

Mamba: Resurrecting RNN for long sequences

Wang Shida

Department of Mathematics

National University of Singapore

March 2, 2024

Table of Contents

1 Introduction

- Sequence modelling

2 State-space models

- What is SSM?
- Why SSM?
- Is SSM good enough?





3 Mamba

- Mamba = SSM + gating + lightspeed kernel
- Lightspeed kernel
- Speed in practice
- Memory in practice

Sequence modelling

- NLP: Machine translation, language modelling, ...
- Time series: Weather prediction, ...
- Heuristic: Image classification, stretch the image into a sequence of pixels. (sMNIST, sCIFAR10, Long range arena¹)
- Video generation: Given a text prompt, generate a video based on the prompt, Sora².

¹Yi Tay et al. "Long Range Arena: A Benchmark for Efficient Transformers". In: *CoRR* abs/2011.04006 (2020). arXiv: 2011.04006.

²<https://openai.com/research/video-generation-models-as-world-simulators>    

Sequence modelling

- NLP: Machine translation, language modelling, ...
- Time series: Weather prediction, ...
- Heuristic: Image classification, stretch the image into a sequence of pixels. (sMNIST, sCIFAR10, Long range arena¹)
- Video generation: Given a text prompt, generate a video based on the prompt, Sora².
 - A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

¹Tay et al., “Long Range Arena: A Benchmark for Efficient Transformers”.

²<https://openai.com/research/video-generation-models-as-world-simulators>



Figure: How many gpu to train sora?

For now, Sora can generate video around 1 minute. If we scale up the transformer³ naively for 1hour movie generation, then it takes 3600x GPU memory to train the model.

³William Peebles and Saining Xie. "Scalable Diffusion Models with Transformers". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 4195–4205.

Inputs $x \in \mathbb{R}^{T \times d}$,

$$q = xQ, k = xK, v = xV, \quad Q, K, V \in \mathbb{R}^{d \times d}, \quad (1)$$

$$\text{Attention}(x) = \text{softmax} \left(\frac{qk^\top}{\sqrt{d}} \right) \in \mathbb{R}^{T \times T}, \quad (2)$$

$$\text{Self-attention}(x) = \text{softmax} \left(\frac{qk^\top}{\sqrt{d}} \right) v. \quad (3)$$

Today, we are focusing on the model design as we want to have a backbone that has better (slower) growth w.r.t. sequence length T . By better we mean **subquadratic**.

Table of Contents

1 Introduction

- Sequence modelling

2 State-space models

- What is SSM?
- Why SSM?
- Is SSM good enough?

3 Mamba

- Mamba = SSM + gating + lightspeed kernel
- Lightspeed kernel
- Speed in practice
- Memory in practice

Single-layer nonlinear recurrent neural networks:

$$\frac{dh_t}{dt} = \sigma(Wh_t + Ux_t), \quad h_{-\infty} = 0 \in \mathbb{R}^m, \quad (4)$$

$$y_t = Ch_t. \quad (5)$$

Single-layer state-space models:

$$\frac{dh_t}{dt} = \Lambda h_t + Ux_t, \quad h_{-\infty} = 0 \in \mathbb{R}^m, \quad (6)$$

$$y_t = \sigma(Ch_t). \quad (7)$$

Trainable parameters: $\Lambda \in \mathbb{R}^{m \times m}$ is a diagonal matrix,
 $U \in \mathbb{R}^{m \times d}$, $C \in \mathbb{R}^{d \times m}$, d is the input dimension while m is the hidden dimension.

	Train latency	Train FLOPs ⁴	Inference FLOPs
RNN	$O(T)$	$O(T)$	$O(1)$
Attn	$O(\log T)$	$O(T^2)$	$O(T)$
SSM	$O(\log T)$	$O(T \log T)$ ⁵	$O(1)$

Table: Comparison of single-layer nonlinear RNN, self-attention and SSM. We focus on the case with training sequence length T and one-step inference of context length T . We consider the idealized setting that we have infinite GPU memory and GPU cores. The bottleneck is marked in red.

⁴FLOPs is the number of floating points calculation in each step.

⁵Work-efficient version costs $O(T)$: https://en.wikipedia.org/wiki/Prefix_sum

- ① Universality:⁶⁷ Any targets that can be learned by RNNs or transformers can also be learned by SSMs.

What is not implied from universality: efficiency.

⁶Shida Wang and Beichen Xue. “State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.

⁷Antonio Orvieto et al. “On the Universality of Linear Recurrences Followed by Nonlinear Projections”. In: (2023). [arXiv: 2307.11888](#).

- ① Universality⁸⁹: Any targets that can be learned by RNNs or transformers can also be learned by SSMs.
- ② Curse of memory¹⁰¹¹¹²: Recurrent models having difficulty learning long-term relationships.

⁸Wang and Xue, “State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory”.

⁹Orvieto et al., “On the Universality of Linear Recurrences Followed by Nonlinear Projections”.

¹⁰Zhong Li et al. “On the Curse of Memory in Recurrent Neural Networks: Approximation and Optimization Analysis”. In: *International Conference on Learning Representations*. Oct. 2020.

¹¹Haotian Jiang et al. “A Brief Survey on the Approximation Theory for Sequence Modelling”. In: *Journal of Machine Learning* 2.1 (June 2023), pp. 1–30. ISSN: 2790-203X, 2790-2048. DOI: 10.4208/jml.221221.

¹²Shida Wang, Zhong Li, and Qianxiao Li. “Inverse Approximation Theory for Nonlinear Recurrent Neural Networks”. In: *The Twelfth International Conference on Learning Representations*. Oct. 2023.

Based on the piece-wise inputs:

$$\mathbf{u}_t^x = \begin{cases} x & t \geq 0, \\ 0 & t < 0. \end{cases} \quad (8)$$

We define the memory function of model:

$$\mathcal{M}(t) := \sup_x \frac{|\frac{dy_t}{dt}(\mathbf{u}^x)|}{|x| + 1}. \quad (9)$$

This memory function is a surrogate to study how the model memorizes the impulse in inputs.

Theoretical results (informal)

If a linear/nonlinear functional is learned by a sequence of linear/nonlinear RNN models, with weights boundedness, memory function of the target functional is decaying exponentially^{ab}: For some $\beta > 0$,

$$\lim_{t \rightarrow \infty} e^{\beta t} \mathcal{M}(t) = 0. \quad (10)$$

^aLi et al., “On the Curse of Memory in Recurrent Neural Networks”.

^bWang, Li, and Li, “Inverse Approximation Theory for Nonlinear Recurrent Neural Networks”.

- ① Universality:
- ② Curse of memory¹³:
 - SSMs without reparameterization (Λ is trainable) has curse of memory.

$$\lim_{t \rightarrow \infty} e^{\beta t} \mathcal{M}(t) = 0. \quad (11)$$

¹³Shida Wang and Qianxiao Li. *StableSSM: Alleviating the Curse of Memory in State-space Models through Stable Reparameterization*. Nov. 2023. [arXiv: 2311.14495](https://arxiv.org/abs/2311.14495).

- ① Universality:
- ② Curse of memory¹³:
 - SSMs without reparameterization (Λ is trainable) has curse of memory.

$$\lim_{t \rightarrow \infty} e^{\beta t} \mathcal{M}(t) = 0. \quad (11)$$

- SSMs with stable reparameterizations such as $\Lambda = -\exp(W)$ or $-\text{softplus}(W)$, $W \in \mathbb{R}^{m \times m}$. Here W is trainable. Then StableSSMs can learn long-term memory such as $\mathcal{M}(t) = 1/t^2$.

¹³Wang and Li, *StableSSM*.

Summary of theory

- ① Why SSM? - They are faster than RNNs and Attns.
- ② Is SSM good enough? - Yes, they are universal approximators. They can overcome the difficulty of learning long-term memory.

Table of Contents

1 Introduction

- Sequence modelling

2 State-space models

- What is SSM?
- Why SSM?
- Is SSM good enough?

3 Mamba

- Mamba = SSM + gating + lightspeed kernel
- Lightspeed kernel
- Speed in practice
- Memory in practice

Mamba¹⁴ is a variant of SSMs with input-dependent gating.

Vanilla	Add gating
RNN	GRU or LSTM
Attn	Gated Attention Unit (GAU) ¹⁵
TCN	Gated Temporal Convolution Network (GTCN) ¹⁶
SSM	GSS ¹⁷ or Mamba

Table: Gating is effective in all seq2seq layers

¹⁴Albert Gu and Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. Dec. 2023. arXiv: 2312.00752.

¹⁵Weizhe Hua et al. “Transformer quality in linear time”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 9099–9117.

¹⁶Yujie Liu et al. “Time series prediction based on temporal convolutional network”. In: *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE. 2019, pp. 300–305.

¹⁷Harsh Mehta et al. *Long Range Language Modeling via Gated State Spaces*. 2022. arXiv: 2206.13947.

Vanilla SSM:

$$h_{k+1} = \Lambda h_k + Ux_k \quad (12)$$

$$= \Lambda^k Ux_0 + \cdots + Ux_k. \quad (13)$$

Input-dependent gating expands the model space:

$$h_{k+1} = \Lambda(x_k)h_k + U(x_k)x_k. \quad (14)$$

No FFT: SSMs with input-dependent gating don't have a convolution kernel with form $\rho(k) = \Lambda^k U$.

Still $O(T \log T)$: Consider following binary operator:

$$(\Lambda_1, x_1) \circ (\Lambda_2, x_2) = (\Lambda_2 \Lambda_1, x_2 + \Lambda_2 x_1). \quad (15)$$

The hidden state of Mamba after k inputs can be retrieved from

$$(\Lambda_k, h_k) = (\Lambda_0, h_0) \circ (\Lambda(x_1), U(x_1)x_1) \circ \cdots \circ (\Lambda(x_k), U(x_k)x_k) \quad (16)$$

As the above binary operator is associative:

$$\left((\Lambda_1, x_1) \circ (\Lambda_2, x_2) \right) \circ (\Lambda_3, x_3) = (\Lambda_1, x_1) \circ \left((\Lambda_2, x_2) \circ (\Lambda_3, x_3) \right). \quad (17)$$

Associative scan¹⁸ can be used to accelerate mamba.

¹⁸Eric Martin and Chris Cundy. "Parallelizing Linear Recurrent Neural Nets Over Sequence Length". In: *International Conference on Learning Representations*. Feb. 2018.

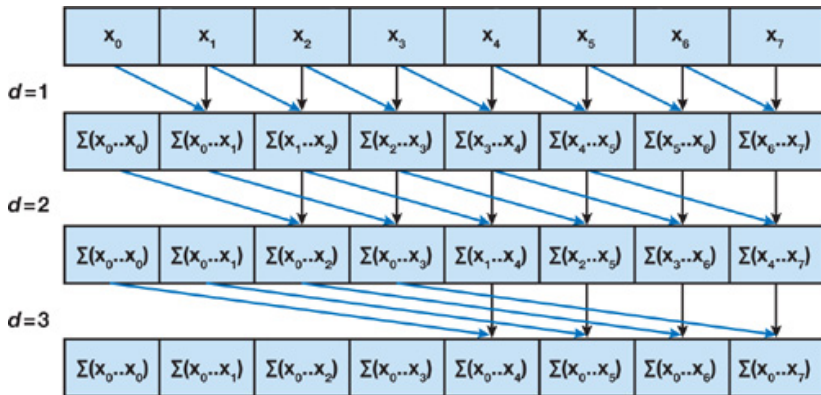
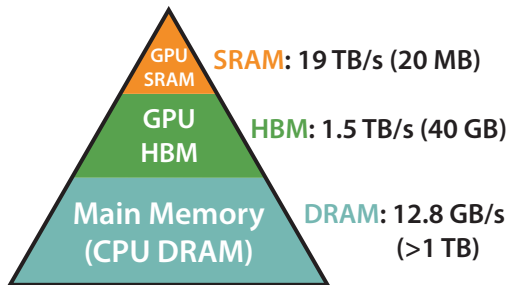


Figure: Associative binary operator only takes $O(T \log T)$ FLOPs: Each row is the addition of two vectors with length $O(T)$. The depth is $O(\log T)$ as the shift of above two vectors are doubled in each layer.¹⁹

¹⁹Figure from <https://developer.nvidia.com/gpugems/gpugems3/part-vi-gpu-computing/chapter-39-parallel-prefix-sum-scan-cuda>



Memory Hierarchy with Bandwidth & Memory Size

Figure: The implementation idea is similar as FlashAttention²⁰: They have better utilization of GPU SRAM and reduce the context switch in the kernel.

²⁰Tri Dao et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. arXiv: 2205.14135.

- Low rank storage. Realized matrix $B * L * D * N$, stored by smaller tensor $B * L * D$ and $D * N$.
- Re-computation in back propogation. (More FLOPs, Less IO).

21

²¹https://github.com/state-spaces/mamba/blob/main/csrc/selective_scan/selective_scan.cpp

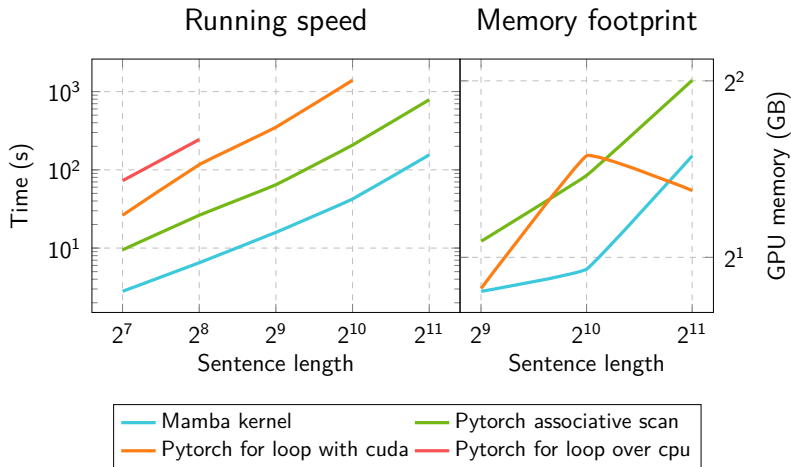


Figure: Speed comparison of different implementations

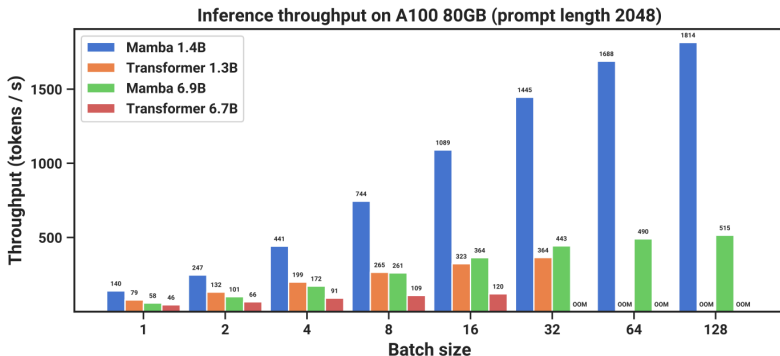


Figure: Inference throughput comparison against transformer. For **training**, if $T \geq 8k$, training of Mamba is faster than Flash-Attn2.^{22 23}

²²Tri Dao. *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*. 2023. [arXiv: 2307.08691](https://arxiv.org/abs/2307.08691).

²³<https://github.com/state-spaces/mamba/issues/156>

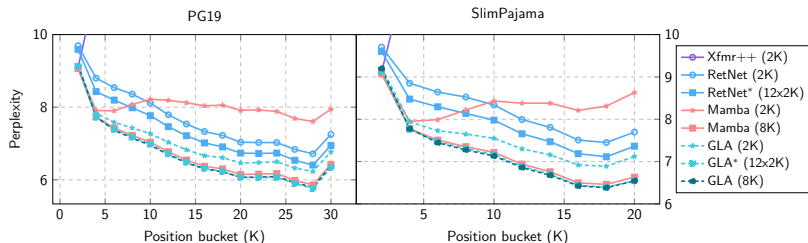


Figure: Length extrapolation performance on the PG19 and SlimPajama test set. They pretrain 1.3B models from scratch on SlimPajama for 100B tokens with different training length. Each bucket is of length 2048. * indicates using truncated BPTT with over 12 segments that are each of 2K-length.²⁴

²⁴Songlin Yang et al. *Gated Linear Attention Transformers with Hardware-Efficient Training*. Dec. 2023. arXiv: 2312.06635.

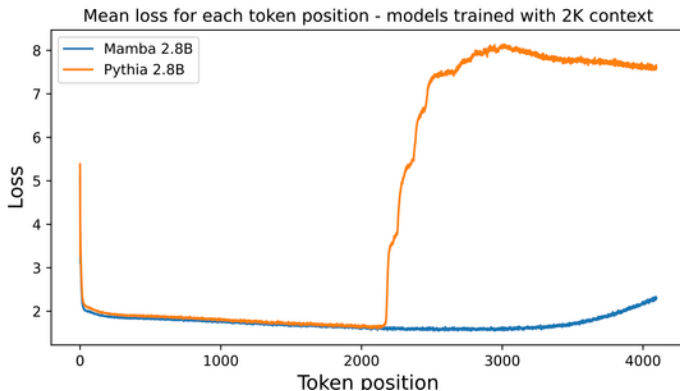











Figure: Length extrapolation of Pythia²⁵ (transformer-based) and Mamba






²⁵Stella Biderman et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. 2023. arXiv: 2304.01373.




Conclusion

- (Speed in theory) SSM layer is faster (in asymptotic sense) than Attention and RNN.
- (Memory in theory) Suitable reparameterization resolves the curse of memory common in recurrent models.
- (Speed in practice) Mamba provides a fast kernel implementation for state-space models with gating.
- (Memory in practice) Mamba and gated linear attention have better length extrapolation beyond training length.

-  Biderman, Stella et al. *Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling*. 2023. [arXiv: 2304.01373](#).
-  Dao, Tri. *FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning*. 2023. [arXiv: 2307.08691](#).
-  Dao, Tri et al. *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. 2022. [arXiv: 2205.14135](#).
-  Gu, Albert and Tri Dao. *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. Dec. 2023. [arXiv: 2312.00752](#).
-  Hua, Weizhe et al. “Transformer quality in linear time”. In: *International Conference on Machine Learning*. PMLR. 2022, pp. 9099–9117.

-  Jiang, Haotian et al. “A Brief Survey on the Approximation Theory for Sequence Modelling”. In: *Journal of Machine Learning* 2.1 (June 2023), pp. 1–30. ISSN: 2790-203X, 2790-2048. DOI: 10.4208/jml.221221.
-  Li, Zhong et al. “On the Curse of Memory in Recurrent Neural Networks: Approximation and Optimization Analysis”. In: *International Conference on Learning Representations*. Oct. 2020.
-  Liu, Yujie et al. “Time series prediction based on temporal convolutional network”. In: *2019 IEEE/ACIS 18th International Conference on Computer and Information Science (ICIS)*. IEEE. 2019, pp. 300–305.
-  Martin, Eric and Chris Cundy. “Parallelizing Linear Recurrent Neural Nets Over Sequence Length”. In: *International Conference on Learning Representations*. Feb. 2018.

-  Mehta, Harsh et al. *Long Range Language Modeling via Gated State Spaces*. 2022. [arXiv: 2206.13947](#).
-  Orvieto, Antonio et al. “On the Universality of Linear Recurrences Followed by Nonlinear Projections”. In: (2023). [arXiv: 2307.11888](#).
-  Peebles, William and Saining Xie. “Scalable Diffusion Models with Transformers”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2023, pp. 4195–4205.
-  Tay, Yi et al. “Long Range Arena: A Benchmark for Efficient Transformers”. In: *CoRR* abs/2011.04006 (2020). [arXiv: 2011.04006](#).
-  Wang, Shida and Qianxiao Li. *StableSSM: Alleviating the Curse of Memory in State-space Models through Stable Reparameterization*. Nov. 2023. [arXiv: 2311.14495](#).

-  Wang, Shida, Zhong Li, and Qianxiao Li. “Inverse Approximation Theory for Nonlinear Recurrent Neural Networks”. In: *The Twelfth International Conference on Learning Representations*. Oct. 2023.
-  Wang, Shida and Beichen Xue. “State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory”. In: *Thirty-seventh Conference on Neural Information Processing Systems*. 2023.
-  Yang, Songlin et al. *Gated Linear Attention Transformers with Hardware-Efficient Training*. Dec. 2023. arXiv: 2312.06635.