西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Mamba

Mamba: Linear-Time Sequence Modeling with Selective State Spaces

Reject By ICLR 2024

Albert Gu (Carnegie Mellon University) and

Tri Dao (Princeton University)

主讲人：阮皓
2024.04.17

Why Reject

一是缺少LRA（Long Range Arena）**评估**，公认的长序列建模基准。

二是仅将困惑度评估作为主要评价指标不行，理由是低困惑度与生成性能不一定正相关。

一篇论文被会议接收与否与它对社区的价值贡献并不挂钩。因为前者很容易依赖于极少数人的判断。

NeurIPS Conference
@NeurIPSConf
Following ···

Test of Time
Distributed Representations of Words and Phrases and their Compositionality

[Submitted on 1 Dec 2023]

Mamba: Linear-time sequence modeling with selective state spaces
A Gu, T Dao
arXiv preprint arXiv:2312.00752, 2023 · arxiv.org

Foundation models, now powering most of the exciting applications in deep learning, are almost universally based on the Transformer architecture and its core attention module. Many subquadratic-time architectures such as linear attention, gated convolution and recurrent models, and structured state space models (SSMs) have been developed to address Transformers' computational inefficiency on long sequences, but they have not performed as well as attention on important modalities such as language. We identify that

展开

☆ 保存 99 引用 被引用次数：172 相关文章 所有 4 个版本 easyScholar文献收藏

Follow Work

U-mamba: Enhancing long-range dependency for biomedical image segmentation
J Ma, F Li, B Wang - arXiv preprint arXiv:2401.04722, 2024 - arxiv.org
Convolutional Neural Networks (CNNs) and Transformers have been the most popular architectures for biomedical image segmentation, but both of them have limited ability to …
☆ 保存 99 引用 被引用次数：41 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Vision mamba: Efficient visual representation learning with bidirectional state space model
L Zhu, B Liao, Q Zhang, X Wang, W Liu… - arXiv preprint arXiv …, 2024 - arxiv.org
Recently the state space models (SSMs) with efficient hardware-aware designs, ie, Mamba, have shown great potential for long sequence modeling. Building efficient and generic …
☆ 保存 99 引用 被引用次数：64 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Parameter-efficient fine-tuning for large models: A comprehensive survey
Z Han, C Gao, J Liu, SQ Zhang - arXiv preprint arXiv:2403.14608, 2024 - arxiv.org
Large models represent a groundbreaking advancement in multiple application fields, enabling remarkable achievements across various tasks. However, their unprecedented …
☆ 保存 99 引用 被引用次数：2 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Vmamba: Visual state space model
Y Liu, Y Tian, Y Zhao, H Yu, L Xie, Y Wang… - arXiv preprint arXiv …, 2024 - arxiv.org
Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) stand as the two most popular foundation models for visual representation learning. While CNNs exhibit …
☆ 保存 99 引用 被引用次数：55 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Vm-unet: Vision mamba unet for medical image segmentation
J Ruan, S Xiang - arXiv preprint arXiv:2402.02491, 2024 - arxiv.org
In the realm of medical image segmentation, both CNN-based and Transformer-based models have been extensively explored. However, CNNs exhibit limitations in long-range …
☆ 保存 99 引用 被引用次数：26 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Characterization of large language model development in the datacenter
Q Hu, Z Ye, Z Wang, G Wang, M Zhang, Q Chen… - arXiv preprint arXiv …, 2024 - arxiv.org
Large Language Models (LLMs) have presented impressive performance across several transformative tasks. However, it is non-trivial to efficiently utilize large-scale cluster …
☆ 保存 99 引用 被引用次数：4 相关文章 ≫ easyScholar文献收藏

Rsmamba: Remote sensing image classification with state space model
K Chen, B Chen, C Liu, W Li, Z Zou, Z Shi - arXiv preprint arXiv …, 2024 - arxiv.org
Remote sensing image classification forms the foundation of various understanding tasks, serving a crucial function in remote sensing image interpretation. The recent advancements …
☆ 保存 99 引用 被引用次数：3 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Clinicalmamba: A generative clinical language model on longitudinal clinical notes
Z Yang, A Mitra, S Kwon, H Yu - arXiv preprint arXiv:2403.05795, 2024 - arxiv.org
The advancement of natural language processing (NLP) systems in healthcare hinges on language model ability to interpret the intricate information contained within clinical notes …
☆ 保存 99 引用 被引用次数：3 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

Videomamba: State space model for efficient video understanding
K Li, X Li, Y Wang, Y He, Y Wang, L Wang… - arXiv preprint arXiv …, 2024 - arxiv.org
Addressing the dual challenges of local redundancy and global dependencies in video understanding, this work innovatively adapts the Mamba to the video domain. The proposed …
☆ 保存 99 引用 被引用次数：2 相关文章 所有 2 个版本 ≫ easyScholar文献收藏

从Transformer复杂度、RNN到SSM

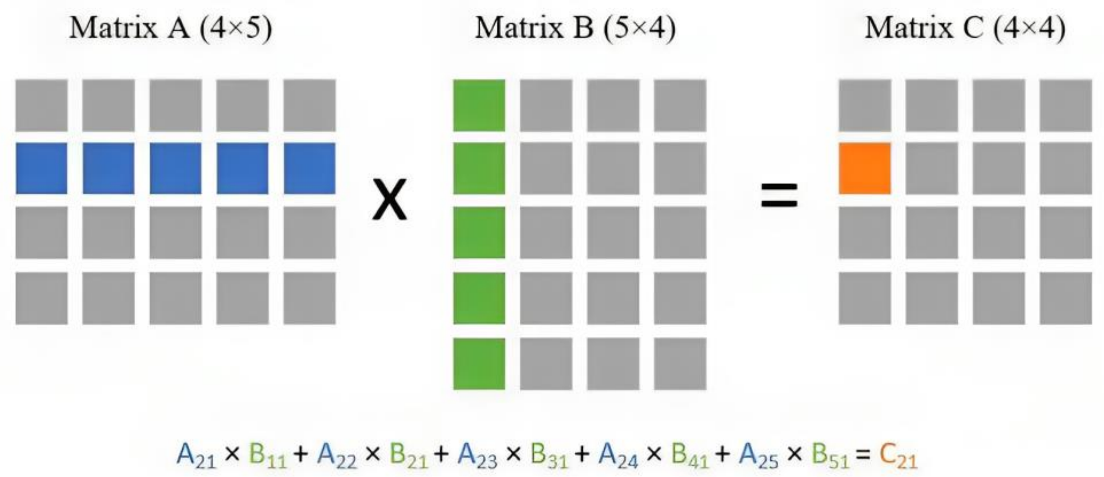
《Attention is All You Need》
NIPS 2017

RetNet, RWKV(Receptance Weighted Key Value), Linear Attention, Flash Attention, Gated Convolution, SSMs(State Space Model)……

SSM→S4→Mamba(S6)

从Transformer复杂度、RNN到SSM

Transformer的二次复杂度

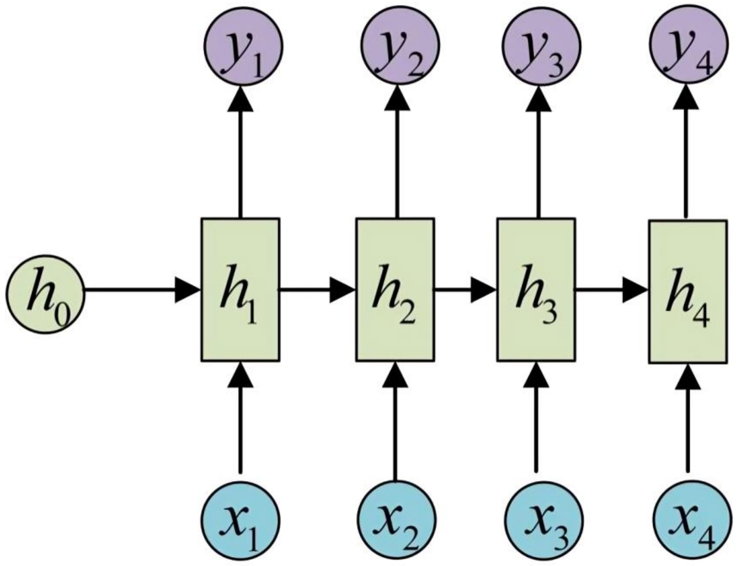


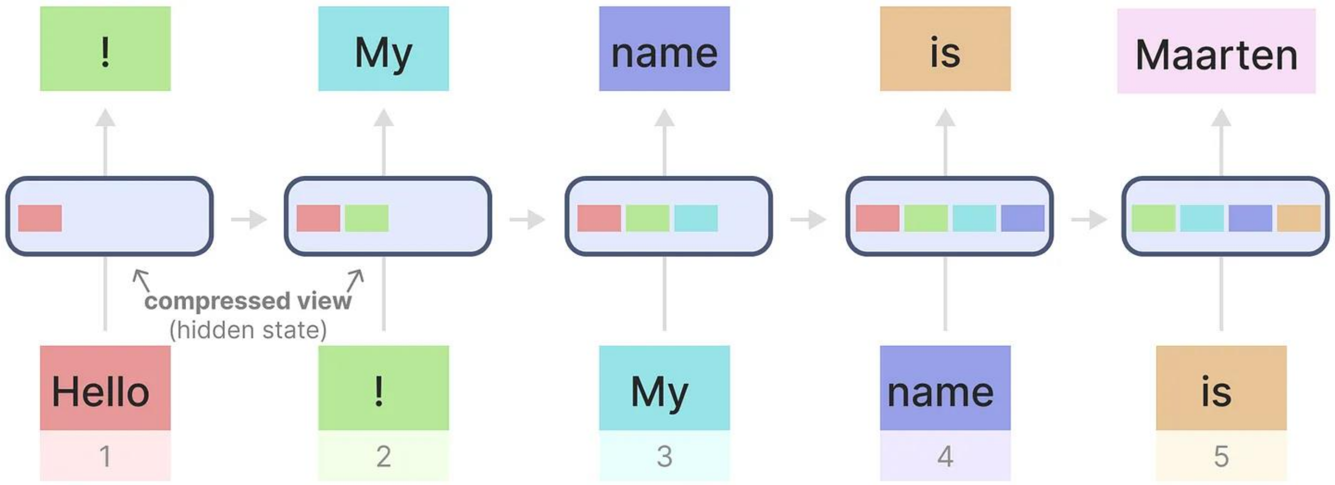
两个相乘的矩阵大小分别为$(N * d)$ 和$(d * N)$，矩阵乘法的一种计算方式是使用第一个矩阵的每一行与第二个矩阵的每一列做点乘

总共就需要 N^2 次点乘。而每次点乘又需要 d 次乘法，所以总复杂度就为 $O(N^2 d)$

- 针对注意力机制的各种所谓魔改，甚至也有S4、FlashAttention及其二代等
- S4、FlashAttention等作者提出了新的序列模型：Mamba，在很多语言任务上击败/匹配Transformer性能，具有线性复杂度和5倍推理吞吐量。

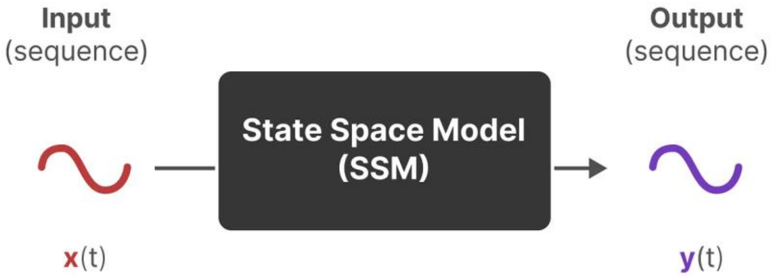
从Transformer复杂度、RNN到SSM





- RNN没法并行训练(串行的结构)，相当于推理快但训练慢
- 遗忘问题，无法有效处理长距离依赖问题
- 为何RNN没法并行训练？

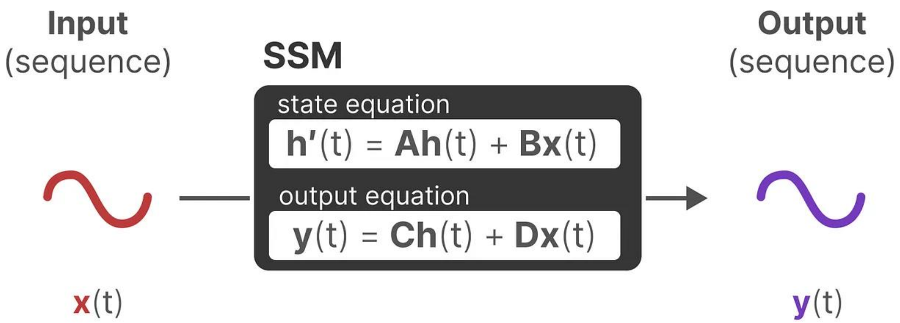
从Transformer复杂度、RNN到SSM

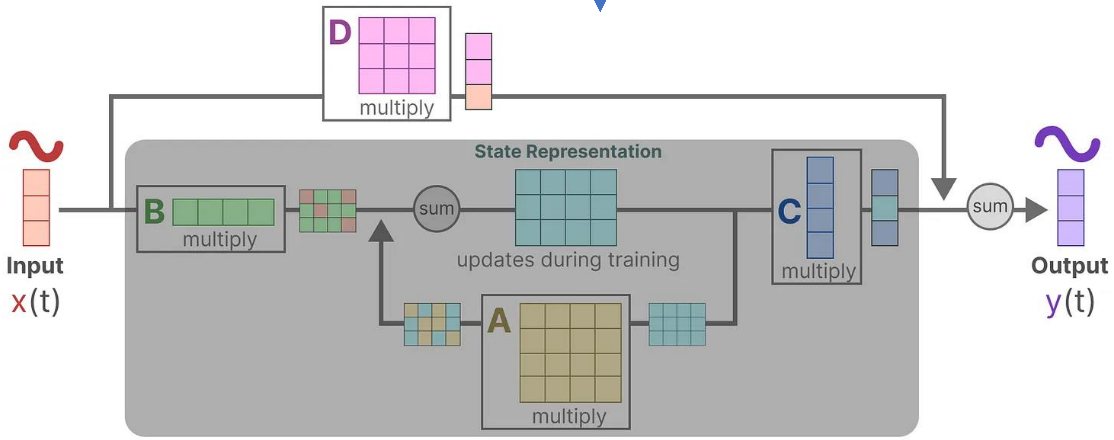


State equation $\quad \mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t)$

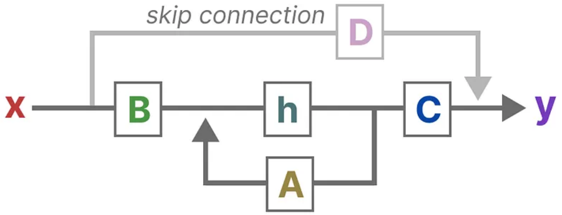
Output equation $\quad \mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t)$

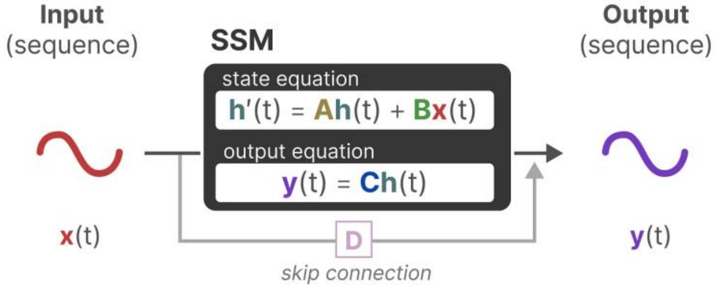
$$h_t = \tanh(W h_{t-1} + U x_t)$$





State Space Model





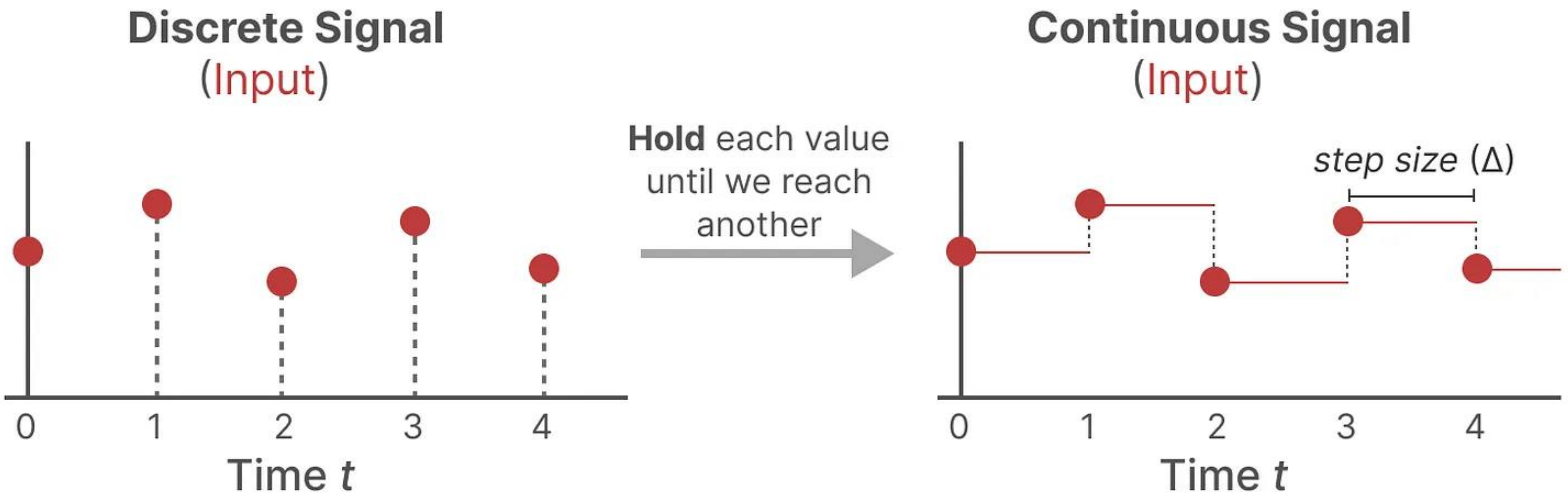
此时在SSM中，即便是在不同的输入之下，矩阵A、B、C、D都是固定不变的。

从SSM到S4

《Efficiently Modeling Long Sequences with Structured State Spaces》ICLR2022 首次提出了结构化状态空间S4

由于除了连续的输入之外，还会通常碰到离散的输入(如文本序列)，不过，就算SSM在离散数据上训练，它仍能学习到底层蕴含的连续信息，因为在SSM眼里，sequence不过是连续信号signal的采样，或者说连续的信号模型是离散的序列模型的概括

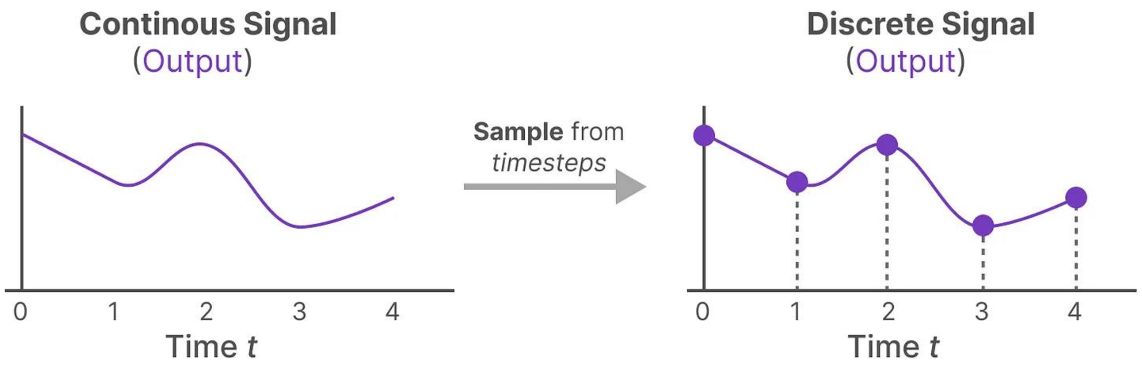
如何处理离散化数据呢？ ***零阶保持技术(Zero-order hold technique)***



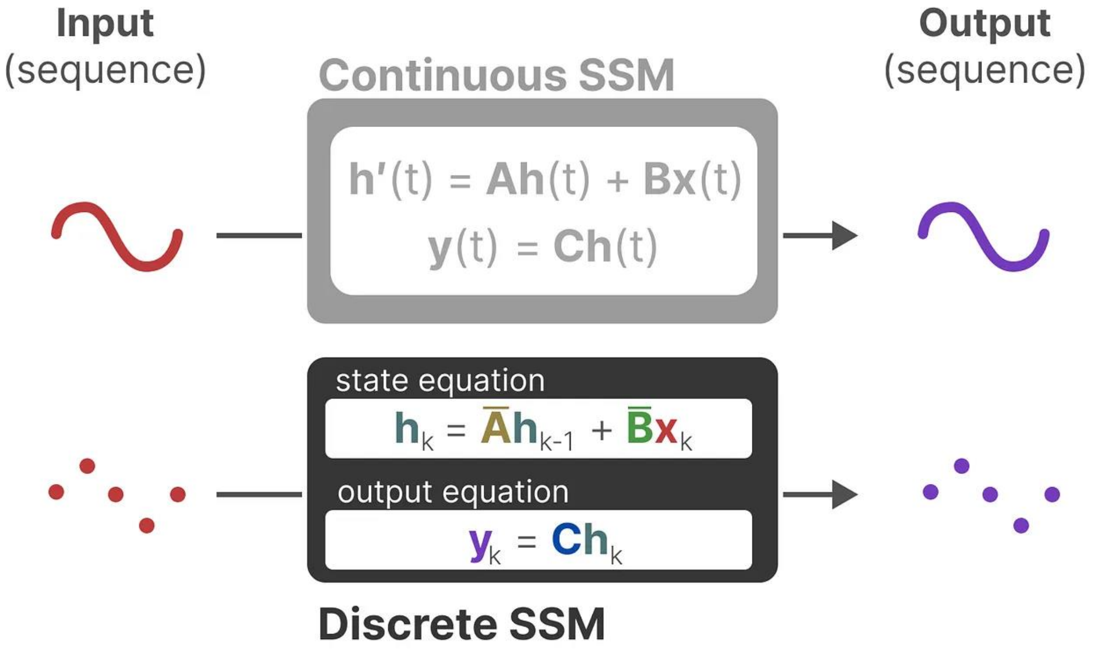
从SSM到S4

Discretized matrix **A** $$\bar{\mathbf{A}} = \exp(\Delta \mathbf{A})$$

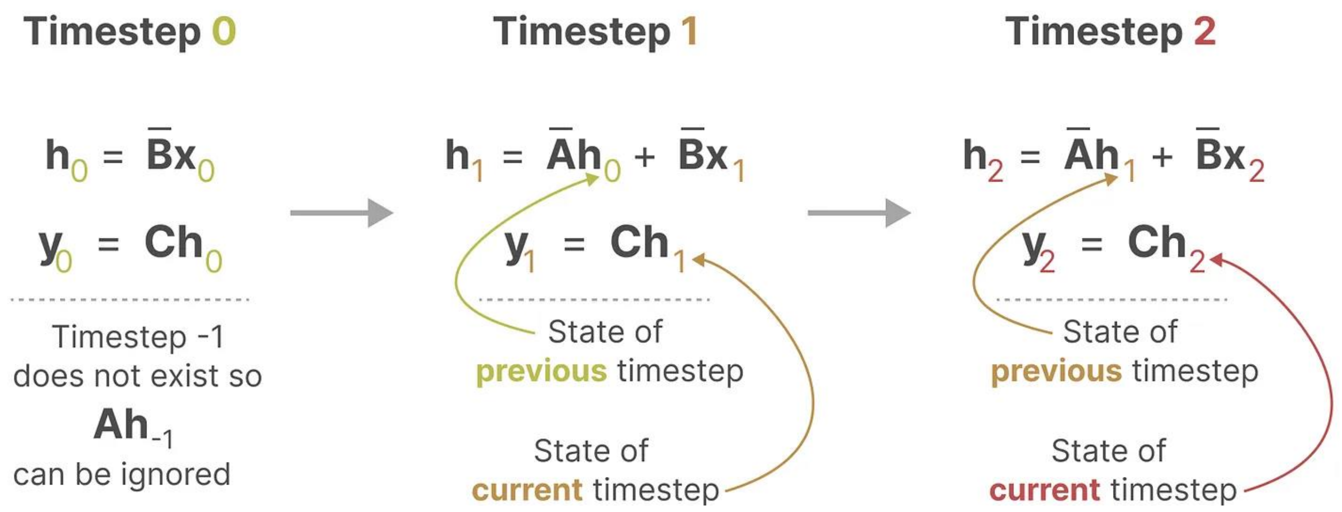
Discretized matrix **B** $$\bar{\mathbf{B}} = (\Delta \mathbf{A})^{-1}(\exp(\Delta \mathbf{A}) - I) \cdot \Delta \mathbf{B}$$



During training, the continuous representation is discretized



从SSM到S4



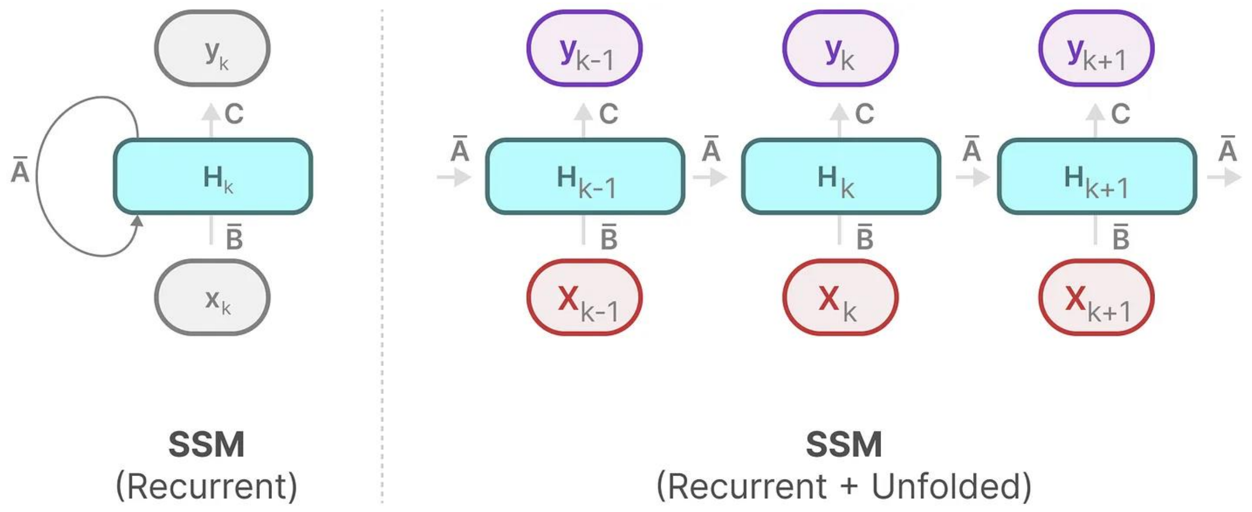
$$
\begin{aligned}
y_2 &= Ch_2 \\
&= C\left(\bar{A}h_1 + \bar{B}x_2\right) \\
&= C\left(\bar{A}\left(\bar{A}h_0 + \bar{B}x_1\right) + \bar{B}x_2\right) \\
&= C\left(\bar{A}\left(\bar{A}\cdot\bar{B}x_0 + \bar{B}x_1\right) + \bar{B}x_2\right) \\
&= C\left(\bar{A}\cdot\bar{A}\cdot\bar{B}x_0 + \bar{A}\cdot\bar{B}x_1 + \bar{B}x_2\right) \\
&= C\cdot\bar{A}^2\cdot\bar{B}x_0 + C\cdot\bar{A}\cdot\bar{B}\cdot x_1 + C\cdot\bar{B}x_2
\end{aligned}
$$

$$y_3 = \mathbf{C}\overline{\mathbf{AAAB}}x_0 + \mathbf{C}\overline{\mathbf{AAB}}x_1 + \mathbf{C}\overline{\mathbf{AB}}x_2 + \mathbf{C}\overline{\mathbf{B}}x_3$$

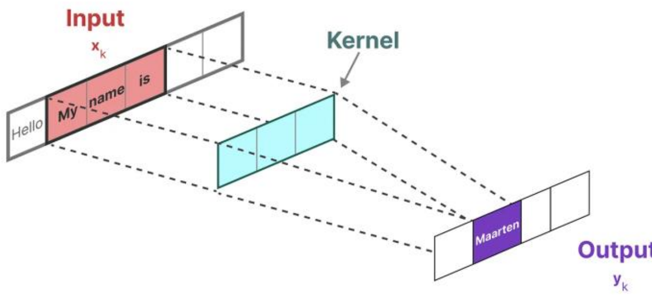
$$y_3 = \begin{pmatrix} \mathbf{C}\overline{\mathbf{AAAB}} & \mathbf{C}\overline{\mathbf{AAB}} & \mathbf{C}\overline{\mathbf{AB}} & \mathbf{C}\overline{\mathbf{B}} \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$$\overline{\mathbf{K}} = \begin{pmatrix} \mathbf{C}\overline{\mathbf{B}} & \mathbf{C}\overline{\mathbf{AB}} & \cdots & \mathbf{C}\overline{\mathbf{A}}^{k}\overline{\mathbf{B}} \end{pmatrix}$$

$$y = \overline{\mathbf{K}} * x$$



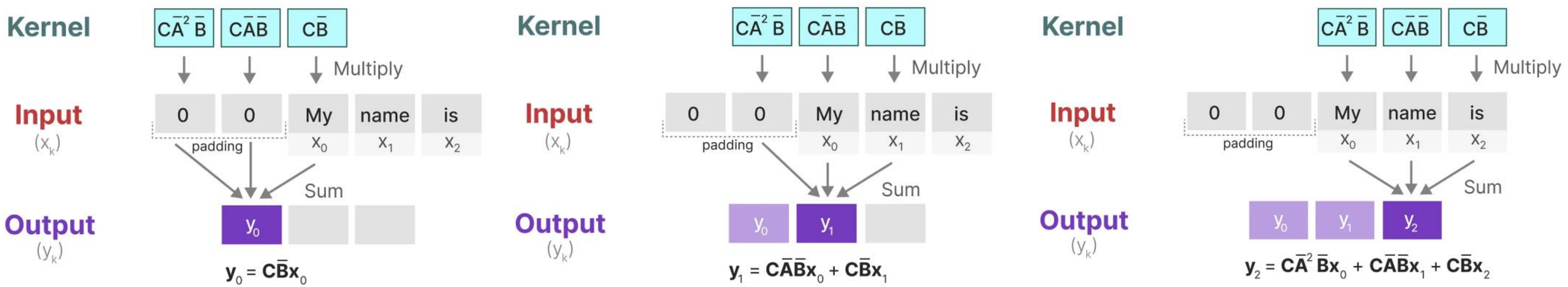
从SSM到S4

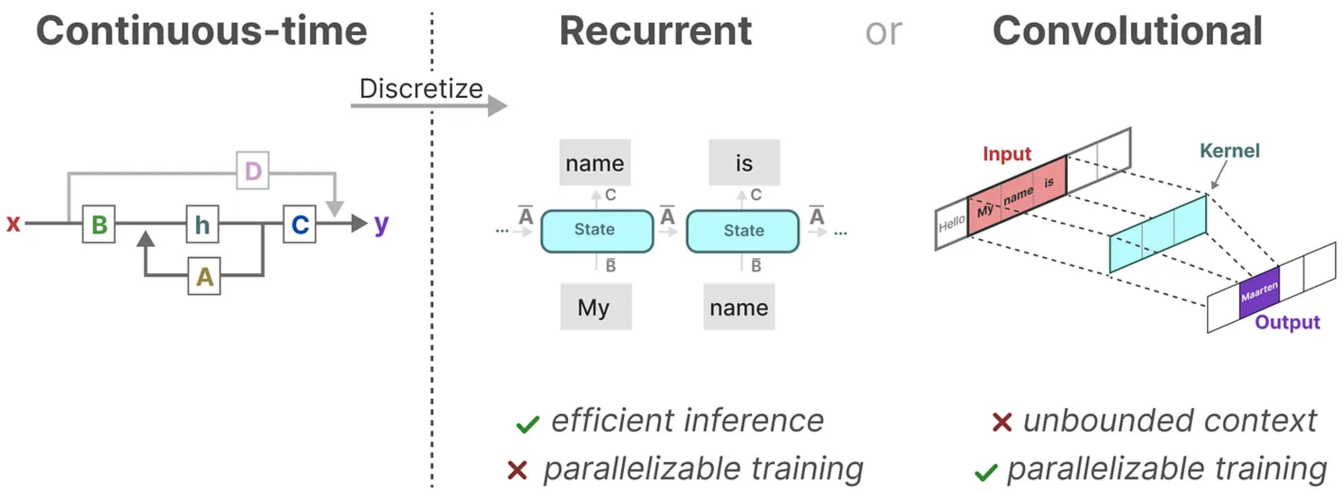


kernel → $\bar{K} = (C\bar{B}, C\bar{A}\bar{B}, \ldots, C\bar{A}^{k}\bar{B}, \ldots)$

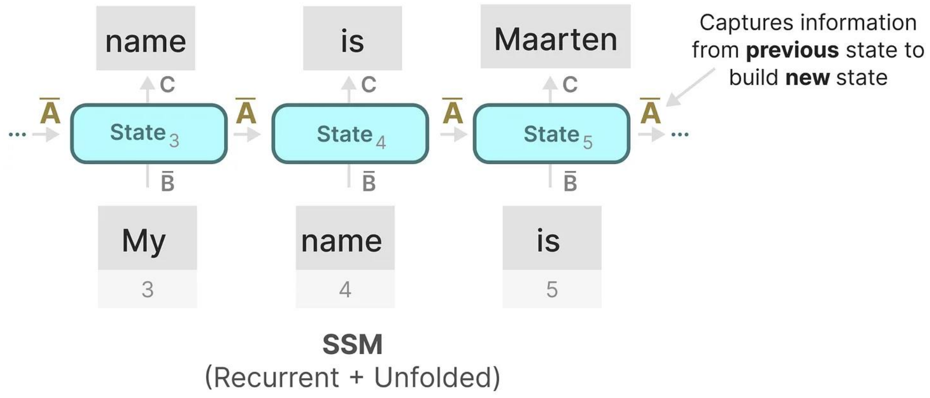
$$y = x * \bar{K}$$

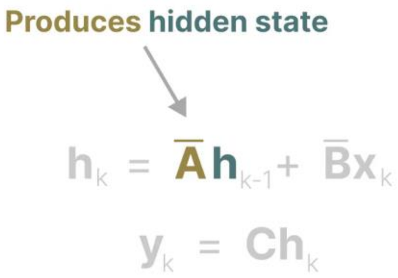
output input kernel





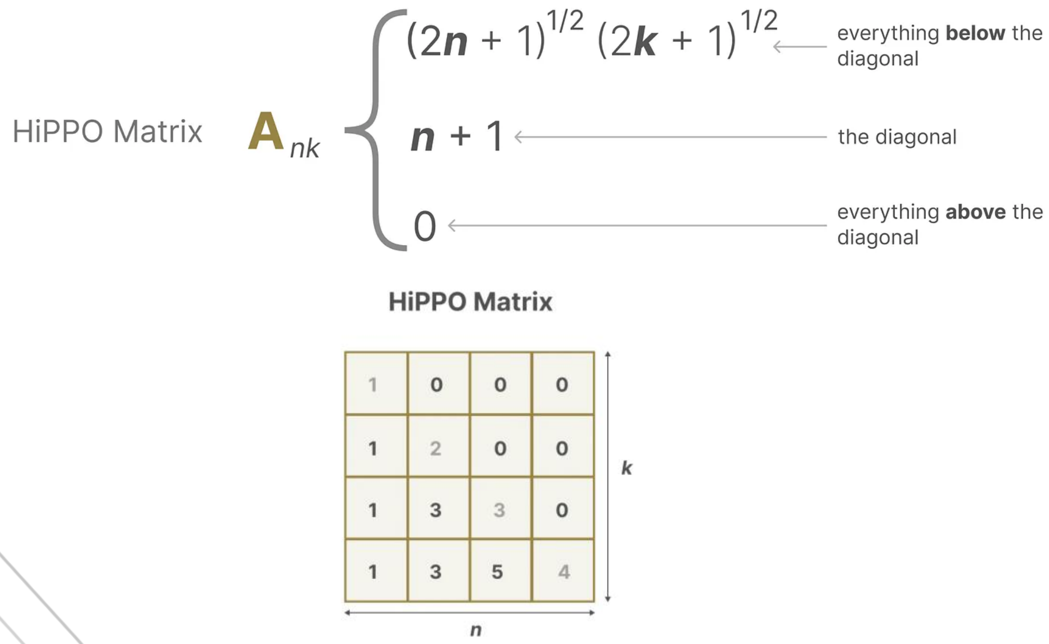
从SSM到S4

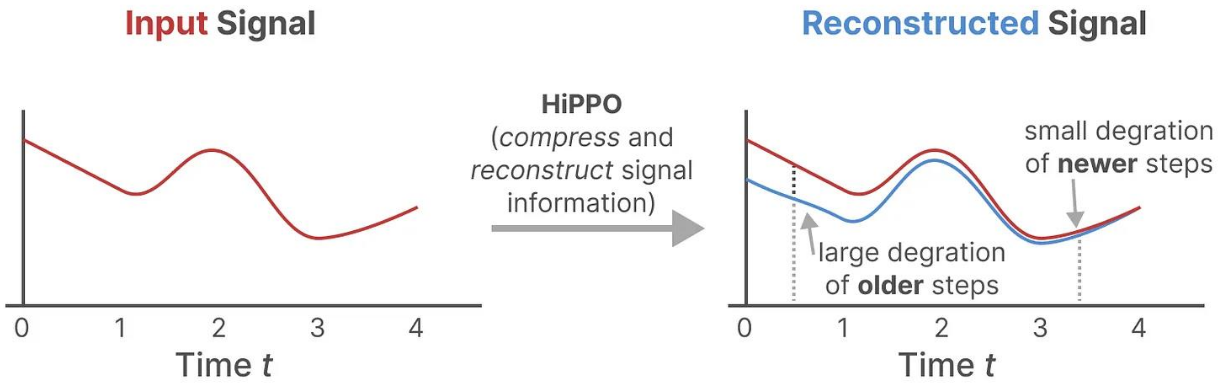




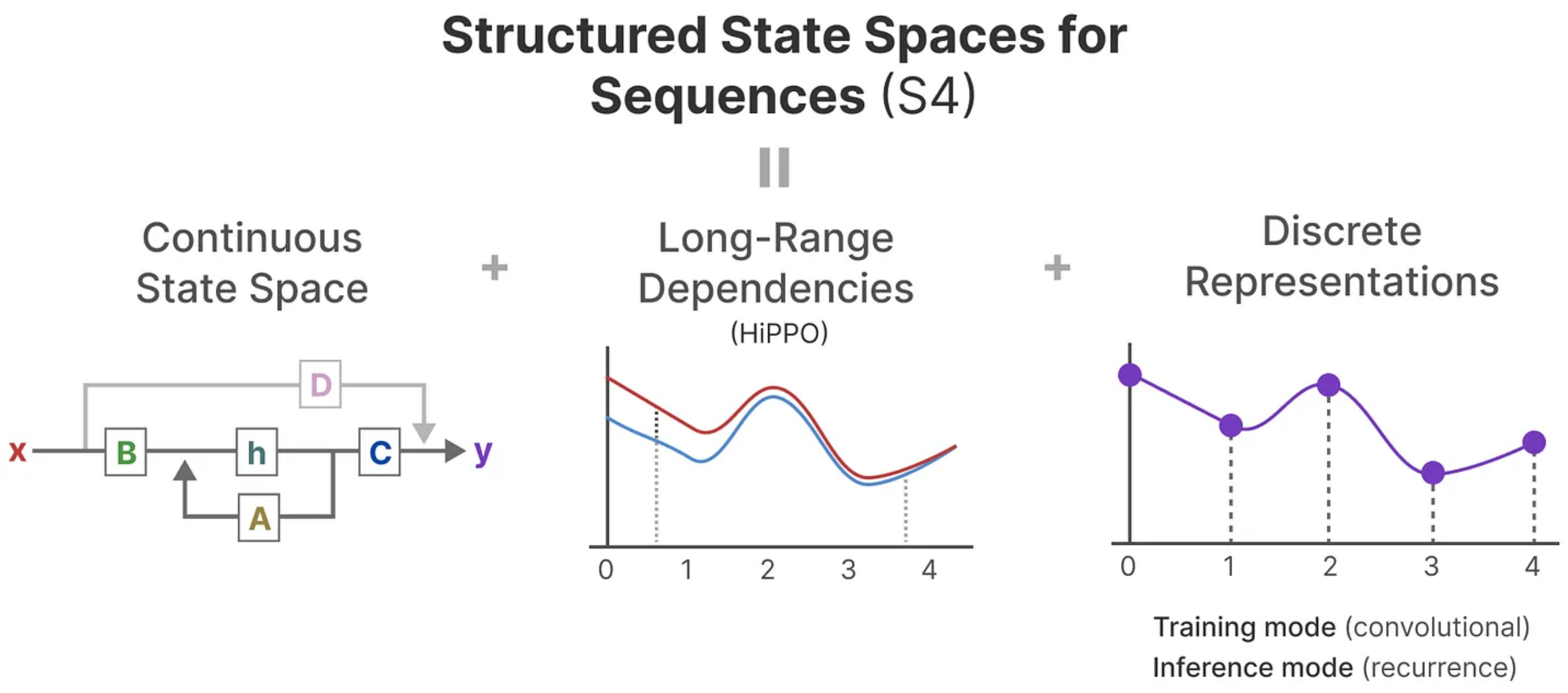
Hippo(Hippo的全称是High-order Polynomial Projection Operator)，解决如何在有限的存储空间中有效地解决序列建模的长距离依赖问题

《HiPPO: Recurrent Memory with Optimal Polynomial Projections》 NIPS2020





从SSM到S4



Theorem 1. *All HiPPO matrices from [16] have a NPLR representation*

$$\boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^* - \boldsymbol{P}\boldsymbol{Q}^\top = \boldsymbol{V}\left(\boldsymbol{\Lambda} - \left(\boldsymbol{V}^*\boldsymbol{P}\right)\left(\boldsymbol{V}^*\boldsymbol{Q}\right)^*\right)\boldsymbol{V}^* \qquad (6)$$

降维!

for unitary $\boldsymbol{V} \in \mathbb{C}^{N\times N}$, diagonal $\boldsymbol{\Lambda}$, and low-rank factorization $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{N\times r}$. These matrices HiPPO- LegS, LegT, LagT all satisfy $r = 1$ or $r = 2$. In particular, equation (2) is NPLR with $r = 1$.

从SSM到S4

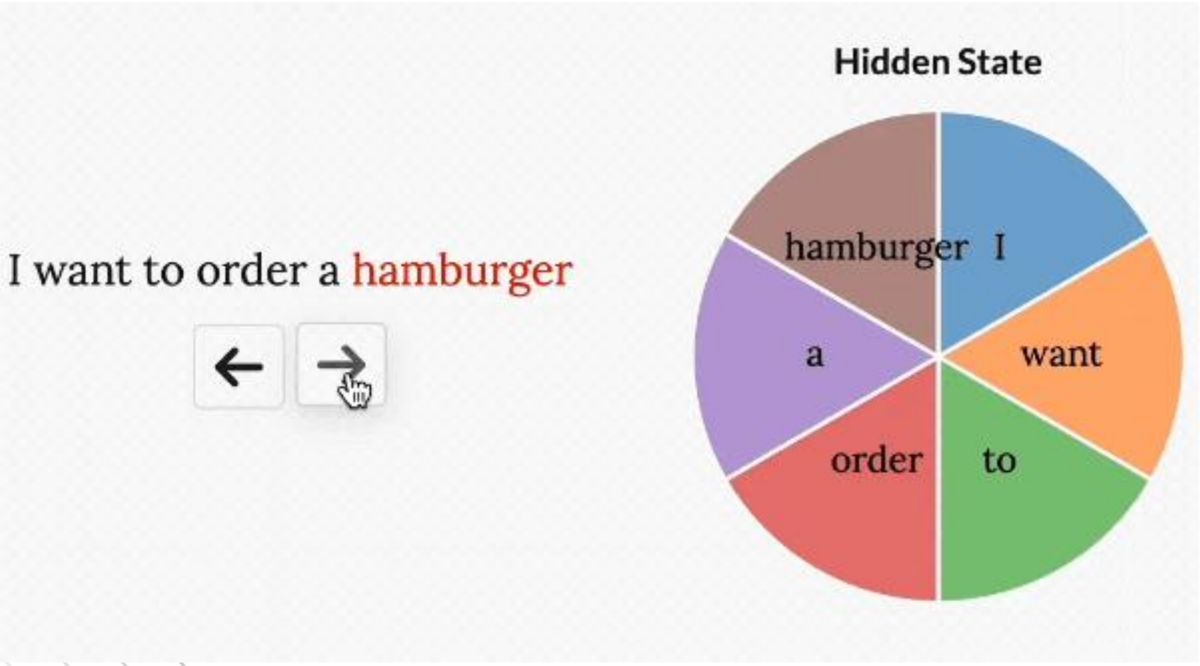
SSM的问题：矩阵不随输入不同而变化，无法针对输入做针对性推理

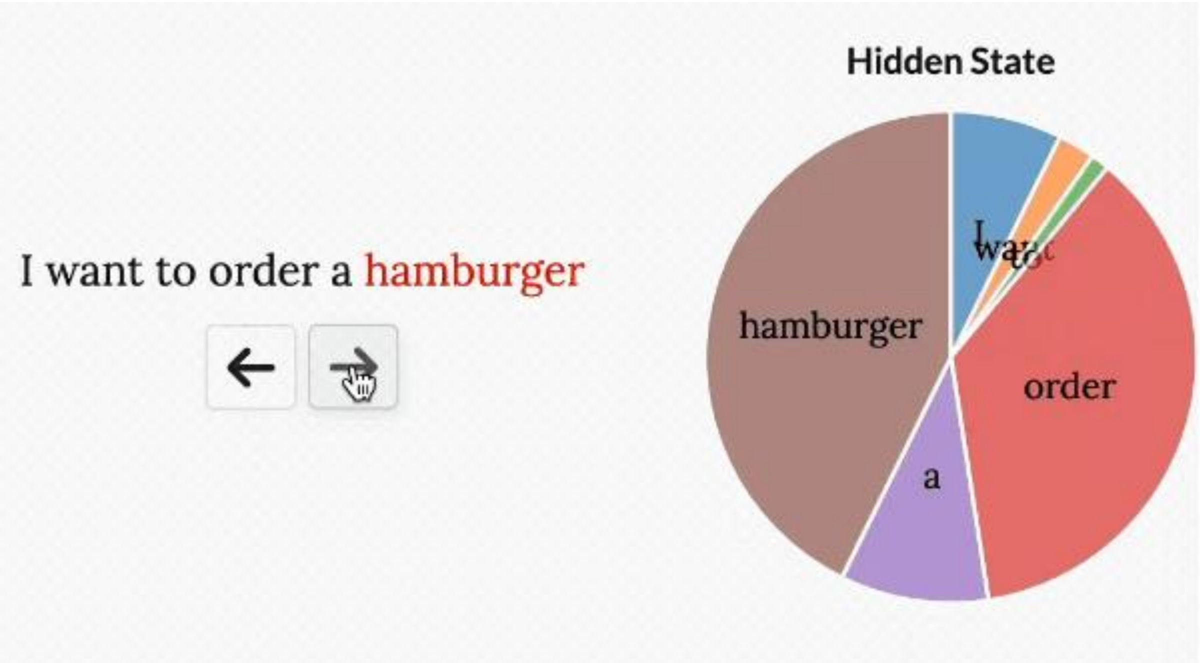
Constant regardless of the input

$$h_k = \bar{A}h_{k-1} + \bar{B}x_k$$

$$y_k = Ch_k$$

Linear Time Invariance (LTI).





从SSM到S4

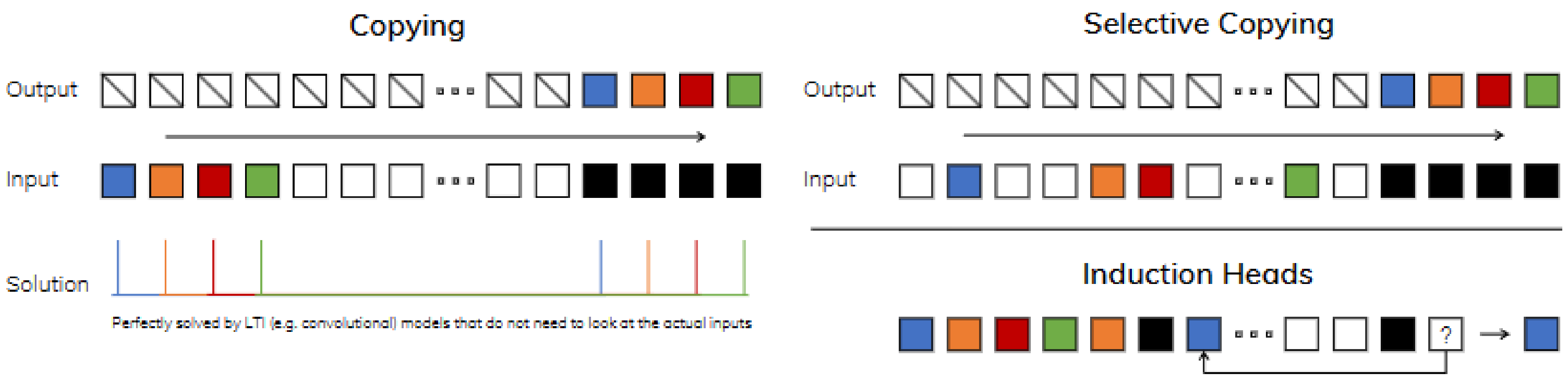
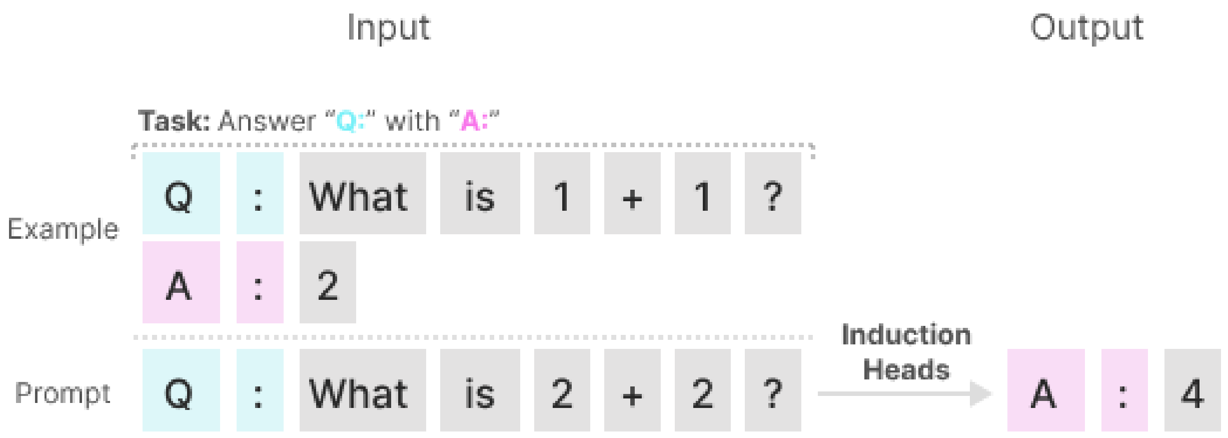


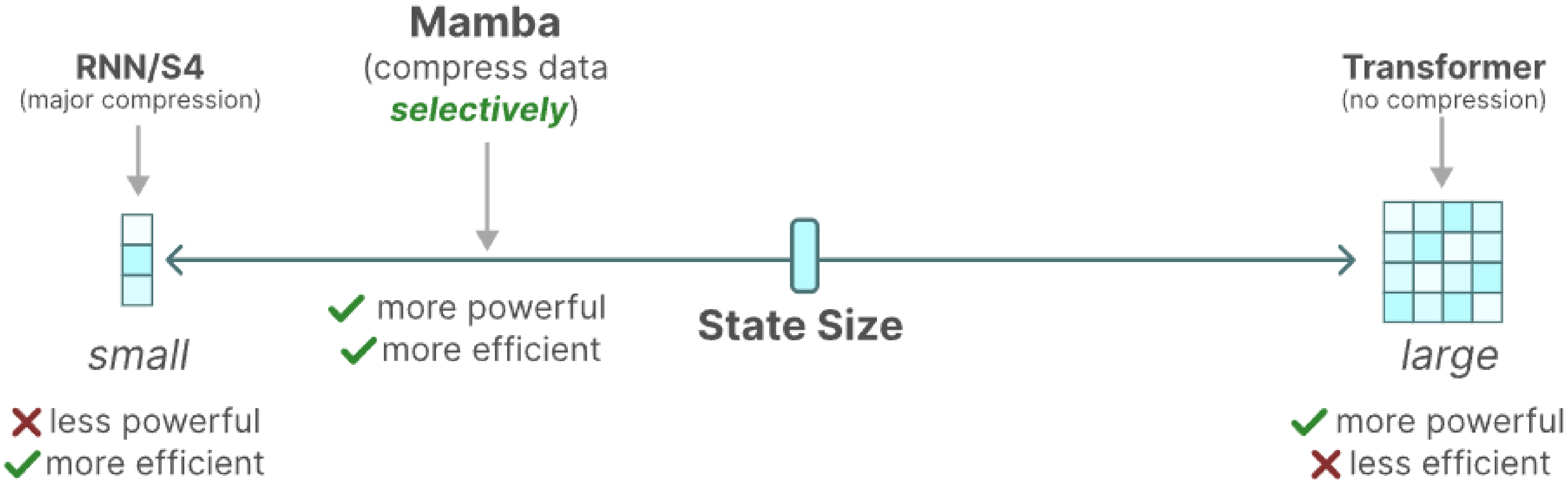
Figure 2: (*Left*) The standard version of the Copying task involves constant spacing between input and output elements and is easily solved by time-invariant models such as linear recurrences and global convolutions. (*Right Top*) The Selective Copying task has random spacing in between inputs and requires time-varying models that can *selectively* remember or ignore inputs depending on their content. (*Right Bottom*) The Induction Heads task is an example of associative recall that requires retrieving an answer based on context, a key ability for LLMs.



Mamba — A Selective SSM

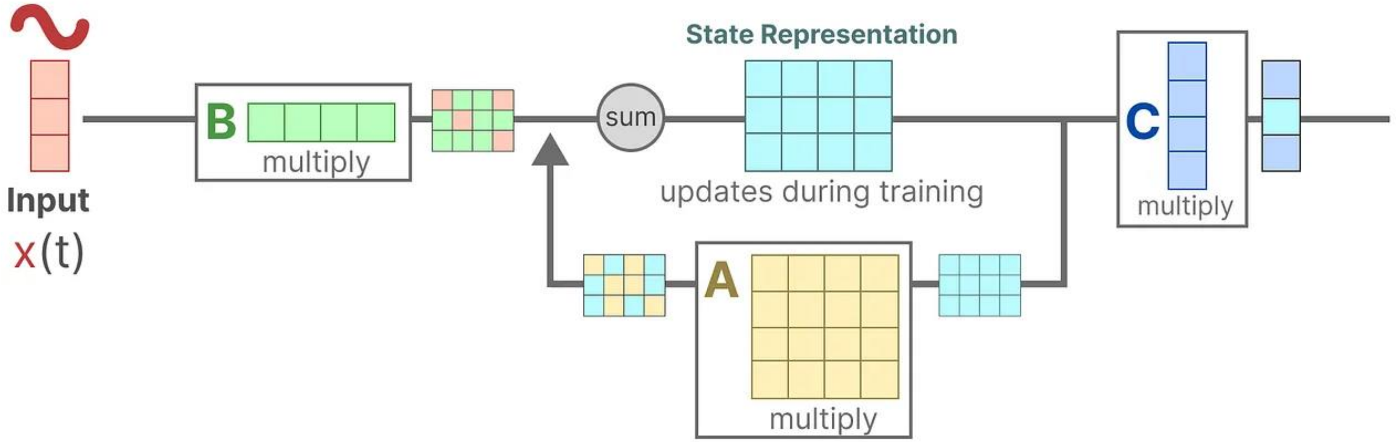
1.A **selective scan algorithm**, which allows the model to filter (ir)relevant information

2.A **hardware-aware algorithm** that allows for efficient storage of (intermediate) results through *parallel scan*, *kernel fusion*, and *recomputation*.

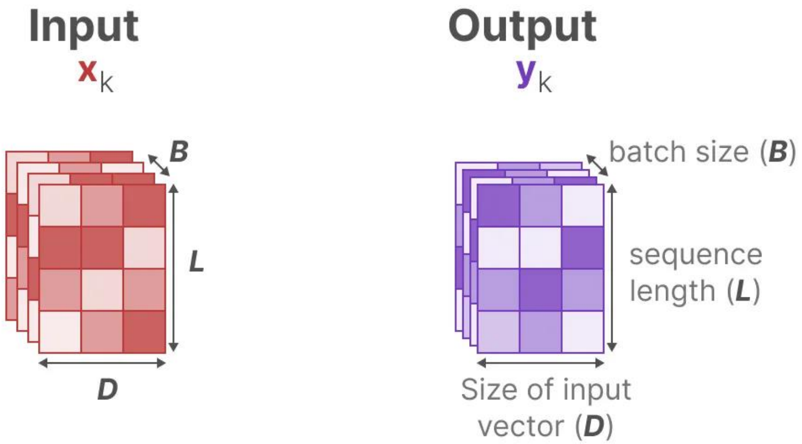


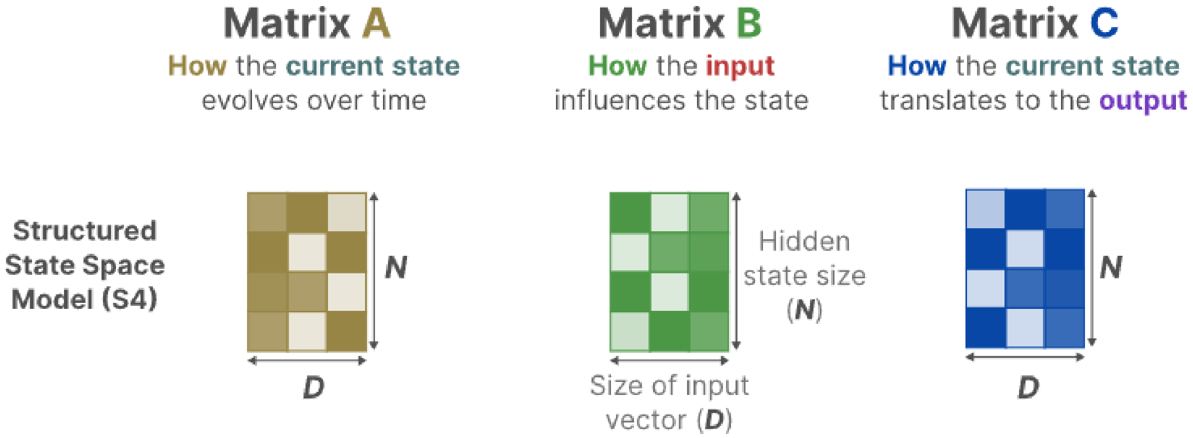
Mamba — A Selective SSM

A **selective scan algorithm**



$$A \in \mathbb{R}^{N\times N}$$
$$B \in \mathbb{R}^{N\times 1}$$
$$C \in \mathbb{R}^{1\times N}$$

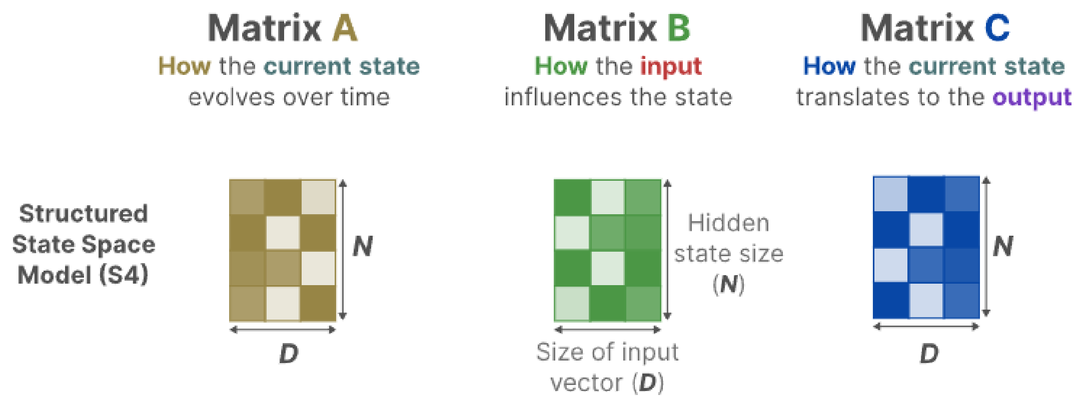


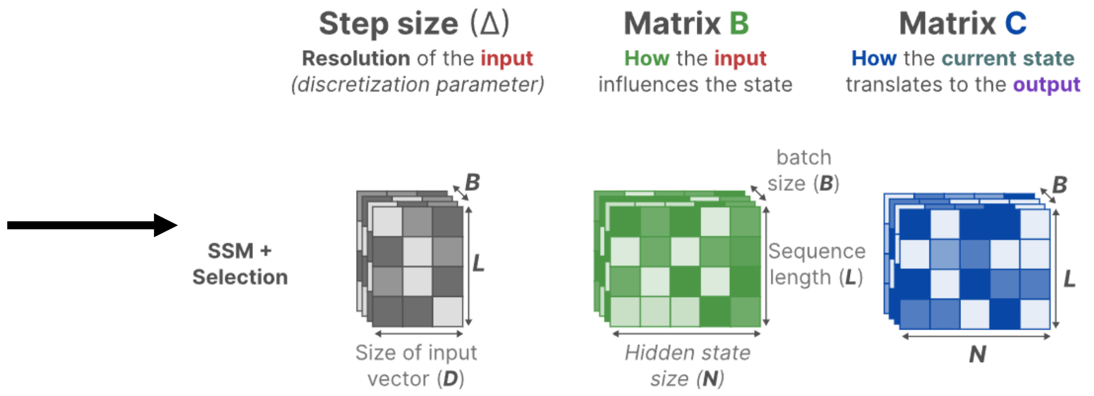


To operate over an input sequence x of batch size B and length L with D channels, the SSM is applied independently to each channel

Mamba — A Selective SSM

A selective scan algorithm





对于每个输入token，现在都有独特不同的Δ、B矩阵、C矩阵

Algorithm 1 SSM (S4)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$

1: $\boldsymbol{A} : (\mathtt{D}, \mathtt{N}) \leftarrow \mathsf{Parameter}$
$\triangleright$ Represents structured $N \times N$ matrix
2: $\boldsymbol{B} : (\mathtt{D}, \mathtt{N}) \leftarrow \mathsf{Parameter}$
3: $\boldsymbol{C} : (\mathtt{D}, \mathtt{N}) \leftarrow \mathsf{Parameter}$
4: $\Delta : (\mathtt{D}) \leftarrow \tau_\Delta(\mathsf{Parameter})$
5: $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} : (\mathtt{D}, \mathtt{N}) \leftarrow \mathsf{discretize}(\Delta, \boldsymbol{A}, \boldsymbol{B})$
6: $y \leftarrow \mathsf{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C})(x)$
$\triangleright$ Time-invariant: recurrence or convolution
7: **return** y

Algorithm 2 SSM + Selection (S6)

Input: $x : (\mathtt{B}, \mathtt{L}, \mathtt{D})$
Output: $y : (\mathtt{B}, \mathtt{L}, \mathtt{D})$

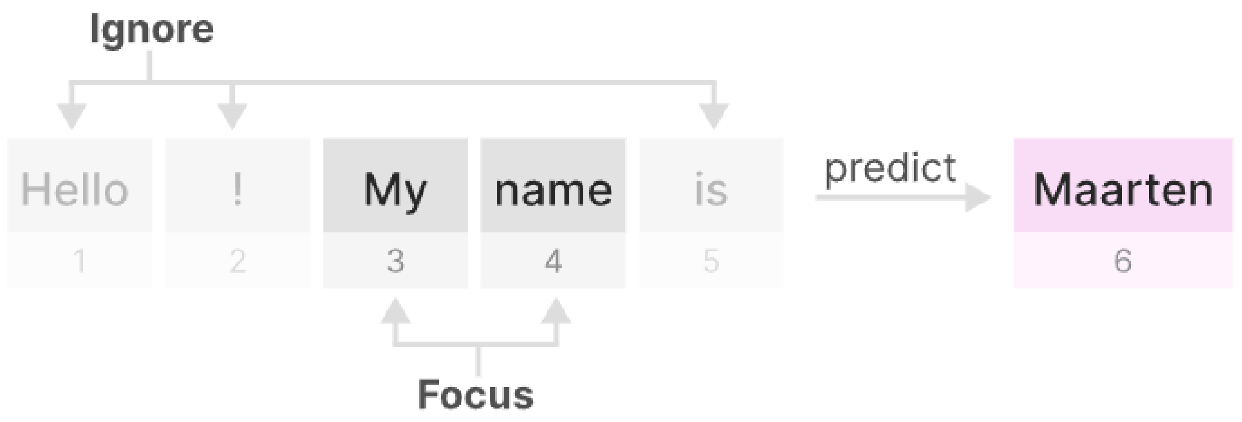
1: $\boldsymbol{A} : (\mathtt{D}, \mathtt{N}) \leftarrow \mathsf{Parameter}$
$\triangleright$ Represents structured $N \times N$ matrix
2: $\boldsymbol{B} : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_B(x)$
3: $\boldsymbol{C} : (\mathtt{B}, \mathtt{L}, \mathtt{N}) \leftarrow s_C(x)$
4: $\Delta : (\mathtt{B}, \mathtt{L}, \mathtt{D}) \leftarrow \tau_\Delta(\mathsf{Parameter} + s_\Delta(x))$
5: $\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}} : (\mathtt{B}, \mathtt{L}, \mathtt{D}, \mathtt{N}) \leftarrow \mathsf{discretize}(\Delta, \boldsymbol{A}, \boldsymbol{B})$
6: $y \leftarrow \mathsf{SSM}(\overline{\boldsymbol{A}}, \overline{\boldsymbol{B}}, \boldsymbol{C})(x)$
$\triangleright$ Time-varying: recurrence (*scan*) only
7: **return** y

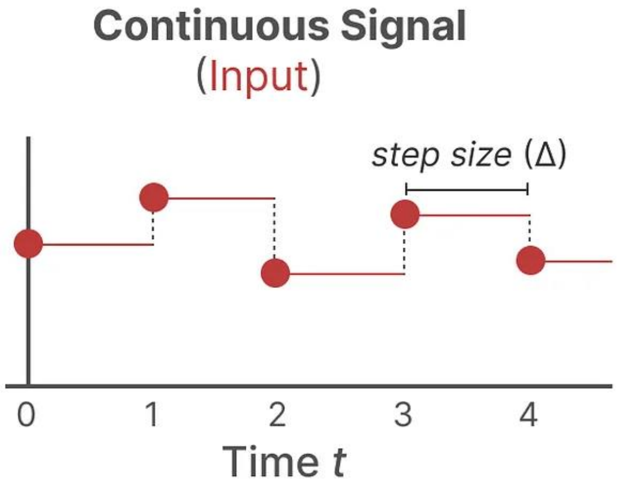
A **selective scan algorithm**

"In general, Δ controls the balance between how much to focus or ignore the current input x_t. It is analogous to the role of the gate g_t in Theorem 1, mechanically, a large Δ resets(重置) the state h and focuses on the current input x, while a small Δ persists(保持) the state and ignores the current input. "

较小的步长Δ会忽略特定单词，而更多地使用先前的上文，而较大的步长Δ会更多地关注输入单词而不是上文

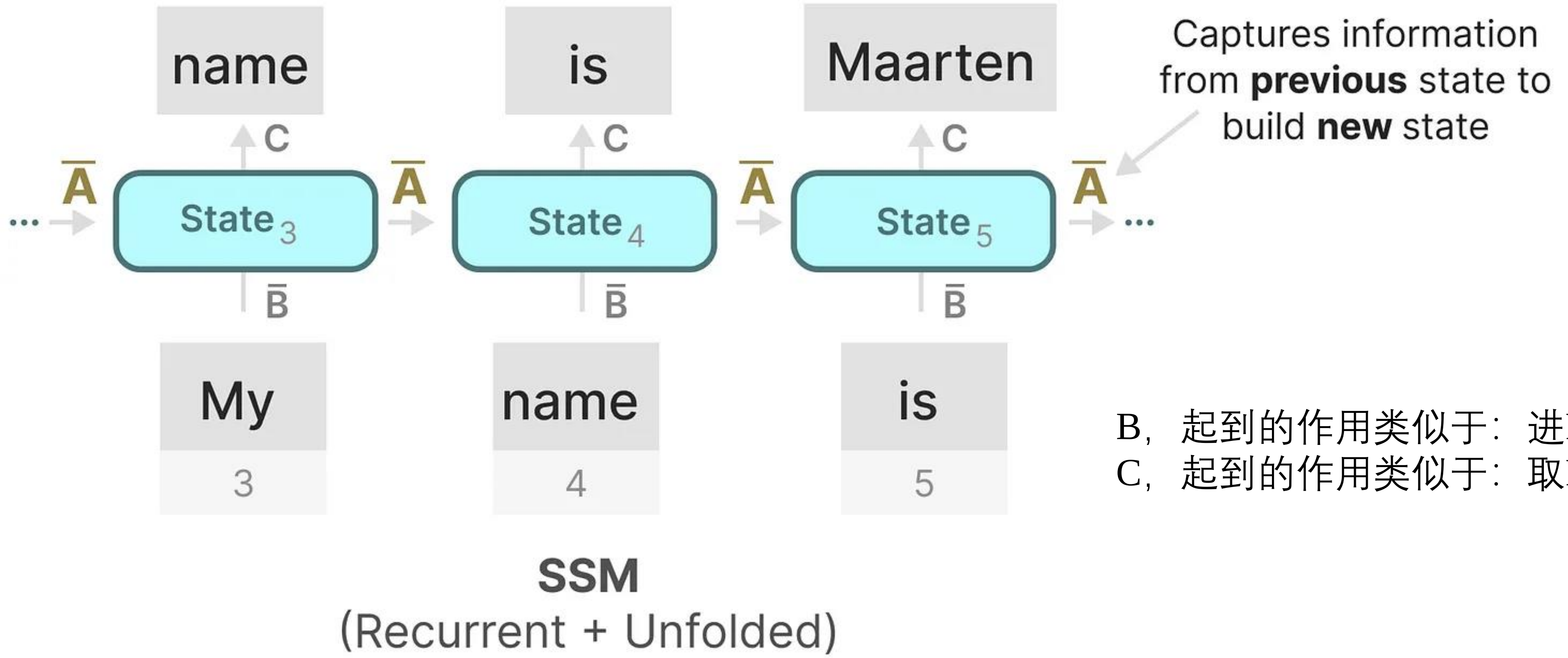
- 如果某个输入比较重要 它的步长就更长些，被重点关注
- 如果某个输入不太重要 它的步长就短，被直接忽略
- 从而对于不同的输入，达到选择性关注或忽略的目标，做到详略得当 主次分明





Mamba — A Selective SSM

A **selective scan algorithm**

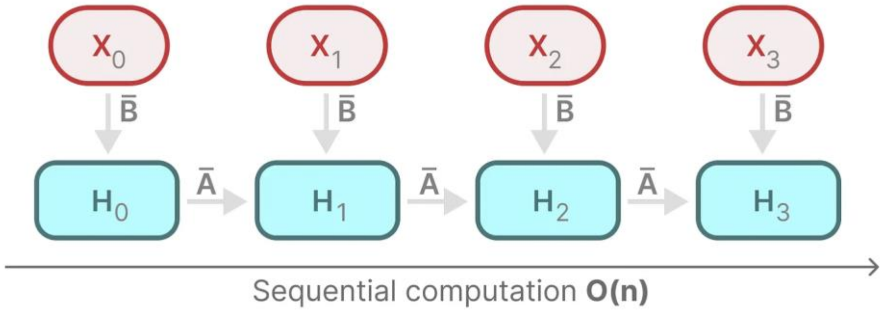


B，起到的作用类似于：进RNN的memory
C，起到的作用类似于：取RNN的memory

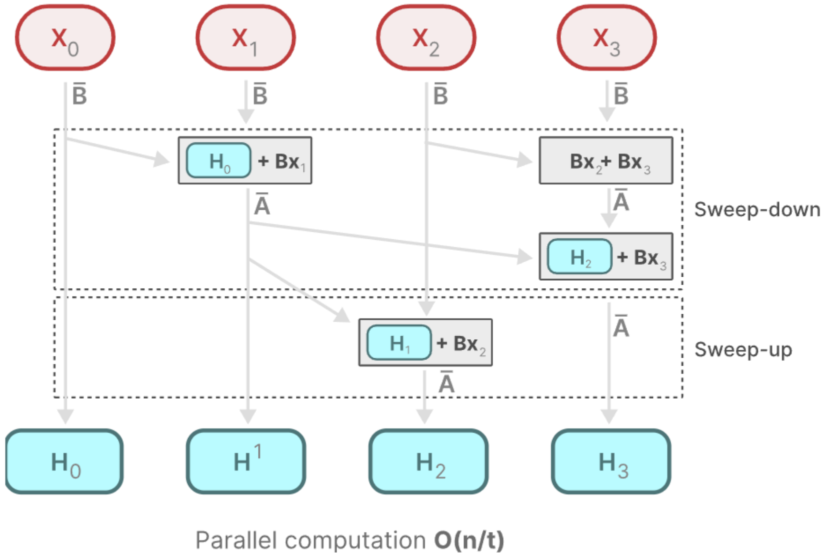
Mamba — A Selective SSM

A **hardware-aware algorithm**

- **并行化——选择性扫描算法**
- **利用SSM本身显存占用小的优势，争取模型和运算过程全部放在SRAM完成**



Mamba, however, makes this possible through the *[parallel scan](https://developer.nvidia.com/gpugems/gpugems3/part-vi-gpu-computing/chapter-39-parallel-prefix-sum-scan-cuda)* algorithm.



输入：$[x_1, x_2, x_3, x_4]$

输出：$[h_1 = x_1, h_2 = x_1 + x_2, h_3 = x_1 + x_2 + x_3, h_4 = x_1 + x_2 + x_3 + x_4]$

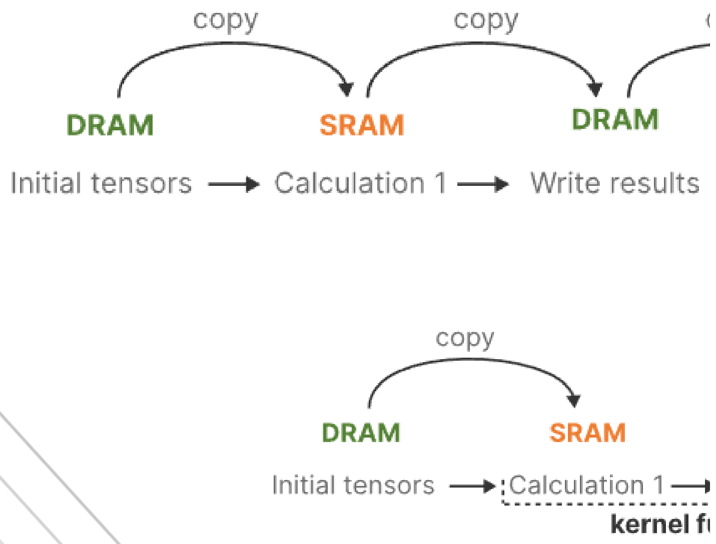
$z_1 = x_1 + x_2 (h_2) \qquad z_2 = x_3 + x_4$

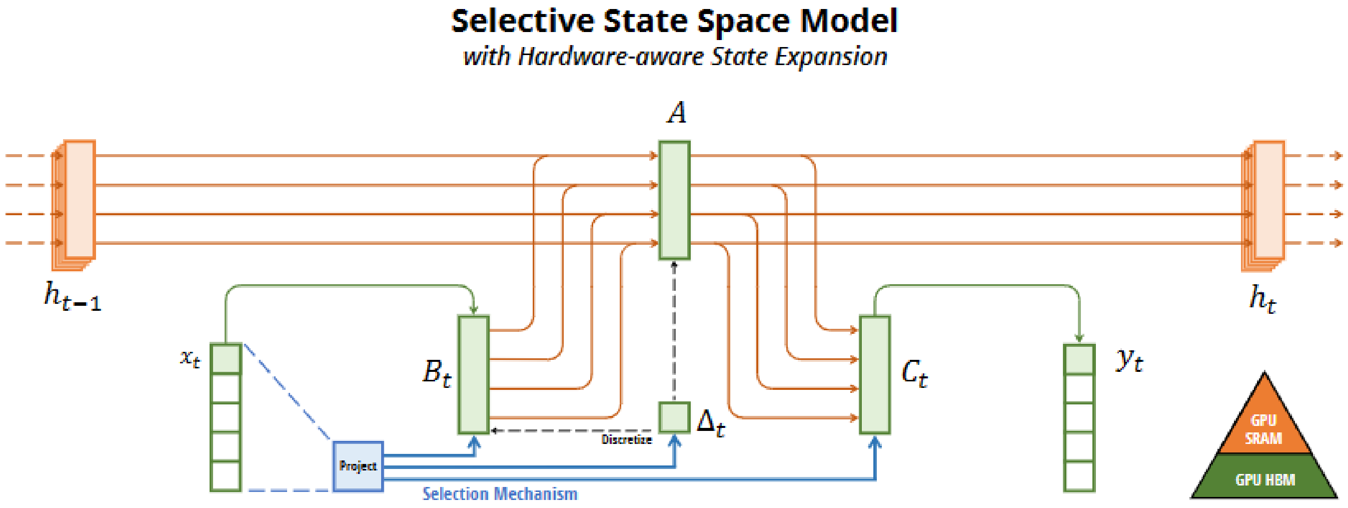
$z_1 + x_3 = x_1 + x_2 + x_3 (h_3) \qquad z_1 + z_2 = x_1 + x_2 + x_3 + x_4 (h_4)$

Mamba — A Selective SSM

A hardware-aware algorithm

- **利用SSM本身显存占用小的优势，争取模型和运算过程全部放在SRAM完成**
  - HBM：显卡的高带宽内存，提供了比传统的GDDR更高的带宽，更低的功耗。当然，相比于SRAM，HBM仍是“低速大容量”的
  - SRAM：显卡的高速缓存区，读取速度非常快
  - Transformer仅注意力层可能就需要把模型各个模块分批次从HBM加载到SRAM去计算，一个模块算完了就从SRAM取出来，再加载下一个模块如，先算QKV，再算注意力分数，注意力分数再与输入相乘
  - SSM的参数（原始的A，B，C，Δ会被直接加载到SRAM，在SRAM里计算$\bar{A}$，$\bar{B}$及后续操作，一步直接得到输出，从SRAM写回HBM）





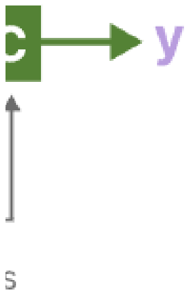
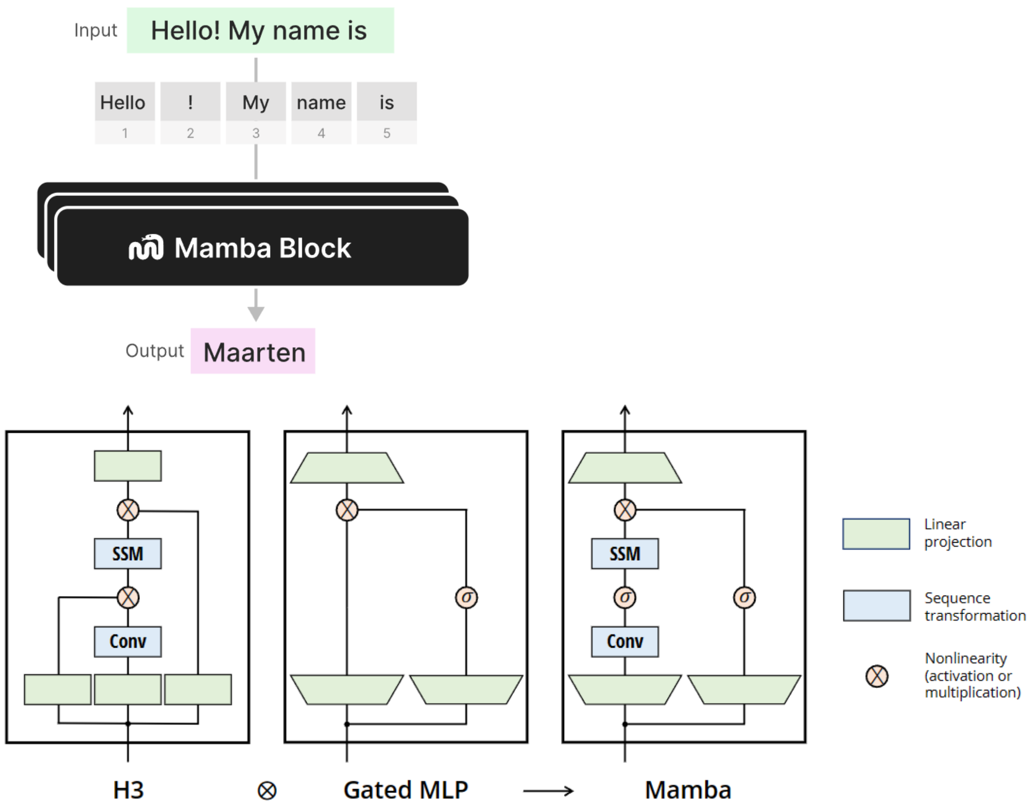
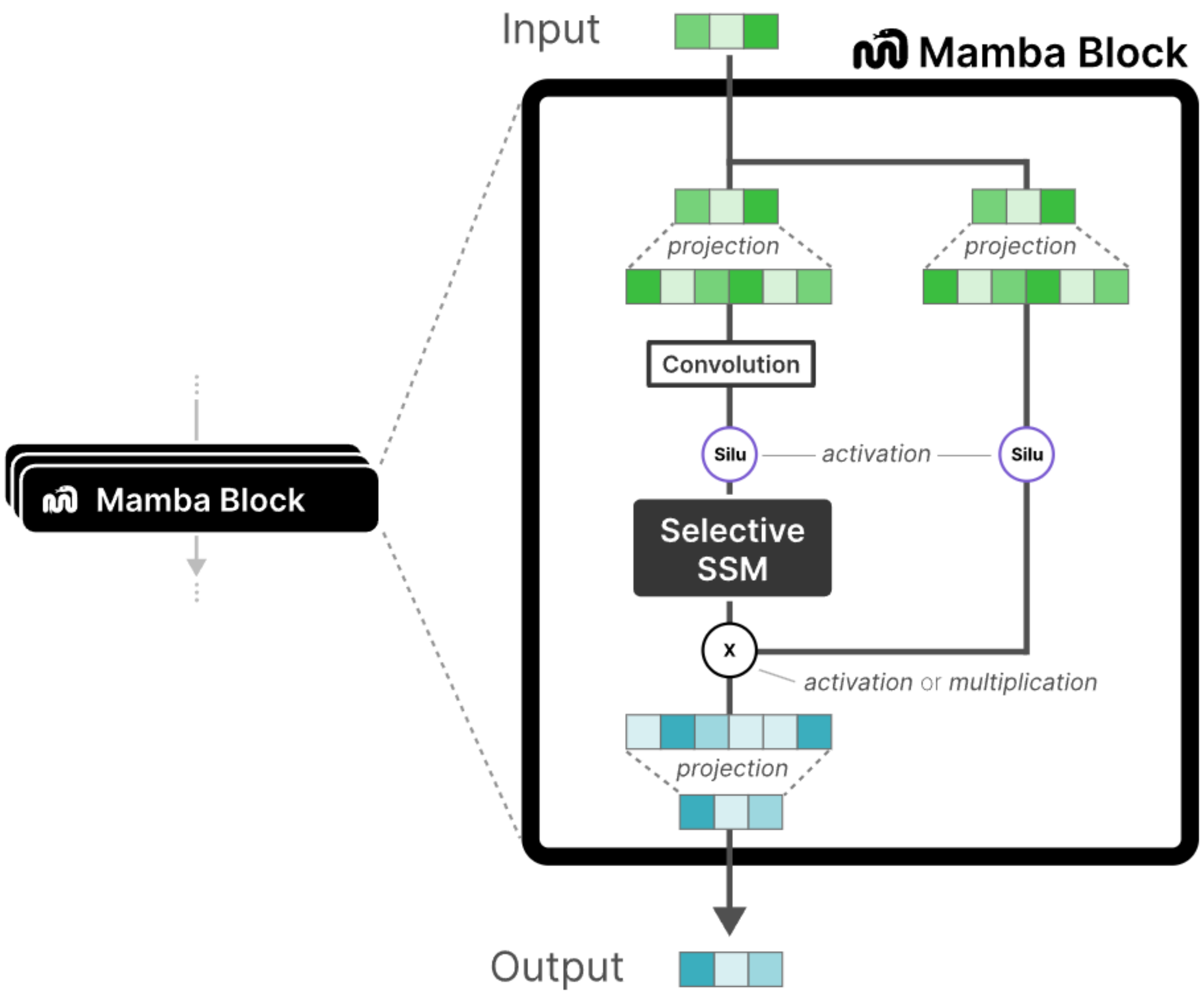


Figure 1: (**Overview**.) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the (Δ, A, B, C) parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU memory hierarchy.

Mamba — A Selective SSM

The Mamba Block





It starts with a linear projection to expand upon the input embeddings. Then, a convolution before the Selective SSM is applied to prevent independent token calculations.

Mamba — A Selective SSM

	Training	Inference
Transformers	**Fast!** (parallelizable)	**Slow...** (scales **quadratically** with sequence length)
RNNs	**Slow...** (not parallelizable)	**Fast!** (scales **linearly** with sequence length)
Mamba	**Fast!** (parallelizable)	**Fast!** (scales **linearly** with sequence length + **unbounded** context)

[GitHub - state-spaces/mamba]()

[GitHub - wzhwzhwzh0921/S-D-Mamba: Code for "Is Mamba Effective for Time Series Forecasting?"]()

西南財經大學
SOUTHWESTERN UNIVERSITY OF FINANCE AND ECONOMICS

Questions and Discussions

主讲人：阮皓
2024.04.17