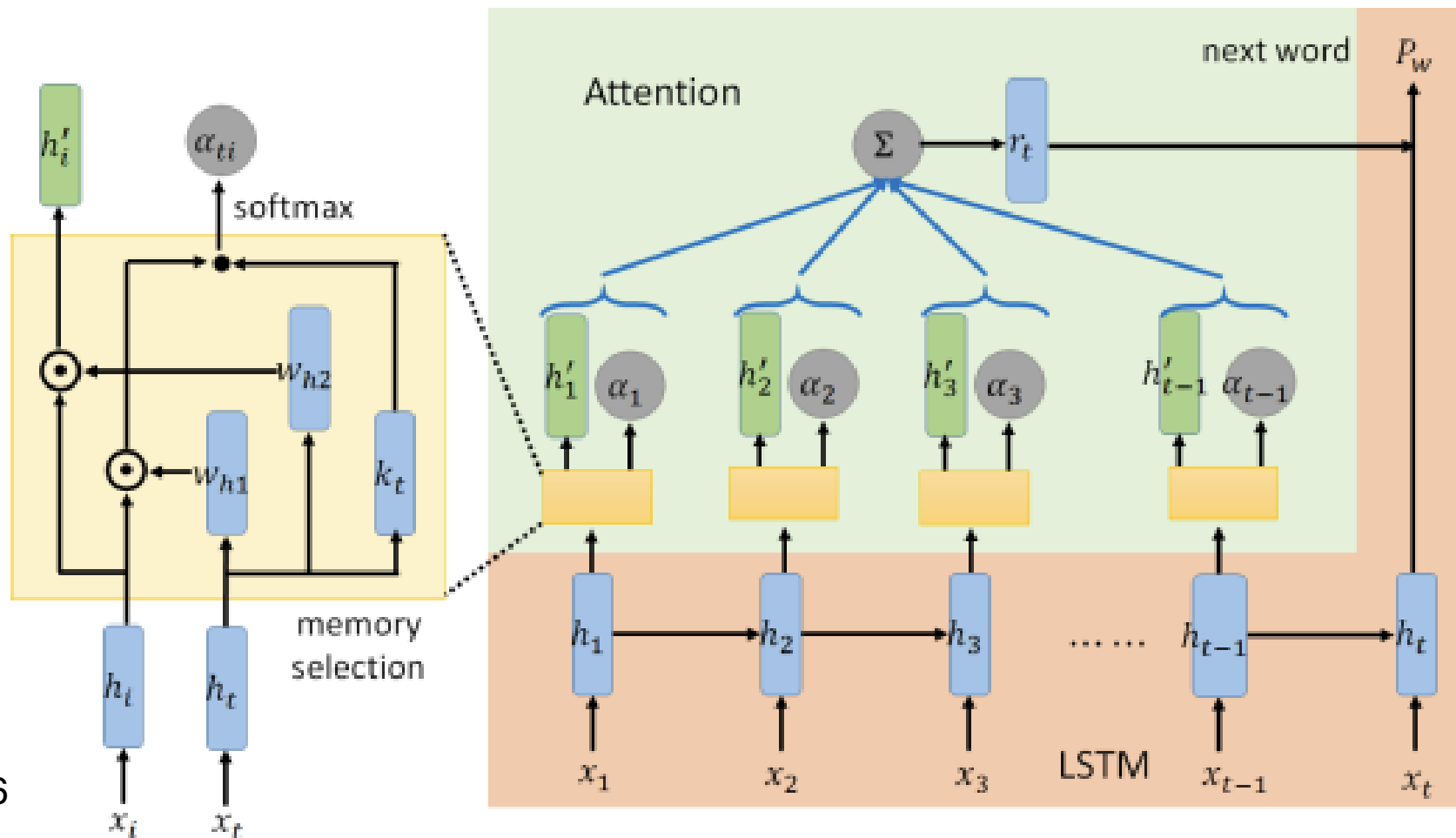
Attention 的概念很早就有了

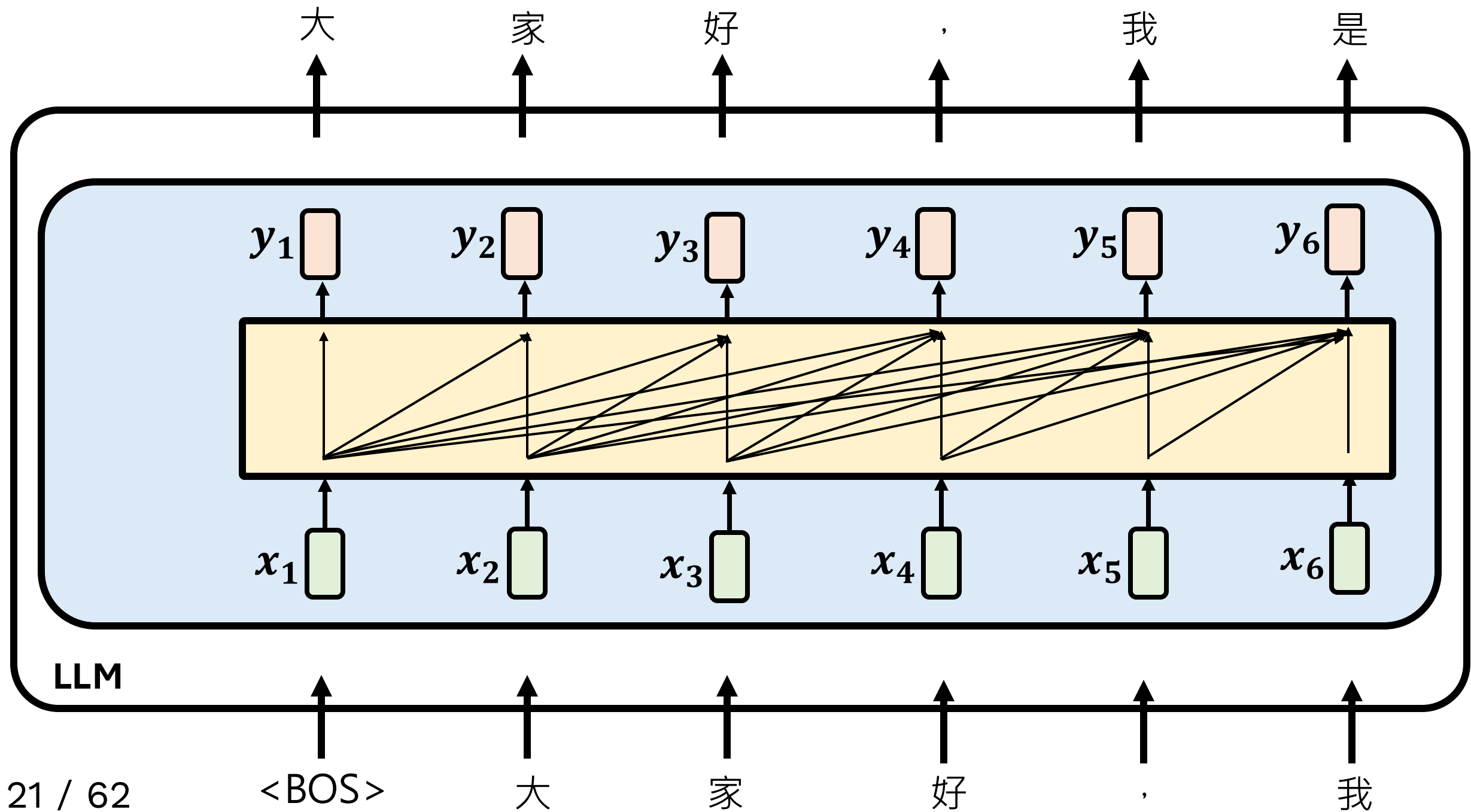


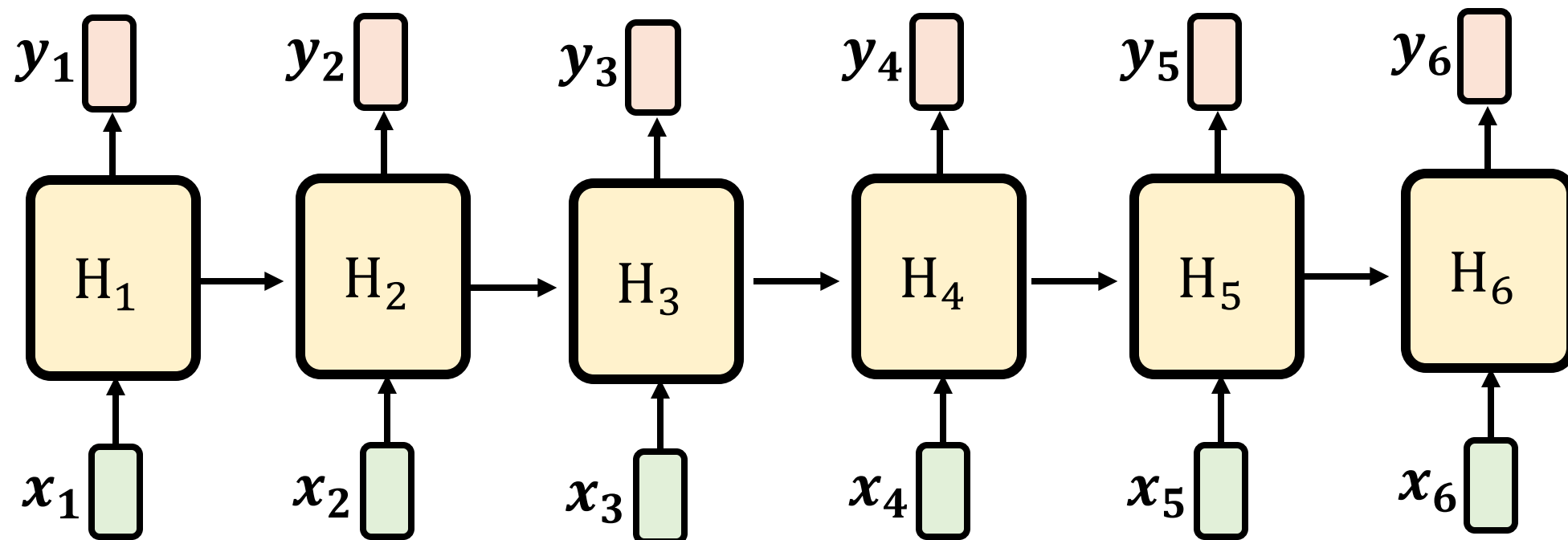
Da-Rong Liu

Attention-based Memory
Selection Recurrent Network
for Language Modeling

<https://arxiv.org/abs/1611.08656>

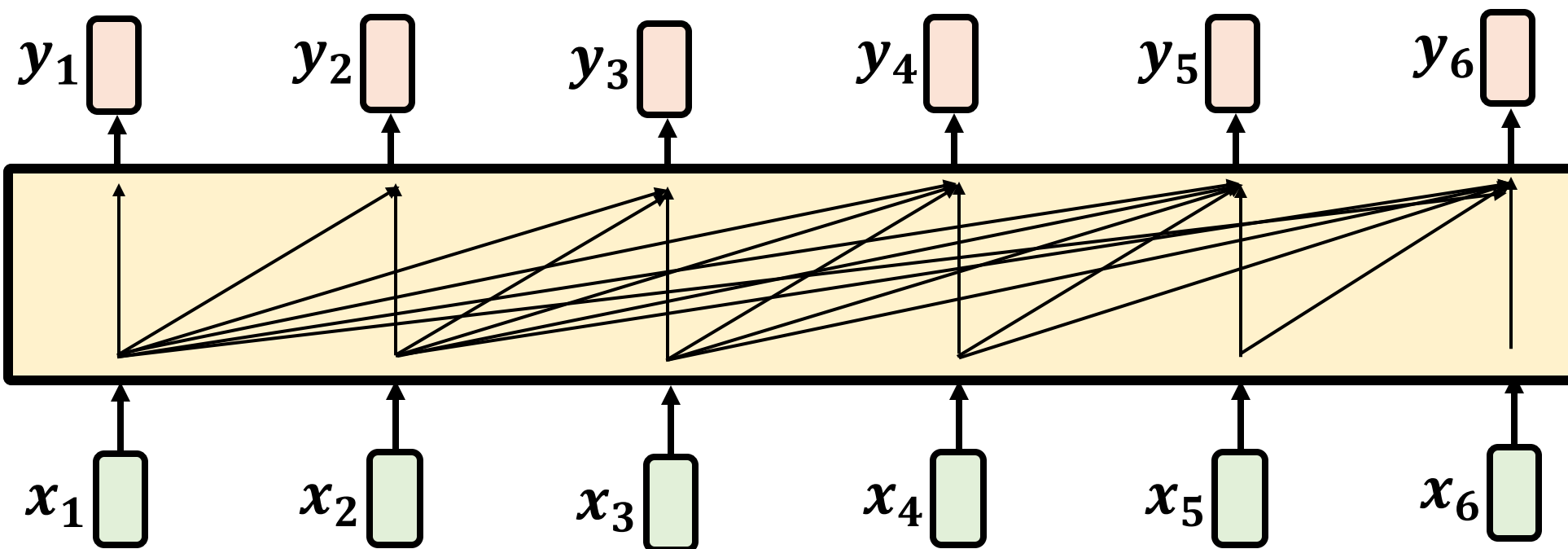






每一步運算量
都一樣

RNN 沒辦法
記大量資訊？

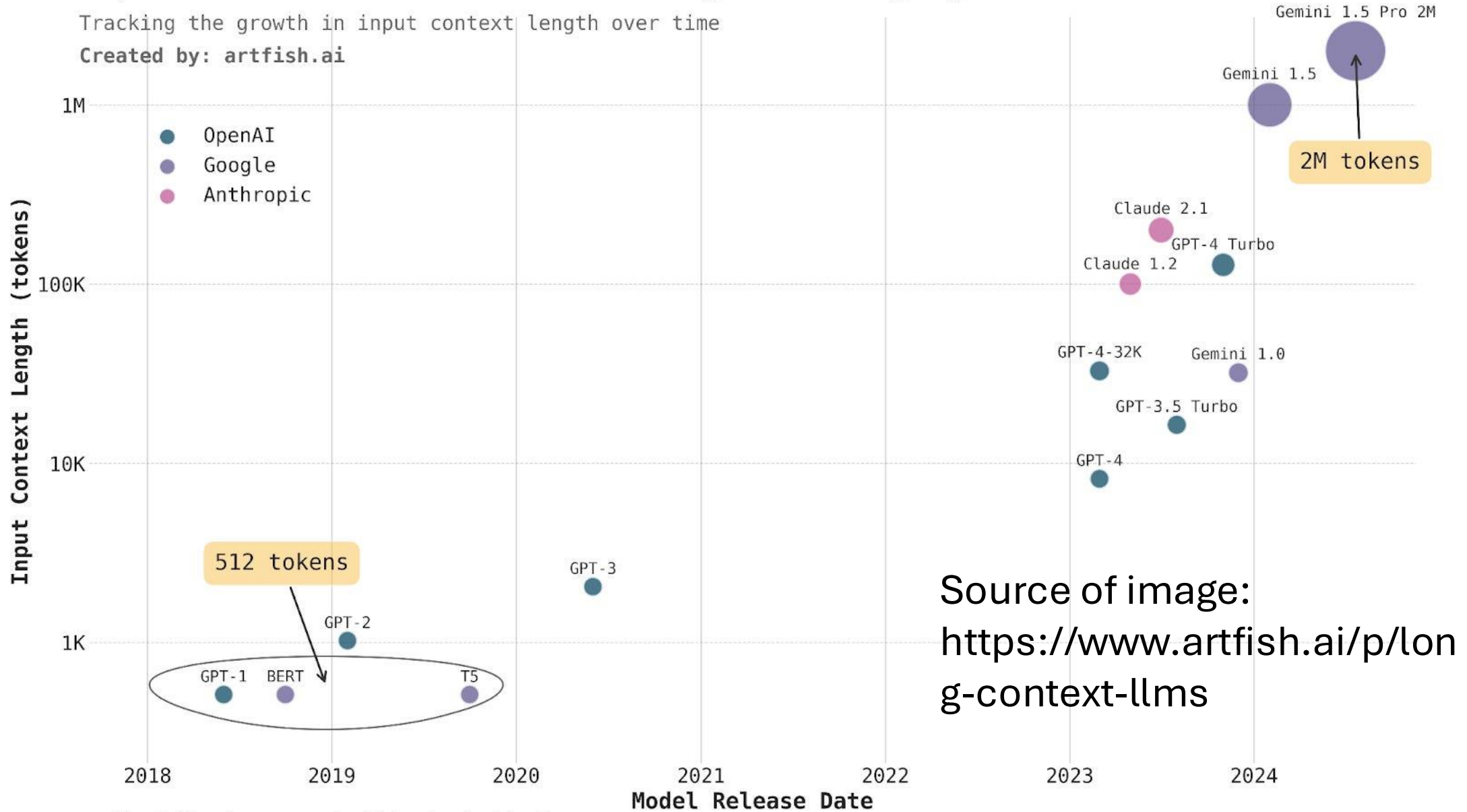


輸入越長，運
算量越來越大

Exponential Growth of Context Length in Language Models

Tracking the growth in input context length over time

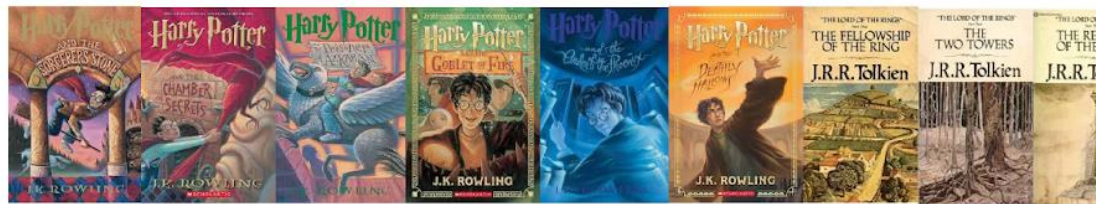
Created by: artfish.ai



Source of image:
<https://www.artfish.ai/p/long-context-llms>

Google's Gemini 1.5 can (almost) fit the entire Harry Potter + Lord of the Ring series in its 2 million context window

Gemini 1.5 2M
(June 2024)



Claude 2.1
(July 2023)



GPT-4 Turbo
(March 2023)



GPT-3.5 Turbo
(March 2022)



RAG、AI Agent 都需要語言模型處理很長的序列

Source of image:
<https://www.artfish.ai/p/long-context-llms>

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

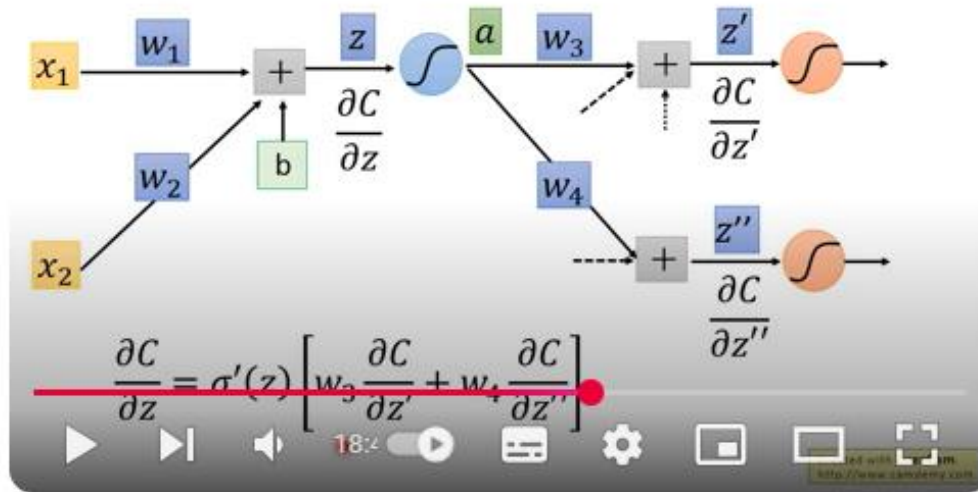
In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

<https://arxiv.org/abs/1706.03762>

語言模型的訓練 (找出參數)

Backpropagation – Backward pass

Compute $\partial C / \partial z$ for all activation function inputs z



ML Lecture 7: Backpropagation

Backpropagation

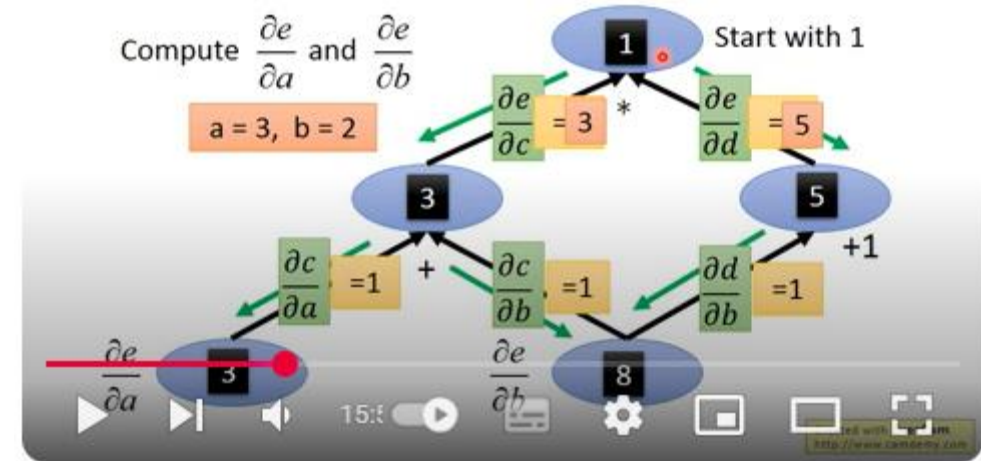
<https://youtu.be/ibJpTrp5mcE>

Computational Graph

- Example: $e = (a+b) * (b+1)$

Reverse mode

What is the benefit?



Computational Graph & Backpropagation

Computational Graph

<https://youtu.be/-yhm3WdGFok?si=2cZOANbtm0Mjd9lT>

語言模型的訓練 (找出參數)

- 更新參數前要先算出自己的答案

$$\{z_1, z_2, \dots, z_{t-1}\} \rightarrow z_t$$

2. 計算差異
1. 得到目前的答案

z_t

???

語言模型

3. 更新參數

z_1

z_2

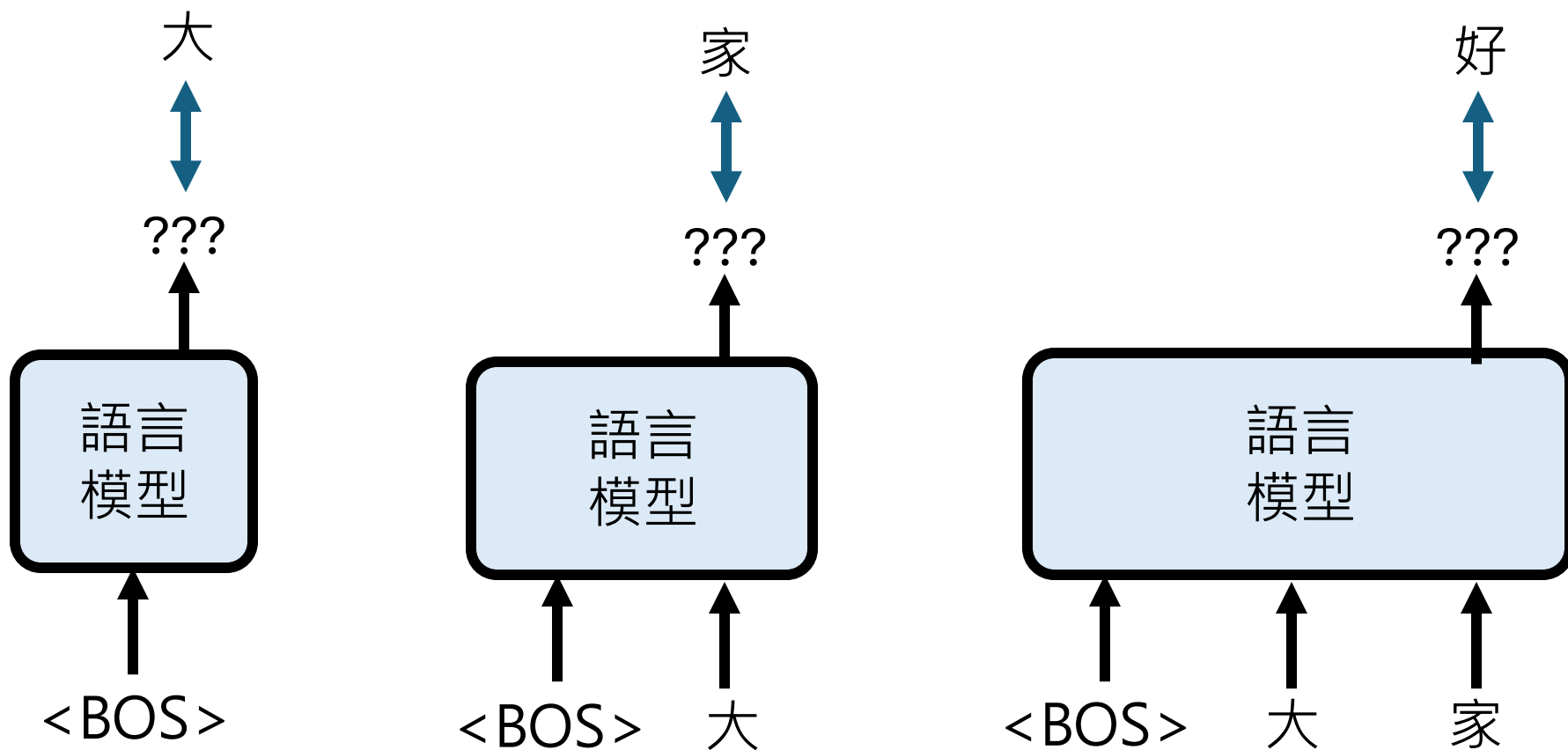
z_3

.....

z_{t-1}

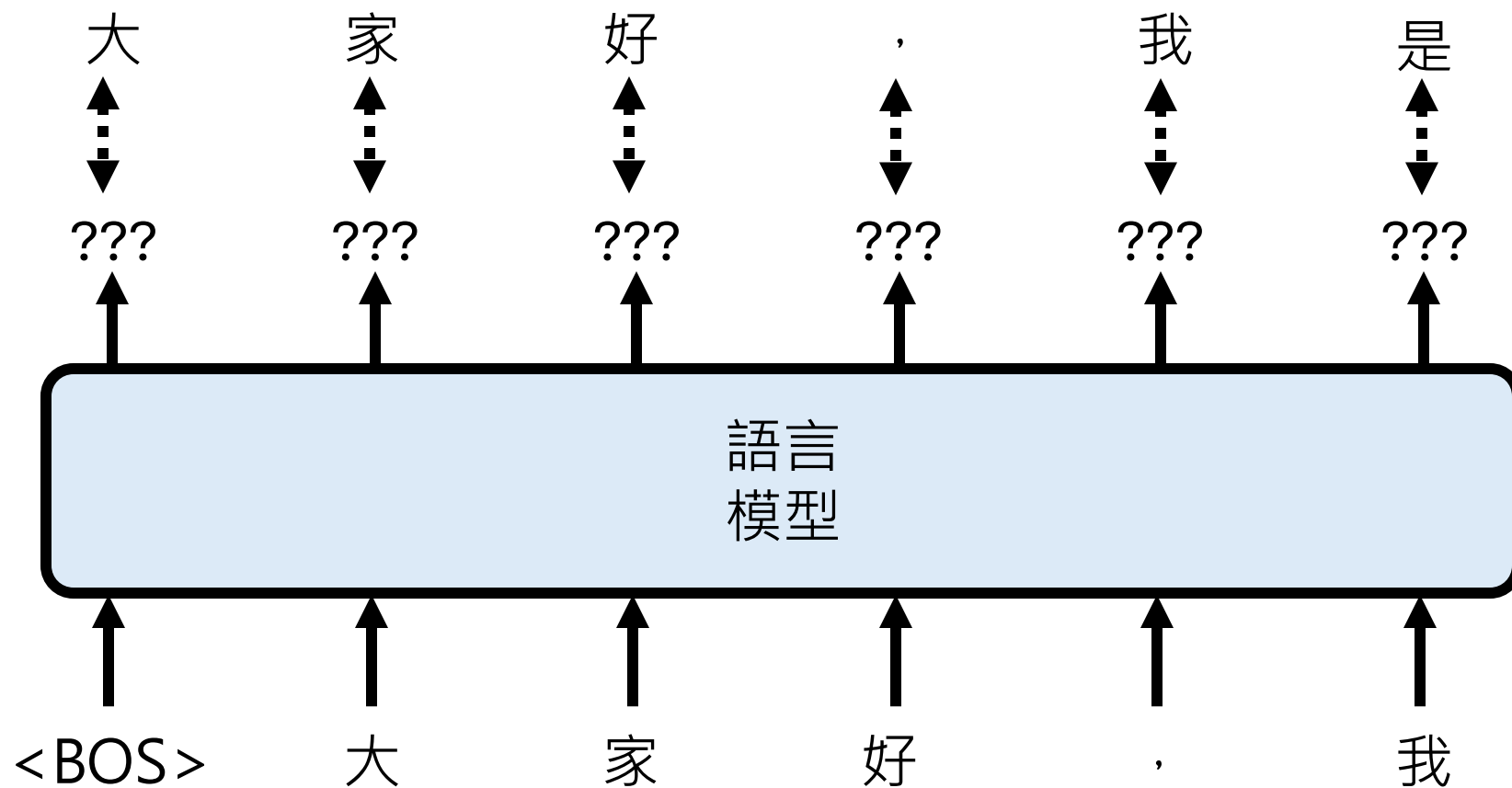
語言模型的訓練 (找出參數)

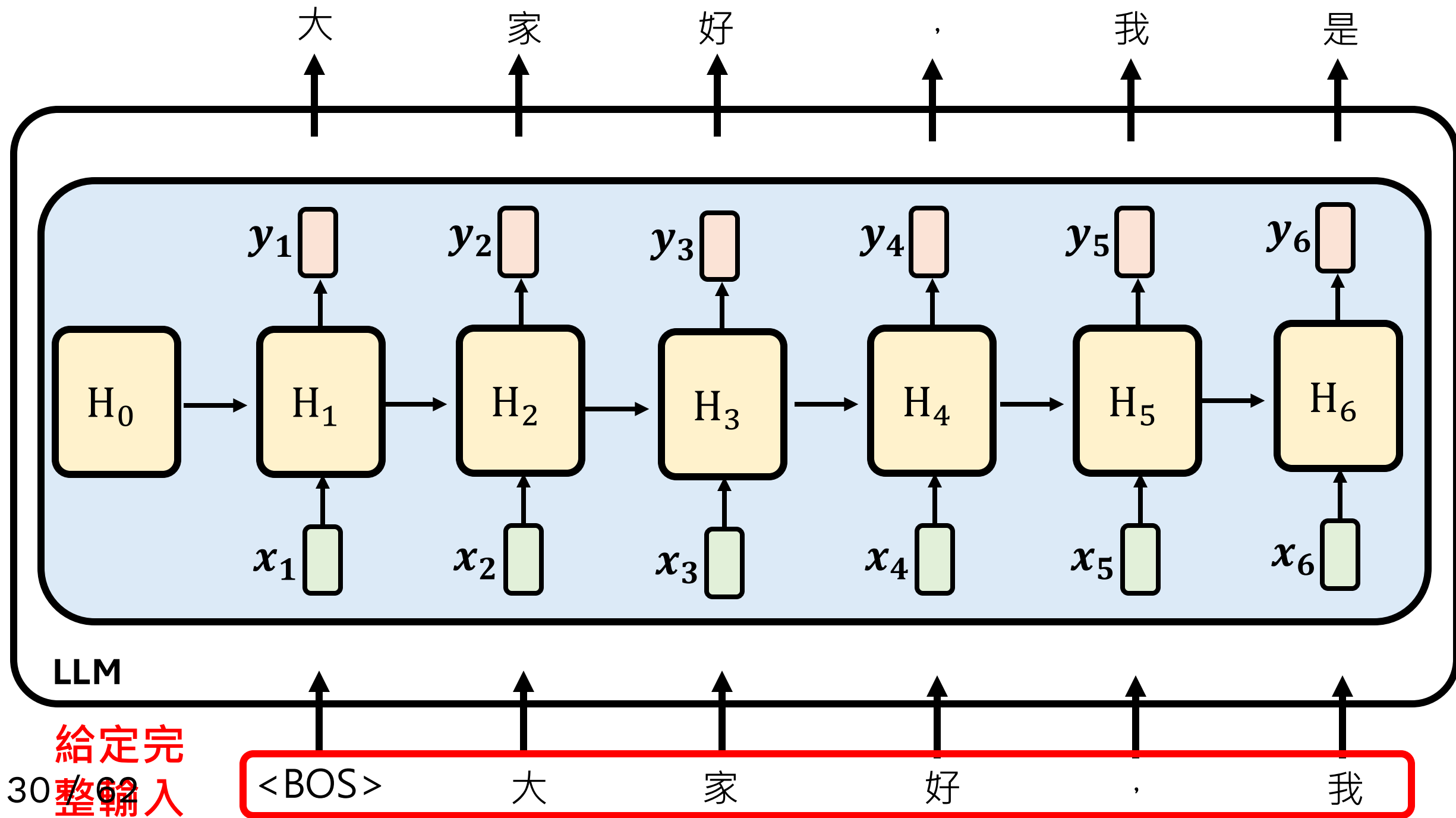
假設我們想要教模型說「大 家 好 ， 我 是」

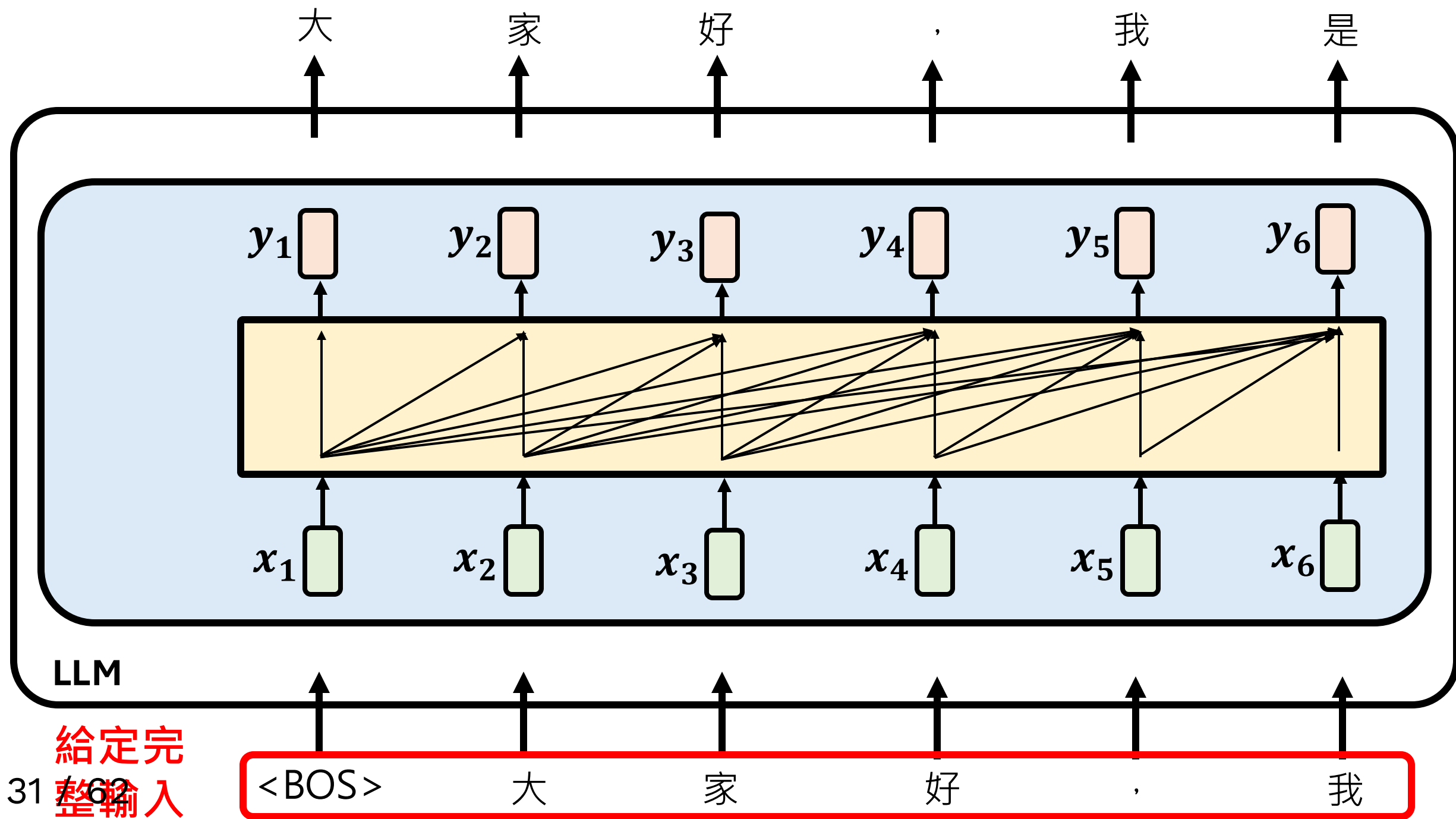


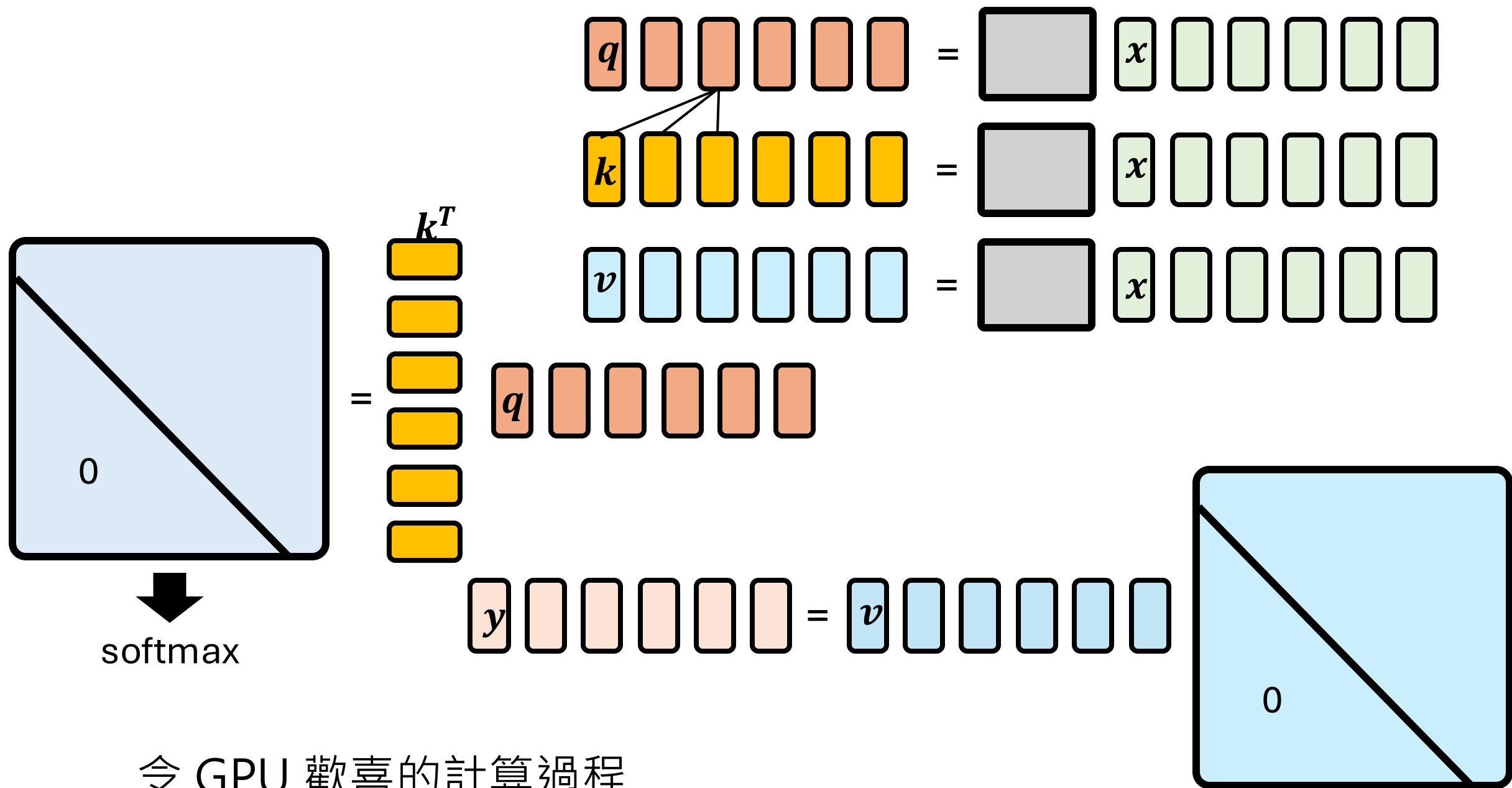
語言模型的訓練 (找出參數)

假設我們想要教模型說「大 家 好 ， 我 是」









令 GPU 歡喜的計算過程

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$H_t = f_{A,t}(H_{t-1}) + f_{B,t}(\mathbf{x}_t)$$

$$\mathbf{y}_t = f_{C,t}(H_t)$$

$$\left\{ \begin{array}{l} H_1 = f_{A,1}(H_0) + f_{B,1}(\mathbf{x}_1) = f_{B,1}(\mathbf{x}_1) \\ H_2 = f_{A,2}(H_1) + f_{B,2}(\mathbf{x}_2) = f_{A,2}(f_{B,1}(\mathbf{x}_1)) + f_{B,2}(\mathbf{x}_2) \\ H_3 = f_{A,3}(H_2) + f_{B,3}(\mathbf{x}_3) = f_{A,3}(f_{A,2}(f_{B,1}(\mathbf{x}_1)) + f_{B,2}(\mathbf{x}_2)) + f_{B,3}(\mathbf{x}_3) \\ \vdots \\ H_t = f_{A,t}(H_{t-1}) + f_{B,t}(\mathbf{x}_t) = \underbrace{f_{A,t}(f_{A,t-1} \dots f_{A,3}(f_{A,2}(f_{B,1}(\mathbf{x}_1) \dots))}_{\dots} + f_{B,t}(\mathbf{x}_t) \end{array} \right.$$

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t)$$

$$\mathbf{y}_t = f_{C,t}(H_t)$$

H_t is a $d \times d$ matrix

$$f_{B,t}(\mathbf{x}_t) = D_t$$

$$\left\{ \begin{array}{l} H_1 = H_0 + f_{B,1}(\mathbf{x}_1) \\ H_2 = H_1 + f_{B,2}(\mathbf{x}_2) \\ H_3 = H_2 + f_{B,3}(\mathbf{x}_3) \\ \vdots \\ H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t) \end{array} \right. \quad \begin{array}{l} = f_{B,1}(\mathbf{x}_1) \\ = f_{B,1}(\mathbf{x}_1) + f_{B,2}(\mathbf{x}_2) \\ = f_{B,1}(\mathbf{x}_1) + f_{B,2}(\mathbf{x}_2) + f_{B,3}(\mathbf{x}_3) \\ \dots \dots + f_{B,t}(\mathbf{x}_t) \end{array}$$

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t)$$

$$\mathbf{y}_t = f_{C,t}(H_t)$$

$$\left\{ \begin{array}{ll} H_1 = D_1 & \mathbf{y}_1 = D_1 \mathbf{q}_1 \\ H_2 = D_1 + D_2 & \mathbf{y}_2 = D_1 \mathbf{q}_2 + D_2 \mathbf{q}_2 \\ H_3 = D_1 + D_2 + D_3 & \mathbf{y}_3 = D_1 \mathbf{q}_3 + D_2 \mathbf{q}_3 + D_3 \mathbf{q}_3 \\ \vdots & \\ H_t = D_1 + D_2 + \dots + D_t & \mathbf{y}_t = D_1 \mathbf{q}_t + D_2 \mathbf{q}_t + \dots + D_t \mathbf{q}_t \end{array} \right.$$

H_t is a $d \times d$ matrix

$$f_{B,t}(\mathbf{x}_t) = D_t$$

$$f_{C,t}(H_t) = H_t \mathbf{q}_t$$

$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$\left\{ \begin{array}{l} y_1 = D_1 \mathbf{q}_1 \\ y_2 = D_1 \mathbf{q}_2 + D_2 \mathbf{q}_2 \\ y_3 = D_1 \mathbf{q}_3 + D_2 \mathbf{q}_3 + D_3 \mathbf{q}_3 \\ \vdots \\ y_t = D_1 \mathbf{q}_t + D_2 \mathbf{q}_t + \cdots + D_t \mathbf{q}_t \end{array} \right.$$

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t)$$

$$\mathbf{y}_t = f_{C,t}(H_t)$$

H_t is a $d \times d$ matrix

$$f_{B,t}(\mathbf{x}_t) = D_t$$

$$D_t = \mathbf{v}_t \mathbf{k}_t^T \quad \begin{array}{l} \mathbf{v}_t = W_v \mathbf{x}_t \\ \mathbf{k}_t = W_k \mathbf{x}_t \end{array}$$

$$f_{C,t}(H_t) = H_t \mathbf{q}_t$$

$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$\left\{ \begin{array}{l} y_1 = v_1 k_1^T q_1 \\ y_2 = v_1 k_1^T q_2 + v_2 k_2^T q_2 \\ y_3 = v_1 k_1^T q_3 + v_2 k_2^T q_3 + v_3 k_3^T q_3 \\ \vdots \\ y_t = v_1 k_1^T q_t + v_2 k_2^T q_t + \cdots + v_t k_t^T q_t \end{array} \right.$$

$$H_t = H_{t-1} + f_{B,t}(x_t)$$

$$y_t = f_{C,t}(H_t)$$

H_t is a $d \times d$ matrix

$$f_{B,t}(x_t) = D_t$$

$$D_t = v_t k_t^T \quad \begin{array}{l} v_t = W_v x_t \\ k_t = W_k x_t \end{array}$$

$$f_{C,t}(H_t) = H_t q_t$$

$$q_t = W_Q x_t$$

RNN 有沒有訓練時平行的可能性

$$f_{A,1}(H_0) = 0$$

$$\mathbf{y}_t = \mathbf{v}_1 \mathbf{k}_1^T \mathbf{q}_t + \mathbf{v}_2 \mathbf{k}_2^T \mathbf{q}_t + \cdots + \mathbf{v}_t \mathbf{k}_t^T \mathbf{q}_t$$

$$= \mathbf{v}_1 a_{t,1} + \mathbf{v}_2 a_{t,2} + \cdots + \mathbf{v}_t a_{t,t}$$

$$= a_{t,1} \mathbf{v}_1 + a_{t,2} \mathbf{v}_2 + \cdots + a_{t,t} \mathbf{v}_t$$

這不就是 Self-attention! (少了 softmax)

叫做 Linear Attention

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t)$$

$$\mathbf{y}_t = f_{C,t}(H_t)$$

H_t is a $d \times d$ matrix

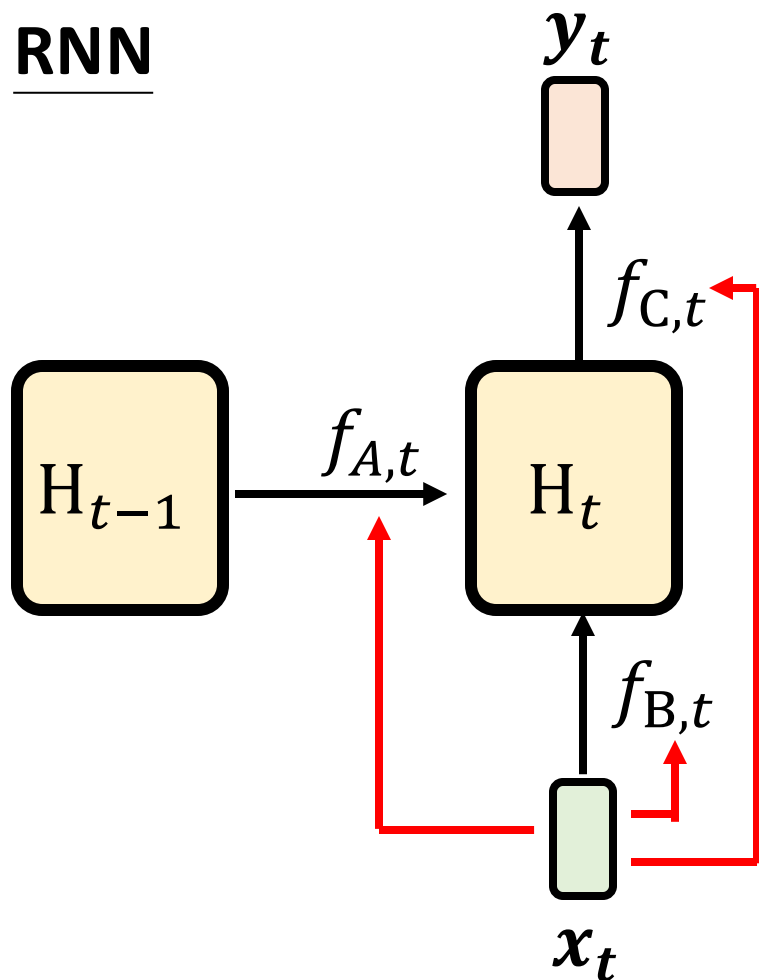
$$f_{B,t}(\mathbf{x}_t) = D_t$$

$$D_t = \mathbf{v}_t \mathbf{k}_t^T \quad \begin{aligned} \mathbf{v}_t &= W_v \mathbf{x}_t \\ \mathbf{k}_t &= W_k \mathbf{x}_t \end{aligned}$$

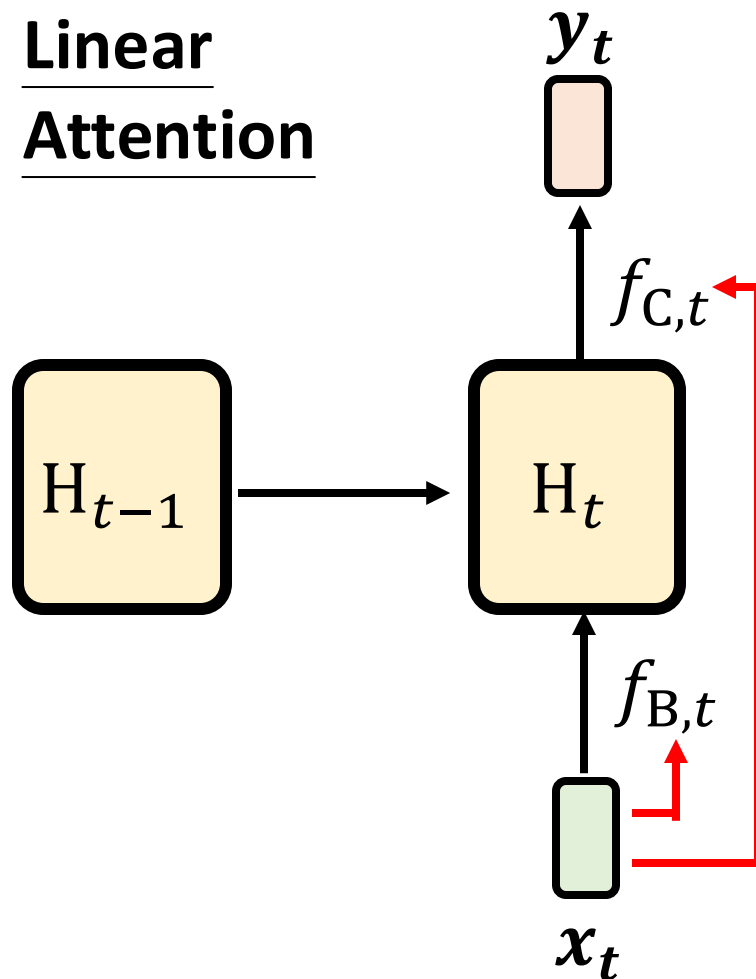
$$f_{C,t}(H_t) = H_t \mathbf{q}_t$$

$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

RNN



Linear Attention



$$f_{C,t}(H_t) = H_t \mathbf{q}_t$$
$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

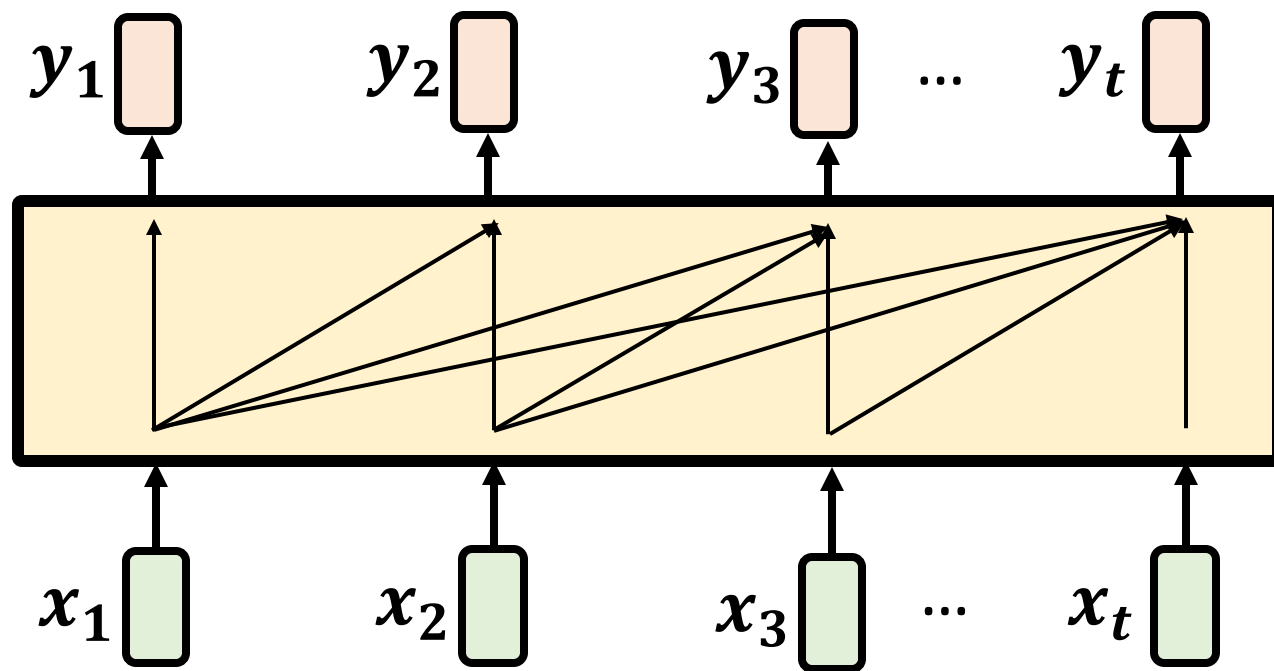
$$f_{B,t}(\mathbf{x}_t) = \mathbf{v}_t \mathbf{k}_t^T$$
$$\mathbf{v}_t = W_v \mathbf{x}_t$$
$$\mathbf{k}_t = W_k \mathbf{x}_t$$

- Linear Attention 就是沒有 “Reflection” $f_{A,t}$ 的 RNN
- RNN 就是 Linear Attention 加上 “Reflection” $f_{A,t}$

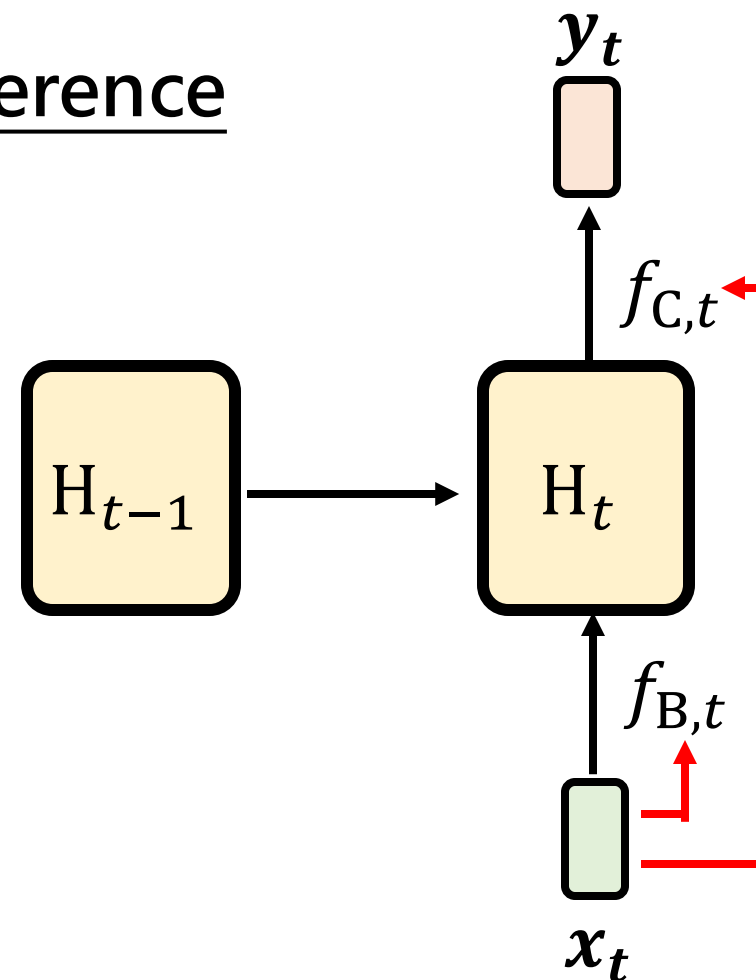
Linear Attention

Training 的時候像 Self-attention
Inference 的時候像 RNN

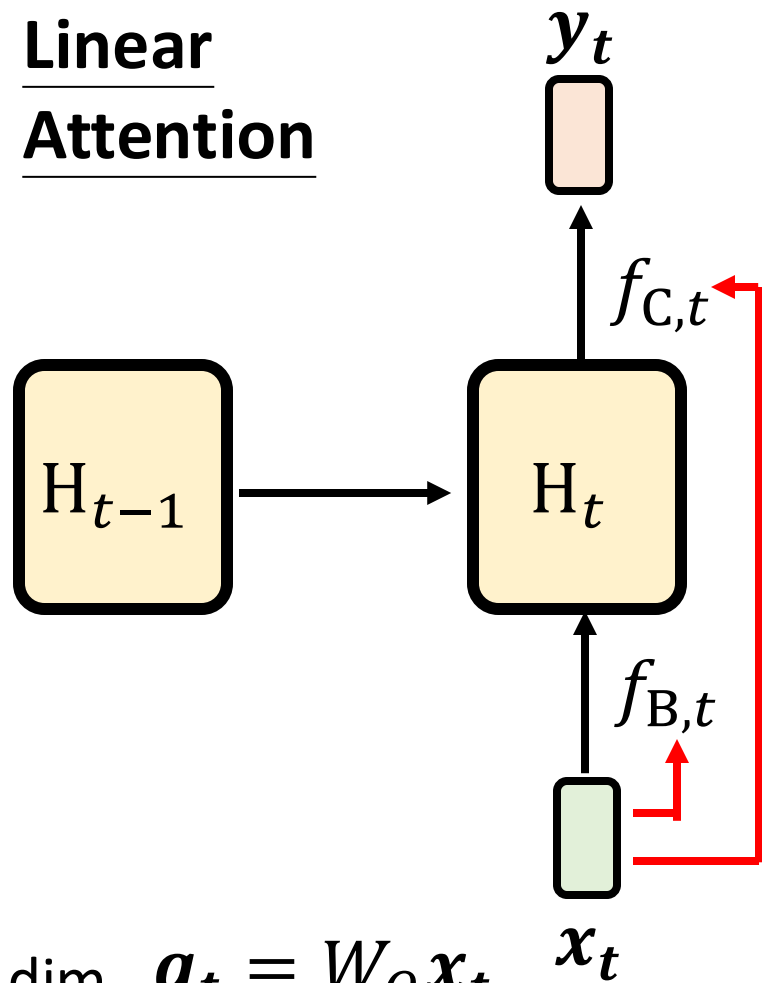
Training



Inference



Linear Attention



$$d \text{ dim } \mathbf{q}_t = W_Q \mathbf{x}_t$$

$$d \text{ dim } \mathbf{k}_t = W_K \mathbf{x}_t$$

$$d' \text{ dim } \mathbf{v}_t = W_V \mathbf{x}_t$$

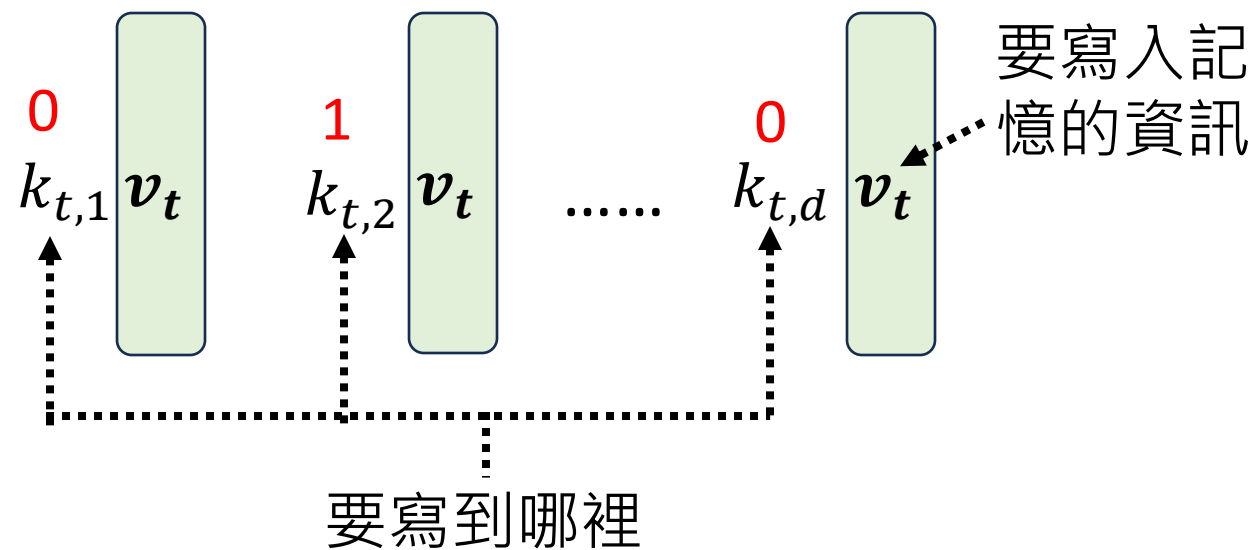
41 / 62

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t) \quad f_{B,t}(\mathbf{x}_t) = \mathbf{v}_t \mathbf{k}_t^T$$

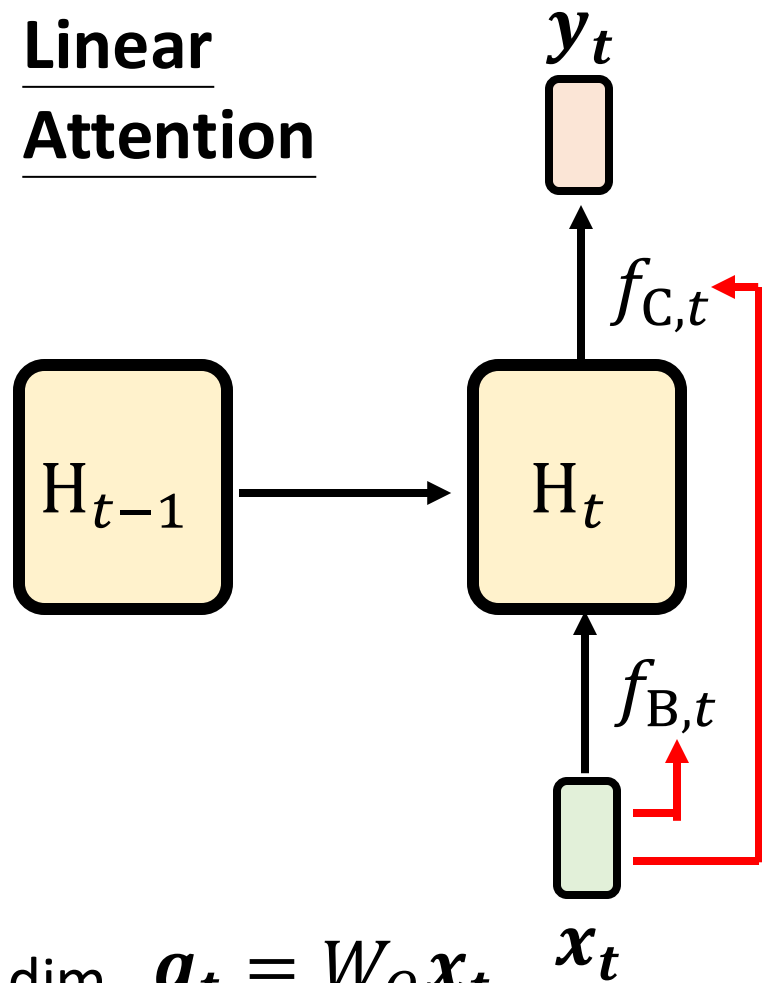
$$\mathbf{y}_t = f_{C,t}(H_t) \quad f_{C,t}(H_t) = H_t \mathbf{q}_t$$

$$H_t = H_{t-1} + d' \mathbf{v}_t \mathbf{k}_t^T$$

把 \mathbf{v}_t 寫入 H 的 2nd column



Linear Attention



$$d \text{ dim } \mathbf{q}_t = W_Q \mathbf{x}_t$$

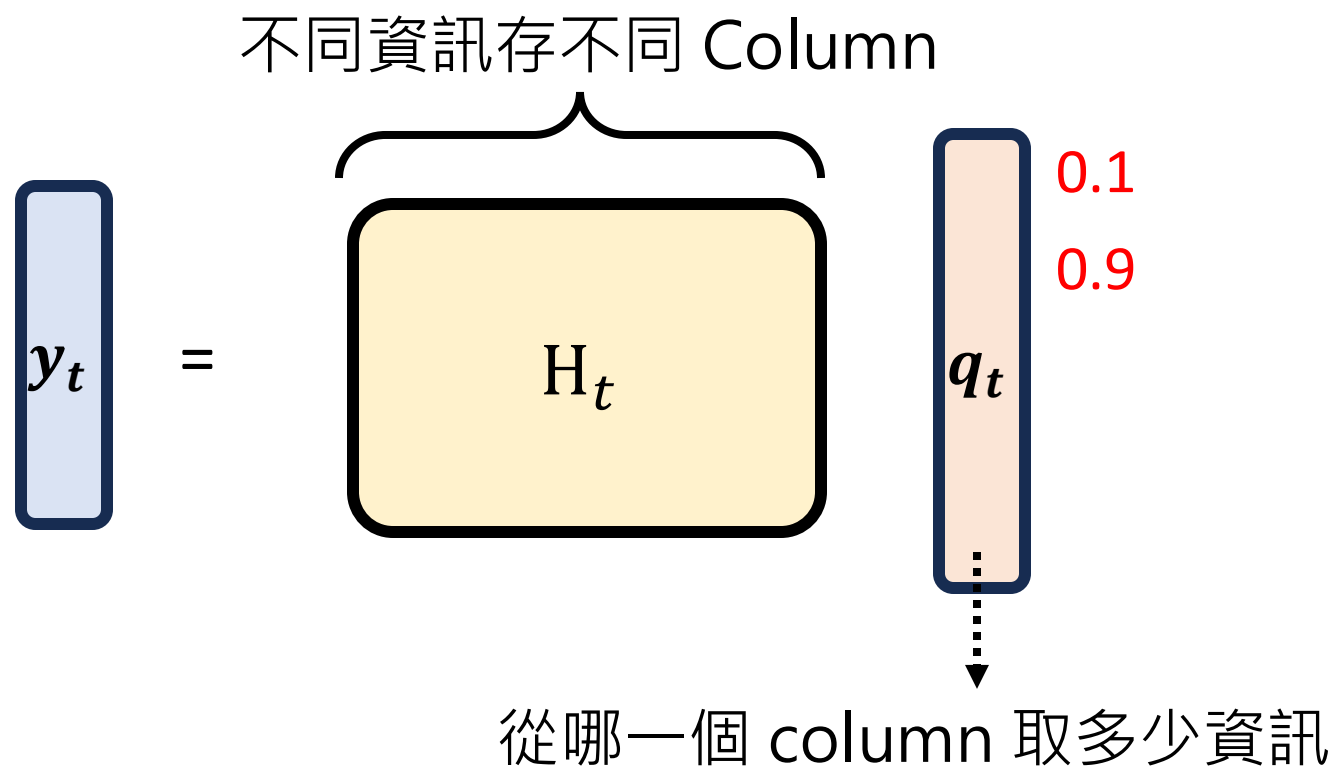
$$d \text{ dim } \mathbf{k}_t = W_K \mathbf{x}_t$$

$$d' \text{ dim } \mathbf{v}_t = W_V \mathbf{x}_t$$

42 / 62

$$H_t = H_{t-1} + f_{B,t}(\mathbf{x}_t) \quad f_{B,t}(\mathbf{x}_t) = \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{y}_t = f_{C,t}(H_t) \quad f_{C,t}(H_t) = H_t \mathbf{q}_t$$



這不是甚麼新想法

Transformers are RNNs: Fast Autoregressive
Transformers with Linear Attention

<https://arxiv.org/abs/2006.16236>

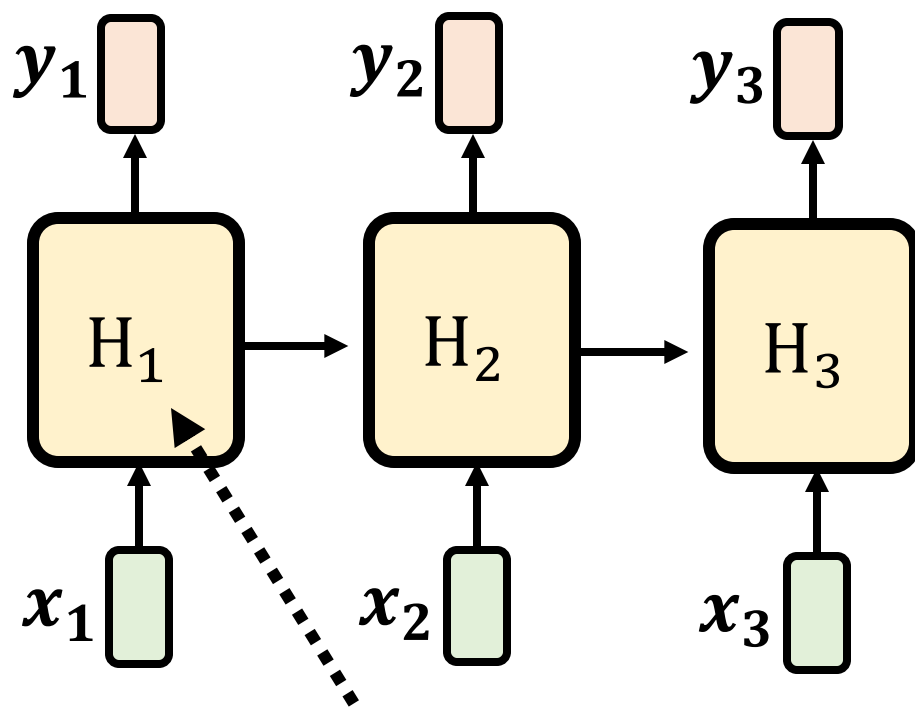
Linear Attention 的變形可
以近似 Softmax

https://youtu.be/yHoAq1IT_og?si=pSymySFnZqQj51lk



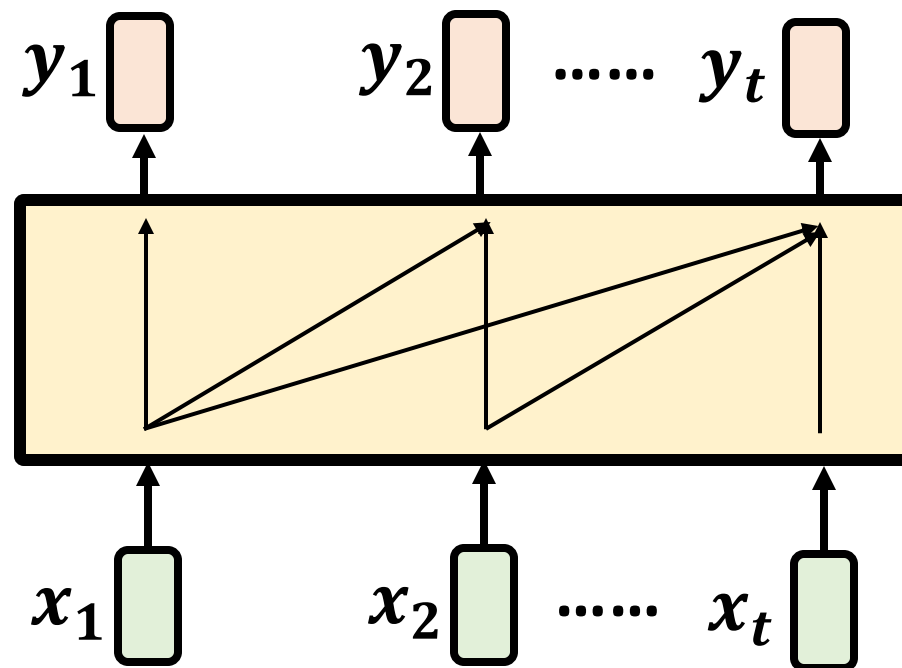
RNN (Linear Attention) 贏不過 Transformer (Self-attention with Softmax) ?

RNN (Linear Attention)



~~記憶太小~~ 記憶有限

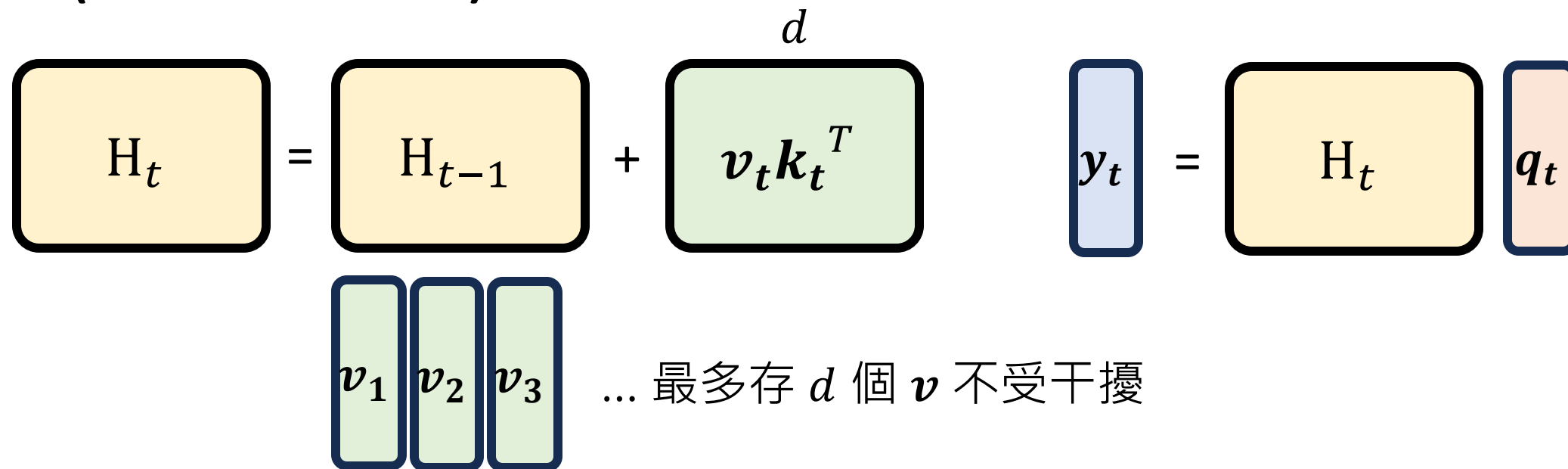
Transformer (Self-attention with softmax)



無限記憶?

RNN (Linear Attention) 贏不過 Transformer (Self-attention with Softmax) ?

RNN (Linear Attention)

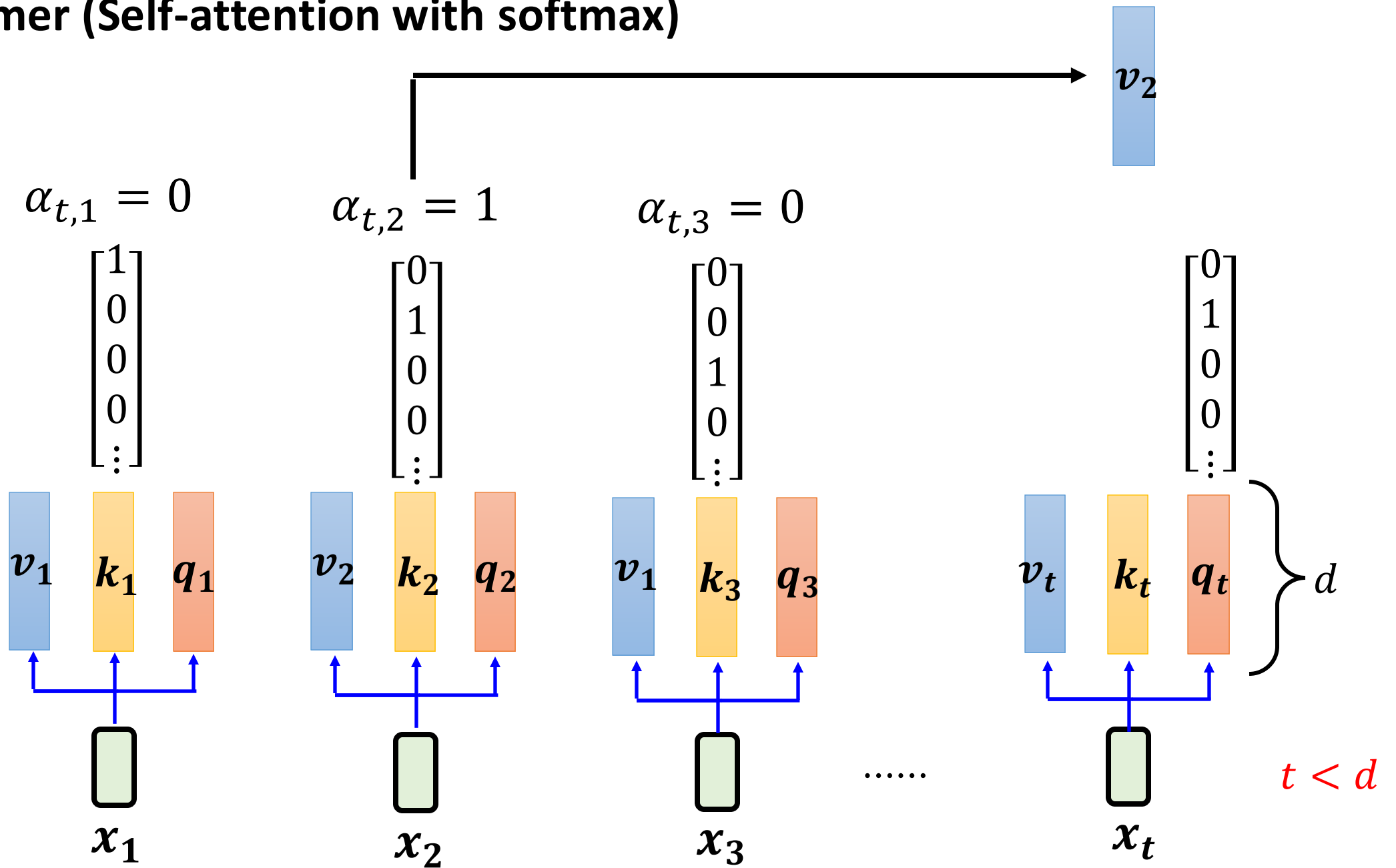


$$k_1^T = [1 \quad 0 \quad \dots]$$

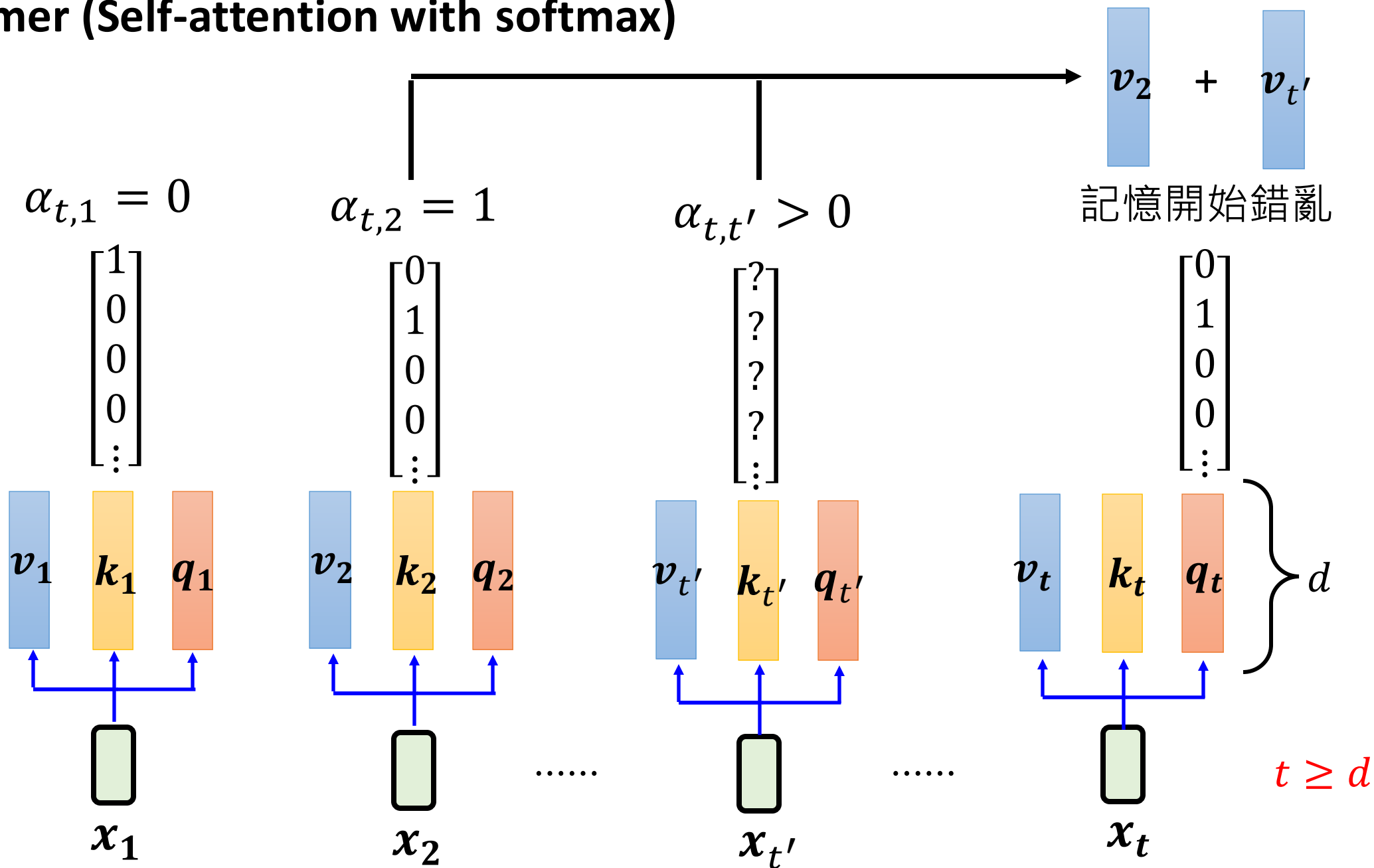
$$k_2^T = [0 \quad 1 \quad \dots]$$

$$k_3^T = [0 \quad 0 \quad 1 \quad \dots]$$

Transformer (Self-attention with softmax)



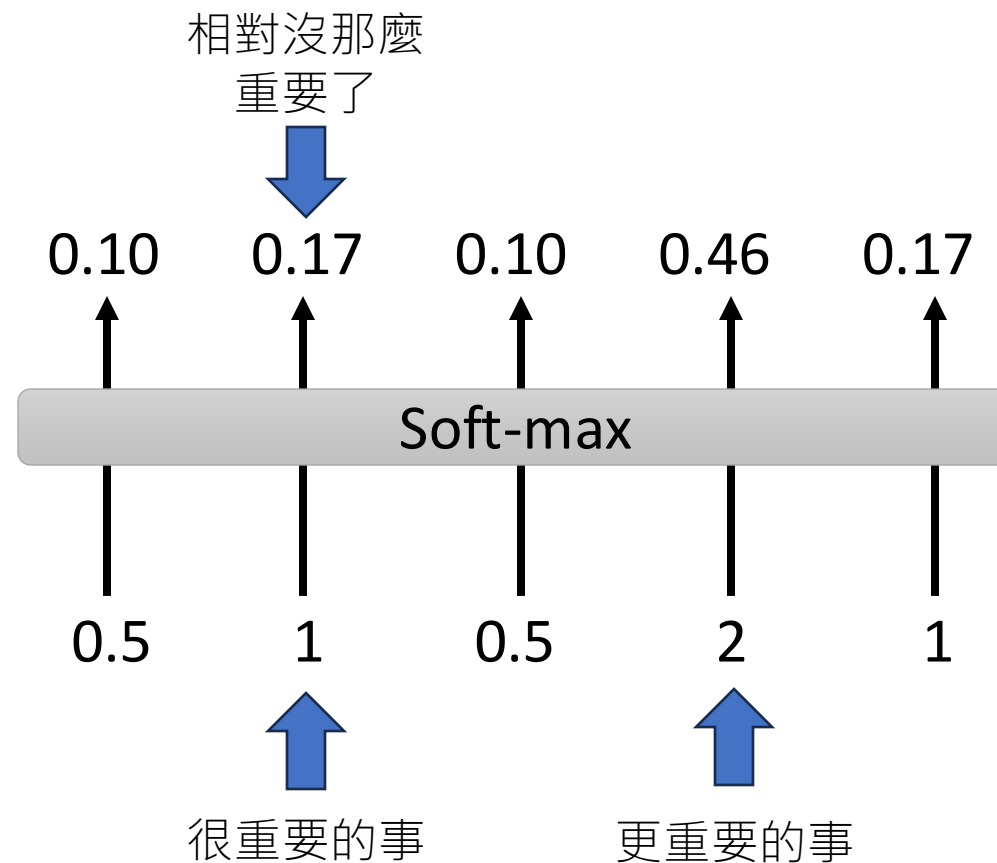
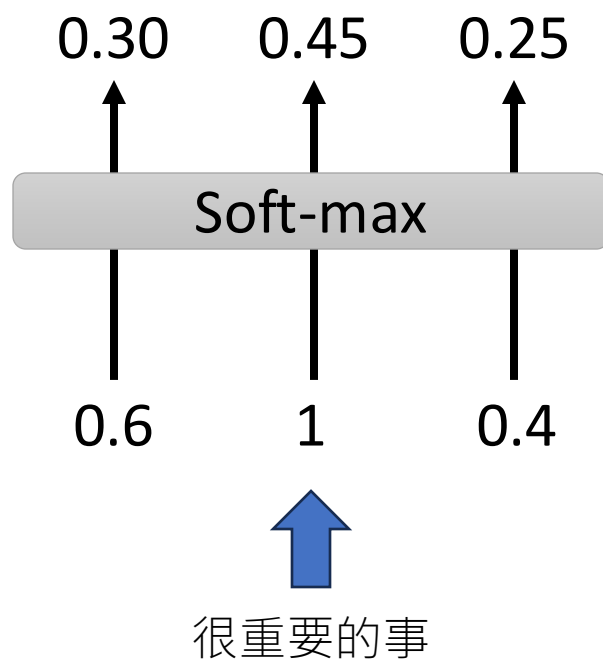
Transformer (Self-attention with softmax)



RNN (Linear Attention) 贏不過 Transformer (Self-attention with Softmax) ?

$$H_t = H_{t-1} + f_{B,t}(x_t)$$

Linear Attention 永不遺忘



加上 Reflection: 逐漸遺忘

Linear Attention

$$\mathbf{H}_t = \mathbf{H}_{t-1} + \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{q}_t$$

$$\mathbf{v}_t = W_v \mathbf{x}_t$$

$$\mathbf{k}_t = W_k \mathbf{x}_t$$

$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

Retention Network (RetNet)

$$\mathbf{H}_t = \gamma \mathbf{H}_{t-1} + \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{q}_t$$

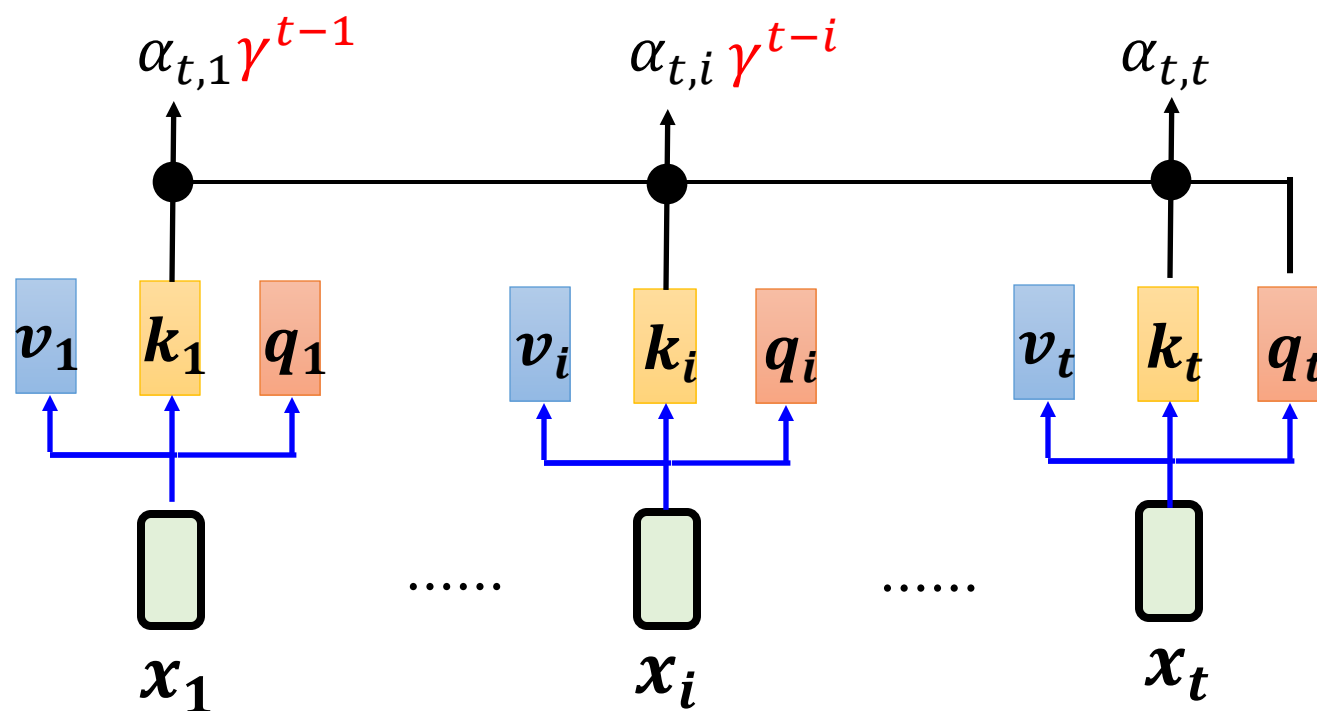
$$\mathbf{v}_t = W_v \mathbf{x}_t$$

$$\mathbf{k}_t = W_k \mathbf{x}_t$$

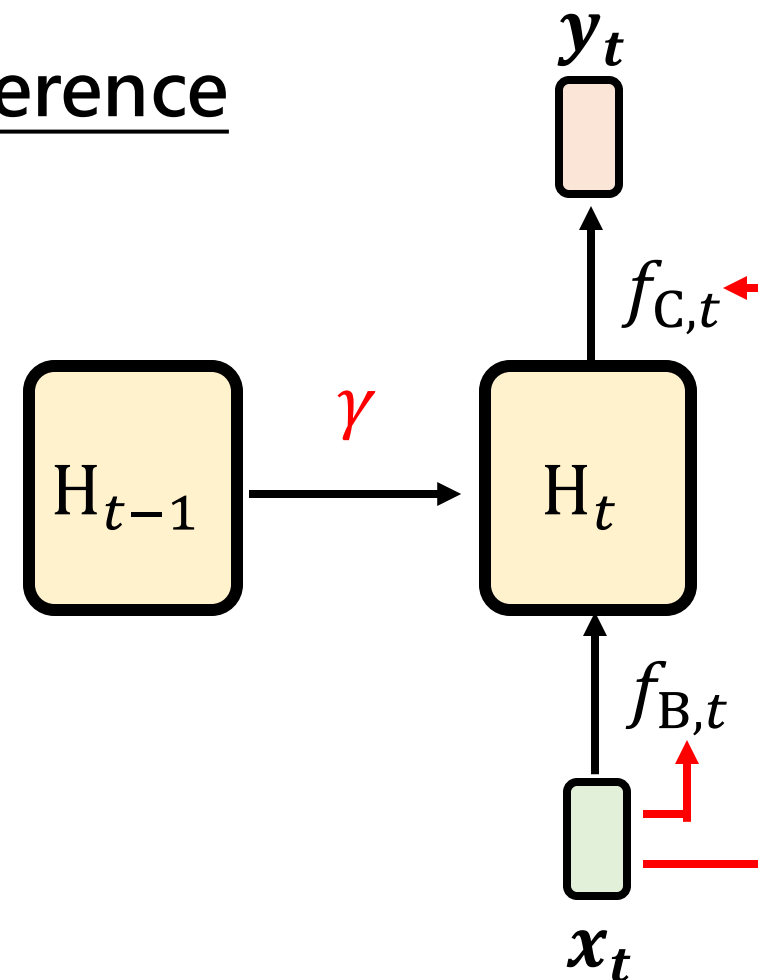
$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

加上 Reflection: 逐漸遺忘

Training



Inference



加上 Reflection: 根據情況遺忘

<https://arxiv.org/abs/2405.05254>

Retention Network (RetNet)

$$\mathbf{H}_t = \gamma \mathbf{H}_{t-1} + \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{q}_t$$

$$\mathbf{v}_t = W_v \mathbf{x}_t$$

$$\mathbf{k}_t = W_k \mathbf{x}_t$$

$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

Gated Retention

$$\mathbf{H}_t = \gamma_t \mathbf{H}_{t-1} + \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{y}_t = \mathbf{H}_t \mathbf{q}_t$$

$$\mathbf{v}_t = W_v \mathbf{x}_t$$

$$\mathbf{k}_t = W_k \mathbf{x}_t$$

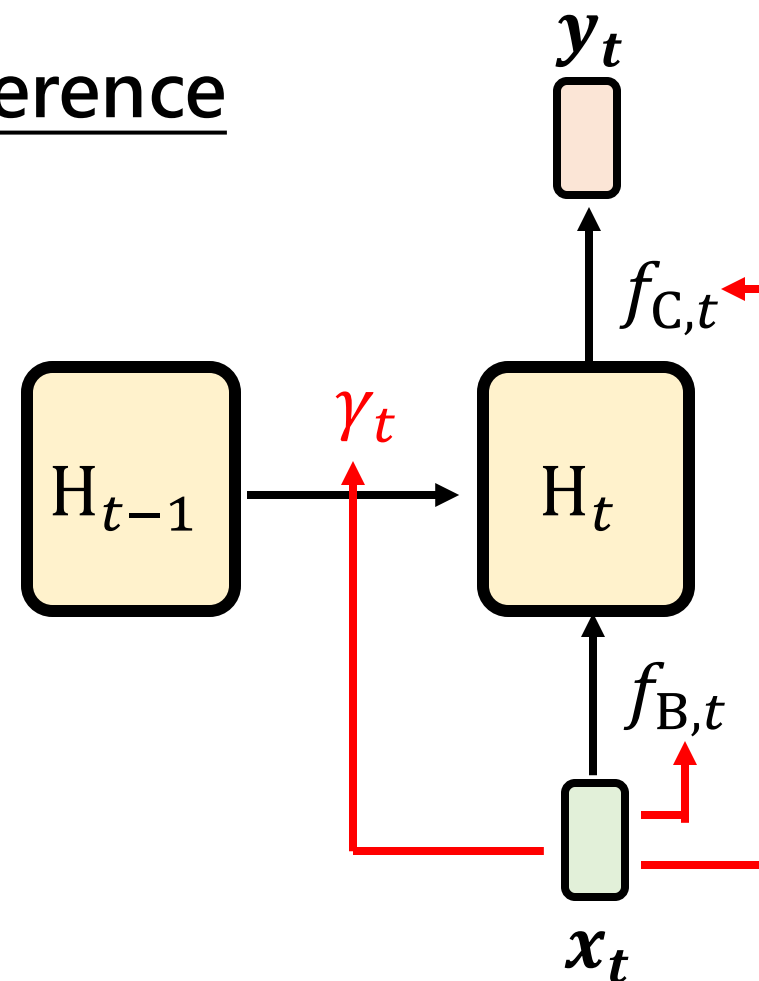
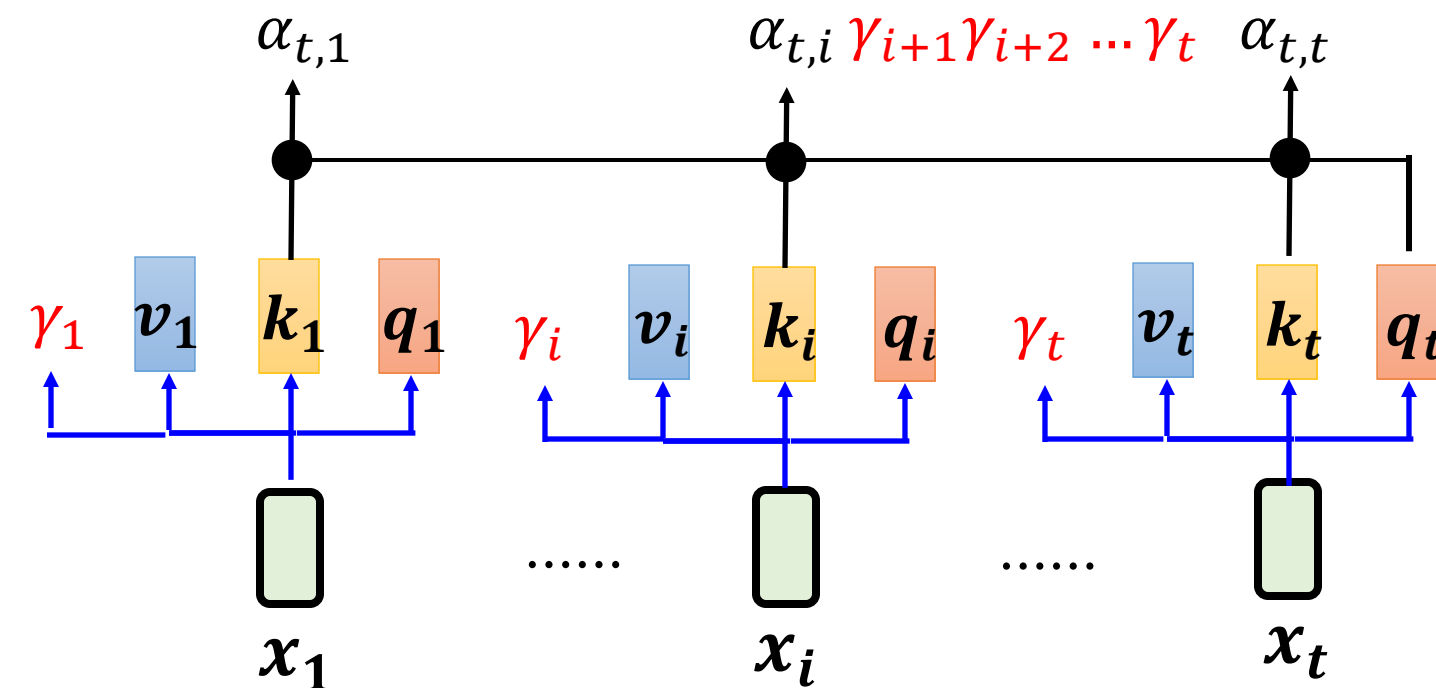
$$\mathbf{q}_t = W_Q \mathbf{x}_t$$

$$\gamma_t = \text{sigmoid}(W_\gamma \mathbf{x}_t)$$

加上 Reflection: 逐漸遺忘

Training

Inference



對 Reflection 做一點限制

$$\mathbf{H}_t = \mathbf{G}_t \odot \mathbf{H}_{t-1} + \mathbf{v}_t \mathbf{k}_t^T$$

$$\mathbf{G}_t = \mathbf{e}_t \mathbf{s}_t^T$$

$$\mathbf{G}_t = \mathbf{1} \mathbf{s}_t^T$$

$$\mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$\mathbf{s}_t^T = [0 \quad 1 \quad 0.1 \quad \dots \dots]$$

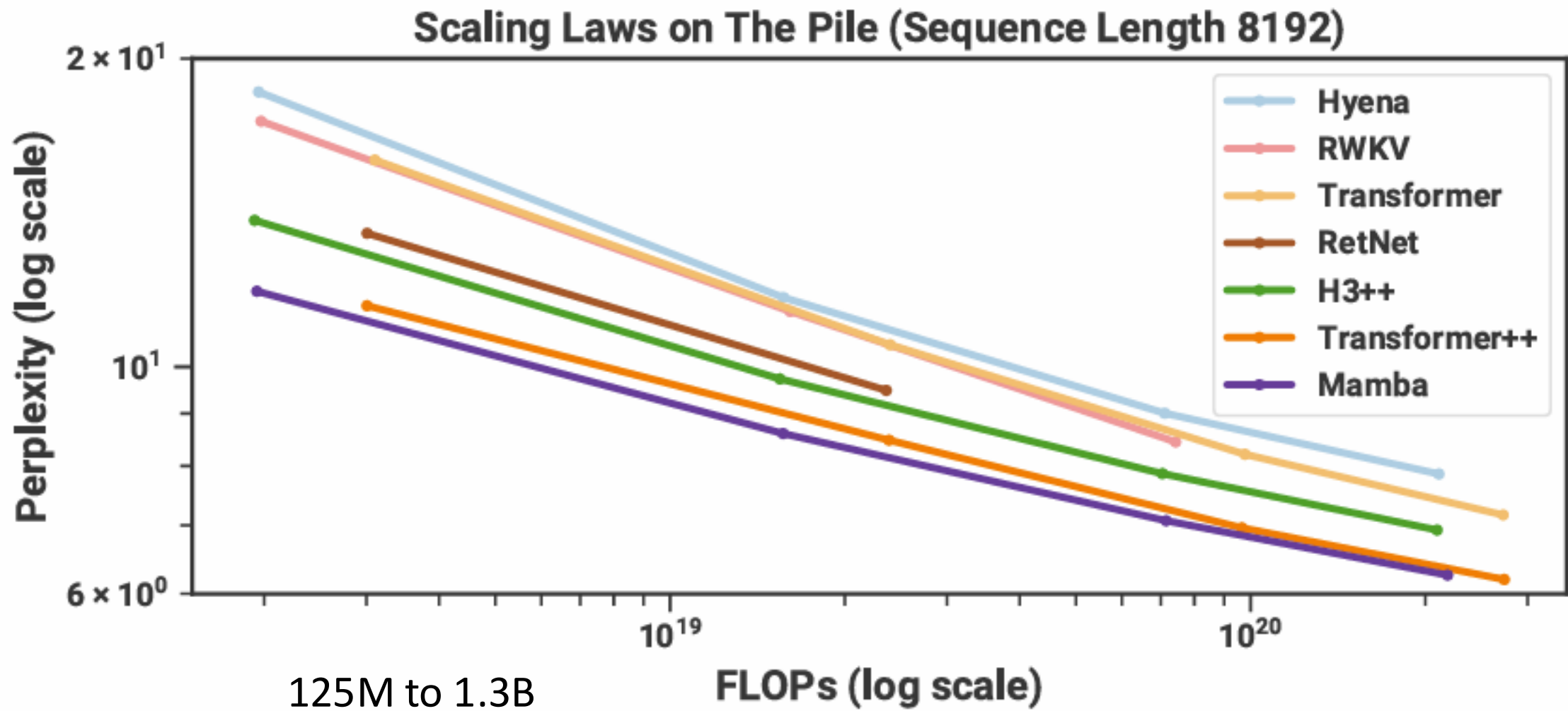
$$\mathbf{1} \mathbf{s}_t^T \odot \mathbf{H}_{t-1}$$

$s_{t,1}$	$s_{t,2}$	$s_{t,3}$	$s_{t,d}$
$s_{t,1}$	$s_{t,2}$	$s_{t,3}$	$s_{t,d}$
\vdots	\vdots	\vdots	\vdots
$s_{t,1}$	$s_{t,2}$	$s_{t,3}$	$s_{t,d}$
0	1	0.1		
抹去	保留	減弱		

Model	Parameterization	Learnable parameters
Mamba (Gu & Dao, 2023)	$\mathbf{G}_t = \exp(-(\mathbf{1}^\top \boldsymbol{\alpha}_t) \odot \exp(\mathbf{A})), \quad \boldsymbol{\alpha}_t = \text{softplus}(\mathbf{x}_t \mathbf{W}_{\alpha_1} \mathbf{W}_{\alpha_2})$	$\mathbf{A} \in \mathbb{R}^{d_k \times d_v}, \quad \mathbf{W}_{\alpha_1} \in \mathbb{R}^{d \times \frac{d}{16}}, \quad \mathbf{W}_{\alpha_2} \in \mathbb{R}^{\frac{d}{16} \times d_v}$
Mamba-2 (Dao & Gu, 2024)	$\mathbf{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}, \quad \gamma_t = \exp(-\text{softplus}(\mathbf{x}_t \mathbf{W}_\gamma) \exp(a))$	$\mathbf{W}_\gamma \in \mathbb{R}^{d \times 1}, \quad a \in \mathbb{R}$
mLSTM (Beck et al., 2024; Peng et al., 2021)	$\mathbf{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}, \quad \gamma_t = \sigma(\mathbf{x}_t \mathbf{W}_\gamma)$	$\mathbf{W}_\gamma \in \mathbb{R}^{d \times 1}$
Gated Retention (Sun et al., 2024)	$\mathbf{G}_t = \gamma_t \mathbf{1}^\top \mathbf{1}, \quad \gamma_t = \sigma(\mathbf{x}_t \mathbf{W}_\gamma)^{\frac{1}{\tau}}$	$\mathbf{W}_\gamma \in \mathbb{R}^{d \times 1}$
DFW (Mao, 2022; Pramanik et al., 2023)	$\mathbf{G}_t = \boldsymbol{\alpha}_t^\top \boldsymbol{\beta}_t, \quad \boldsymbol{\alpha}_t = \sigma(\mathbf{x}_t \mathbf{W}_\alpha), \quad \boldsymbol{\beta}_t = \sigma(\mathbf{x}_t \mathbf{W}_\beta)$	$\mathbf{W}_\alpha \in \mathbb{R}^{d \times d_k}, \quad \mathbf{W}_\beta \in \mathbb{R}^{d \times d_v}$
GateLoop (Katsch, 2023)	$\mathbf{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}, \quad \boldsymbol{\alpha}_t = \sigma(\mathbf{x}_t \mathbf{W}_{\alpha_1}) \exp(\mathbf{x}_t \mathbf{W}_{\alpha_2} \mathbf{i})$	$\mathbf{W}_{\alpha_1} \in \mathbb{R}^{d \times d_k}, \quad \mathbf{W}_{\alpha_2} \in \mathbb{R}^{d \times d_k}$
HGRN-2 (Qin et al., 2024b)	$\mathbf{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}, \quad \boldsymbol{\alpha}_t = \gamma + (1 - \gamma) \sigma(\mathbf{x}_t \mathbf{W}_\alpha)$	$\mathbf{W}_\alpha \in \mathbb{R}^{d \times d_k}, \quad \gamma \in (0, 1)^{d_k}$
RWKV-6 (Peng et al., 2024)	$\mathbf{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}, \quad \boldsymbol{\alpha}_t = \exp(-\exp(\mathbf{x}_t \mathbf{W}_\alpha))$	$\mathbf{W}_\alpha \in \mathbb{R}^{d \times d_k}$
Gated Linear Attention (GLA)	$\mathbf{G}_t = \boldsymbol{\alpha}_t^\top \mathbf{1}, \quad \boldsymbol{\alpha}_t = \sigma(\mathbf{x}_t \mathbf{W}_{\alpha_1} \mathbf{W}_{\alpha_2})^{\frac{1}{\tau}}$	$\mathbf{W}_{\alpha_1} \in \mathbb{R}^{d \times 16}, \quad \mathbf{W}_{\alpha_2} \in \mathbb{R}^{16 \times d_k}$

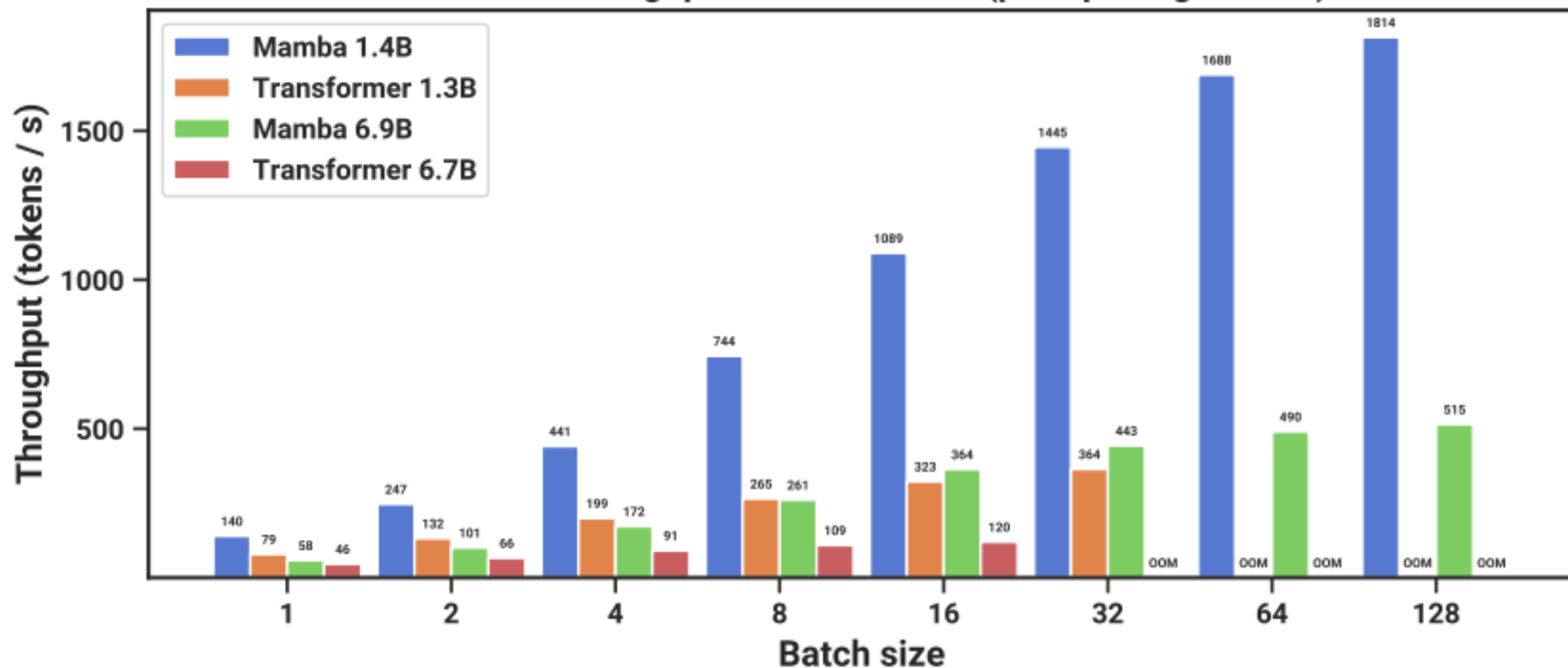
<https://arxiv.org/abs/2312.06635>

Model	Recurrence	Memory read-out
Linear Attention [48, 47]	$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
+ Kernel	$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{v}_t \phi(\mathbf{k}_t)^\top$	$\mathbf{o}_t = \mathbf{S}_t \phi(\mathbf{q}_t)$
+ Normalization	$\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{v}_t \phi(\mathbf{k}_t)^\top, \mathbf{z}_t = \mathbf{z}_{t-1} + \phi(\mathbf{k}_t)$	$\mathbf{o}_t = \mathbf{S}_t \phi(\mathbf{q}_t) / (\mathbf{z}_t^\top \phi(\mathbf{q}_t))$
DeltaNet [101]	$\mathbf{S}_t = \mathbf{S}_{t-1}(\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
Gated RFA [81]	$\mathbf{S}_t = g_t \mathbf{S}_{t-1} + (1 - g_t) \mathbf{v}_t \mathbf{k}_t^\top, \mathbf{z}_t = g_t \mathbf{z}_{t-1} + (1 - g_t) \mathbf{k}_t$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t / (\mathbf{z}_t^\top \mathbf{q}_t)$
S4 [32, 106]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot \exp(-(\boldsymbol{\alpha} \mathbf{1}^\top) \odot \exp(\mathbf{A})) + \mathbf{B} \odot (\mathbf{v}_t \mathbf{1}^\top)$	$\mathbf{o}_t = (\mathbf{S}_t \odot \mathbf{C}) \mathbf{1} + \mathbf{d} \odot \mathbf{v}_t$
ABC [82]	$\mathbf{S}_t^k = \mathbf{S}_{t-1}^k + \mathbf{k}_t \phi_t^\top, \mathbf{S}_t^v = \mathbf{S}_{t-1}^v + \mathbf{v}_t \phi_t^\top$	$\mathbf{o}_t = \mathbf{S}_t^v \text{softmax}(\mathbf{S}_t^k \mathbf{q}_t)$
DFW [63]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot (\beta_t \boldsymbol{\alpha}_t^\top) + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
RetNet [108]	$\mathbf{S}_t = \gamma \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
Mamba [31]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot \exp(-(\boldsymbol{\alpha}_t \mathbf{1}^\top) \odot \exp(\mathbf{A})) + (\boldsymbol{\alpha}_t \odot \mathbf{v}_t) \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t + \mathbf{d} \odot \mathbf{v}_t$
GLA [124]	$\mathbf{S}_t = \mathbf{S}_{t-1} \odot (\mathbf{1} \boldsymbol{\alpha}_t^\top) + \mathbf{v}_t \mathbf{k}_t^\top = \mathbf{S}_{t-1} \text{Diag}(\boldsymbol{\alpha}_t) + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
RWKV-6 [79]	$\mathbf{S}_t = \mathbf{S}_{t-1} \text{Diag}(\boldsymbol{\alpha}_t) + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = (\mathbf{S}_{t-1} + (\mathbf{d} \odot \mathbf{v}_t) \mathbf{k}_t^\top) \mathbf{q}_t$
HGRN-2 [92]	$\mathbf{S}_t = \mathbf{S}_{t-1} \text{Diag}(\boldsymbol{\alpha}_t) + \mathbf{v}_t (\mathbf{1} - \boldsymbol{\alpha}_t)^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
mLSTM [9]	$\mathbf{S}_t = f_t \mathbf{S}_{t-1} + i_t \mathbf{v}_t \mathbf{k}_t^\top, \mathbf{z}_t = f_t \mathbf{z}_{t-1} + i_t \mathbf{k}_t$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t / \max\{1, \mathbf{z}_t^\top \mathbf{q}_t \}$
Mamba-2 [19]	$\mathbf{S}_t = \gamma_t \mathbf{S}_{t-1} + \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$
GSA [131]	$\mathbf{S}_t^k = \mathbf{S}_{t-1}^k \text{Diag}(\boldsymbol{\alpha}_t) + \mathbf{k}_t \phi_t^\top, \mathbf{S}_t^v = \mathbf{S}_{t-1}^v \text{Diag}(\boldsymbol{\alpha}_t) + \mathbf{v}_t \phi_t^\top$	$\mathbf{o}_t = \mathbf{S}_t^v \text{softmax}(\mathbf{S}_t^k \mathbf{q}_t)$
Gated DeltaNet [125]	$\mathbf{S}_t = \mathbf{S}_{t-1} \left(\boldsymbol{\alpha}_t (\mathbf{I} - \beta_t \mathbf{k}_t \mathbf{k}_t^\top) \right) + \beta_t \mathbf{v}_t \mathbf{k}_t^\top$	$\mathbf{o}_t = \mathbf{S}_t \mathbf{q}_t$

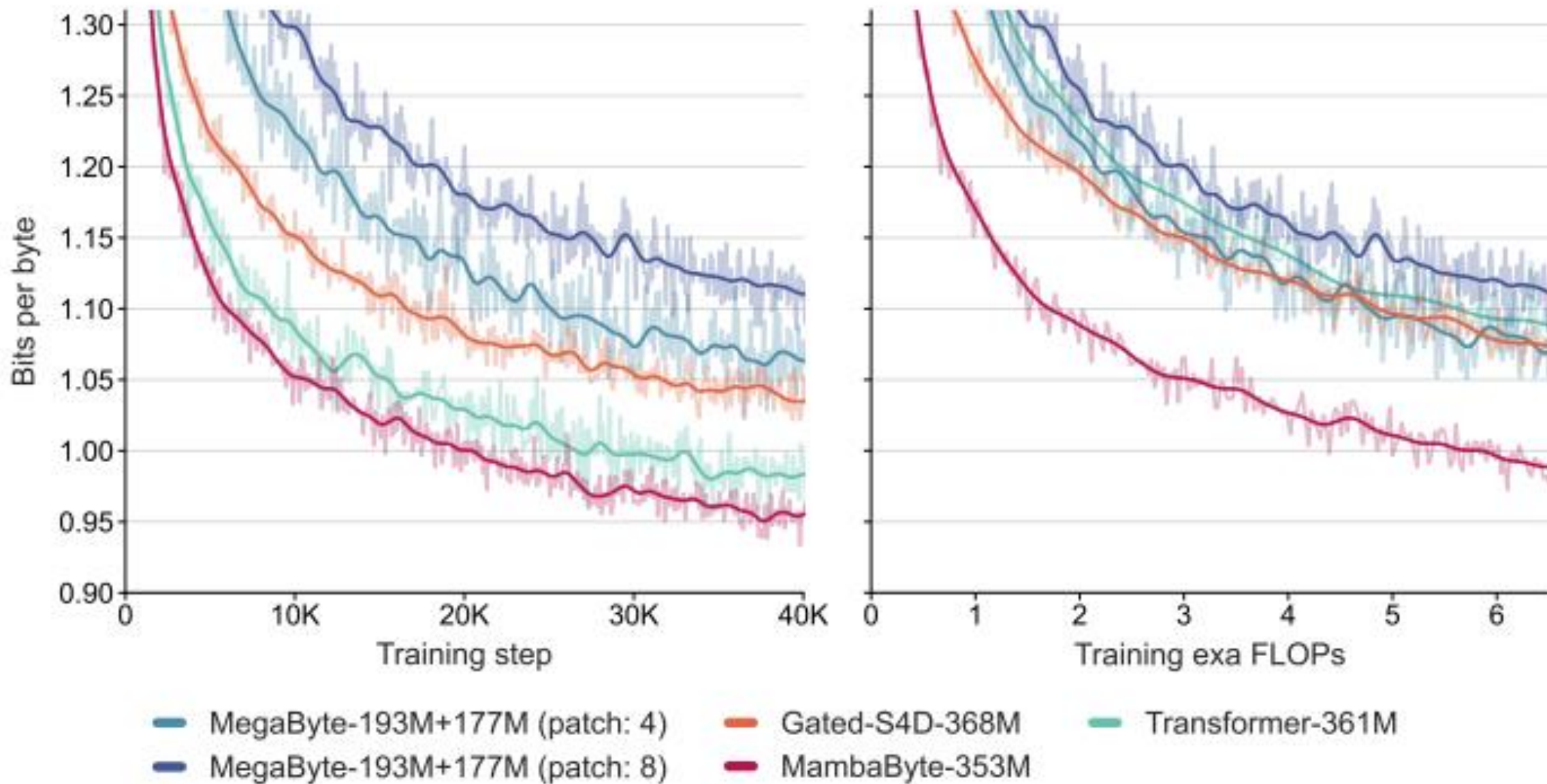


<https://arxiv.org/abs/2312.00752>

Inference throughput on A100 80GB (prompt length 2048)



<https://arxiv.org/abs/2312.00752>



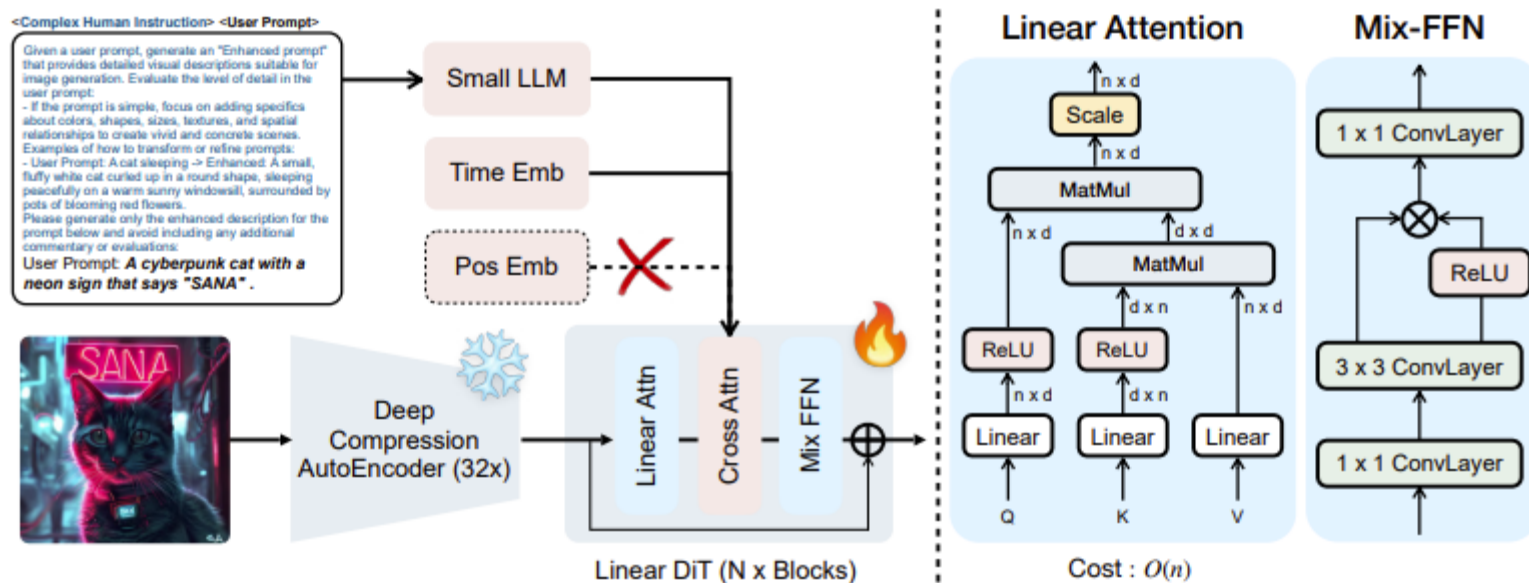
Name	Modality	Affiliations	Sizes	Access Link
Mamba 1&2	Language	Carnegie Mellon University & Princeton University	130M-2.8B	1
Falcon Mamba 7B	Language	Technology Innovation Institute	7B	2
Mistral 7B	Language	Mistral AI & NVIDIA	7B	3
Jamba	Language	AI21 Lab	12B/52B	4
Vision Mamba	Vision	Huazhong University of Science and Technology	7M-98M	5
VideoMamba	Video	OpenGVLab, Shanghai AI Laboratory	28M-392M	6
Codestral Mamba	Code	Mistral AI	7B, 22B	7

1. <https://github.com/state-spaces/mamba>
2. <https://huggingface.co/tiiuae/falcon-mamba-7b>
3. <https://huggingface.co/mistralai/Mistral-7B-v0.1>
4. <https://huggingface.co/ai21labs/Jamba-v0.1>
5. <https://huggingface.co/hustvl/Vim-base-midclstok>
6. <https://huggingface.co/OpenGVLab/VideoMamba>
7. <https://mistral.ai/news/codestral-mamba/>

<https://arxiv.org/abs/2408.01129>

Minimax-01

<https://arxiv.org/abs/2501.08313>



(a). Architecture overview of our Sana.

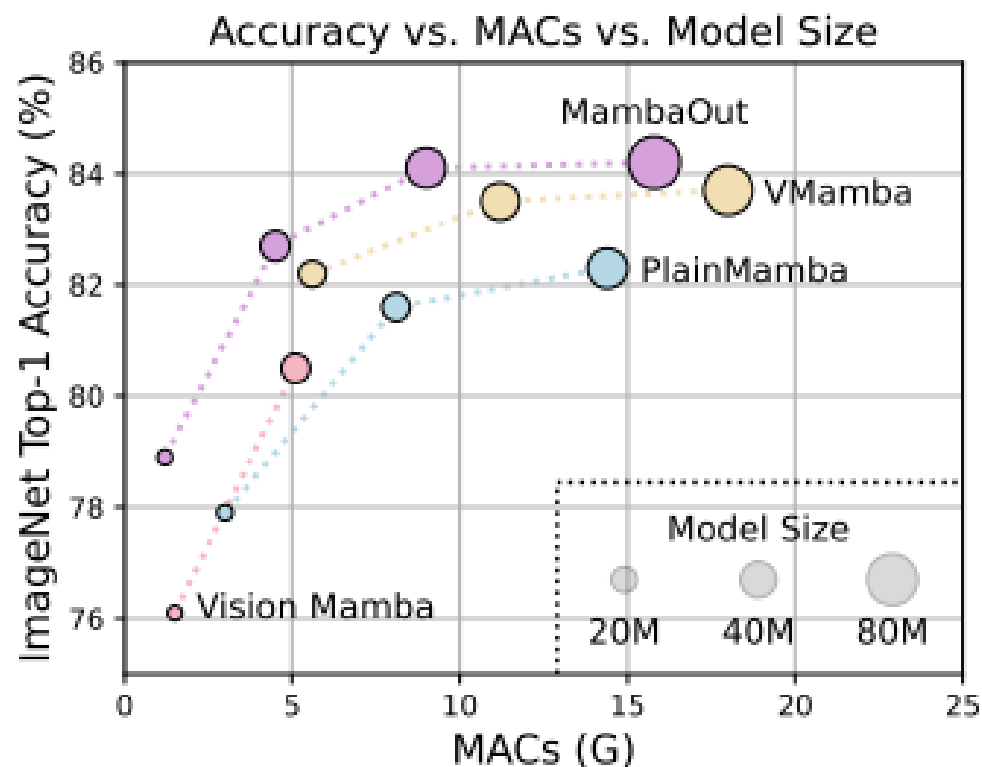
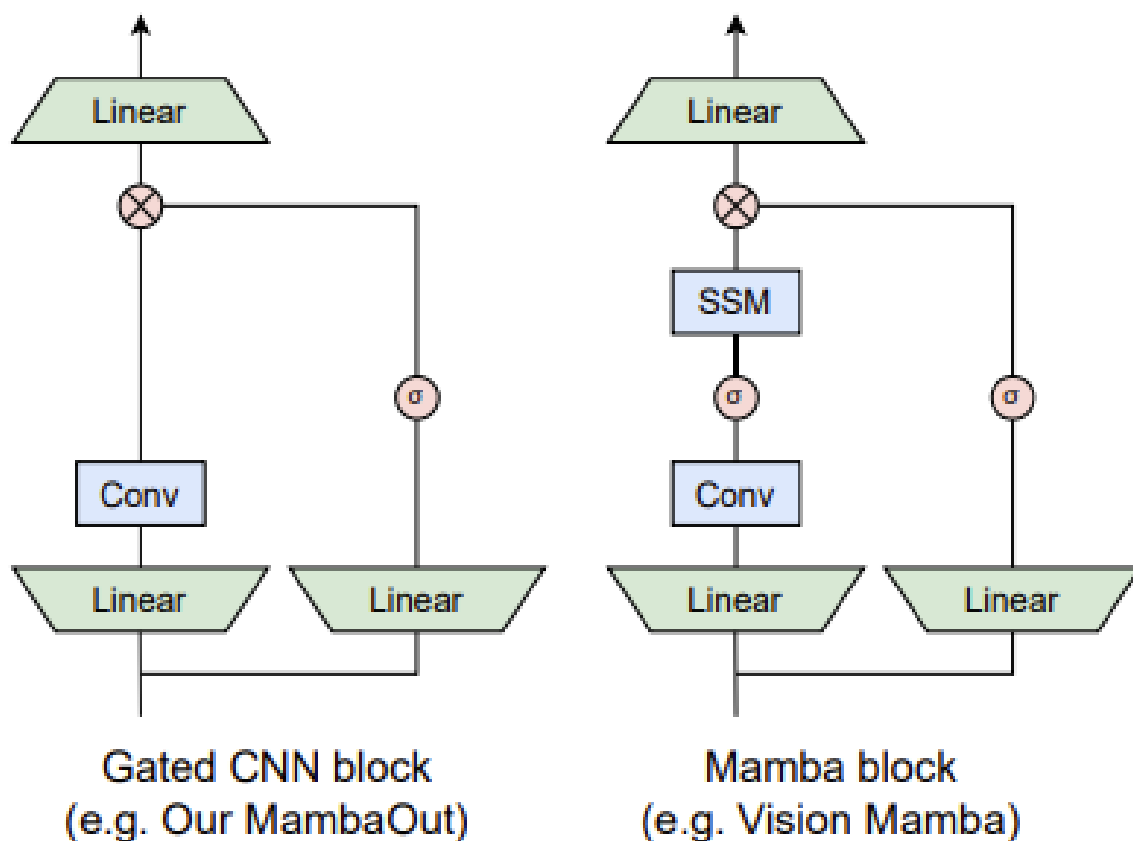
(b). Linear DiT Module.

MambaOut: Do We Really Need Mamba for Vision?

<https://arxiv.org/abs/2405.07992>

In memory of Kobe Bryant

“What can I say, Mamba out.” — Kobe Bryant’s NBA farewell speech in 2016.



Do not train from scratch

Low-rank Linear Conversion via Attention Transfer (LoLCATs), <https://arxiv.org/abs/2410.10254>

The Mamba in the Llama,
<https://arxiv.org/abs/2408.15237>

Transformers to SSMs,
<https://arxiv.org/abs/2408.10189>

Linger, <https://arxiv.org/abs/2503.01496>

