Gu & Dao, 2024

STATS 631, Winter 2026

Impact

Gu, A., & Dao, T. (2024, May). Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*.

- cited 8426 times
- Sometimes considered an advance over transformer architectures.

Insights

" letting the SSM parameters be functions of the input addresses their weakness with discrete modalities, allowing the model to selectively propagate or forget information along the sequence length dimension depending on the current token"

"We integrate these selective SSMs into a simplified end-to-end neural network architecture without attention or even MLP blocks (Mamba). Mamba enjoys fast inference (5$\times$ higher throughput than Transformers) and linear scaling in sequence length, and its performance improves on real data up to million-length sequences."

Selective state space models

- Obtained by parameterizing the SSM based on the input
- *"hardware-aware algorithm that computes the model recurrently with a scan instead of convolution"*
  - Gemini: A "scan" (prefix sum) is a parallel algorithm that computes cumulative sums of elements. It is foundational to parallel algorithms
  - Sengupta, S., Harris, M., Zhang, Y., & Owens, J. D. (2007). Scan primitives for GPU computing. https://doi.org/10.2312/EGGH/EGGH07/097-106

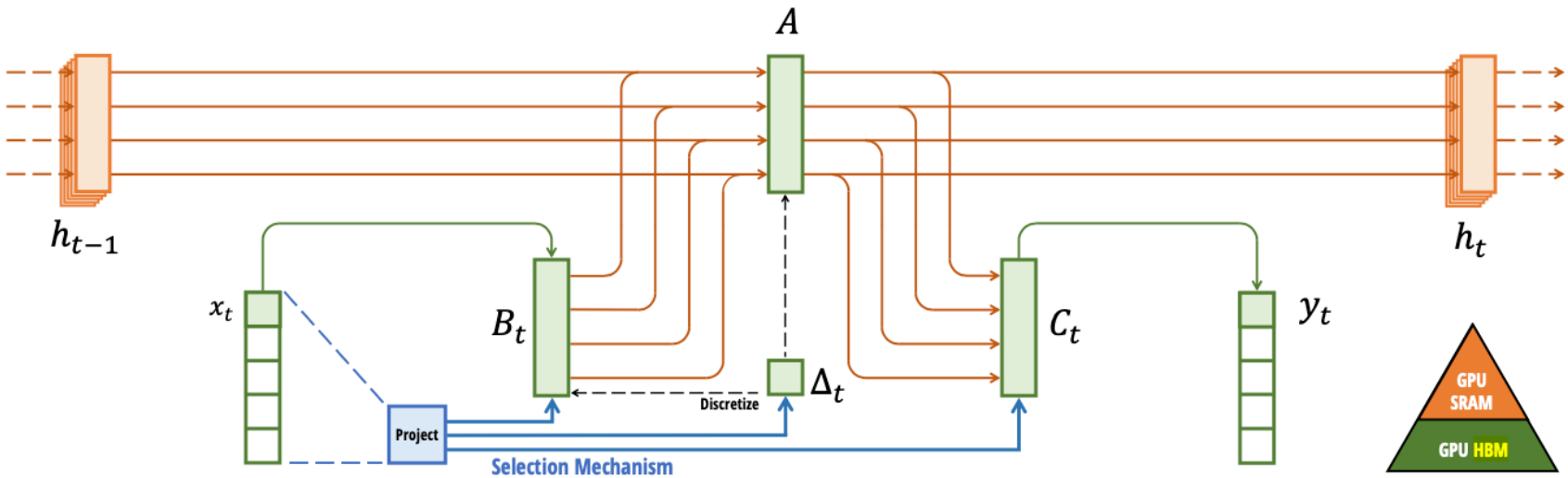

Figure 1: (**Overview.**) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time-invariance: the (Δ, $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$) 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.

Selection

- In a similar category as gating in LSTM or attention in LLM.
- Δ, B, C are learned functions of the input

Evaluation

- How did the authors convince people that their method works well?
- Is Mamba being used in current AI models?
- It took 5yr from "Attention is all you need" to the first ChatGPT product

Gemini response I

AI21 Labs (Jamba): AI21 released Jamba, the first production-grade, large-scale hybrid Transformer–Mamba model. It combines Mamba layers with attention mechanisms to handle 256K-token context lengths.

Mistral AI (Codestral Mamba): Mistral released Codestral Mamba, a 7.3B parameter model built specifically on the Mamba2 architecture for code generation and reasoning, supporting 256K context tokens.

NVIDIA (Nemotron-H): NVIDIA introduced Nemotron-H, a family of hybrid Mamba–Transformer models that replace up to 92% of attention layers with Mamba2 blocks to increase throughput, with models ranging from 8B to 56B parameters.

Gemini response II

IBM (Granite 4.0 / Bamba): IBM's Granite 4.0 and the Bamba-9B model are open-source hybrid models combining Transformers with Mamba2, optimized for high-performance, long-context inference.
Together AI (Mamba-Llama): Together AI developed Mamba-Llama, a hybrid model that replaces 75% of a LLaMA-3-8B model's attention layers with Mamba blocks, reducing inference latency by up to 5×.

Technology Innovation Institute (Falcon Mamba): TII released Falcon Mamba 7B, a pure, non-transformer Mamba-based language model.

Tencent (Hunyuan-TurboS): Tencent introduced this hybrid Transformer-Mamba2-MoE model, which uses 560B total parameters (56B active) to improve speed and context handling.

Gemini response III

Microsoft (Phi-4-mini-flash-reasoning): This 3.8B model uses a "SambaY" architecture, a hybrid of Mamba and attention, designed for low-latency, edge-device reasoning.

Why have lead AI models not incorpotated mamba I

Performance Trade-offs: While Mamba excels in processing long sequences with lower memory overhead, it can struggle with "memory-intensive" tasks where detailed, precise recall is required compared to Transformers.

Lack of Production-Grade Maturity: Mamba is a newer architecture, whereas Transformer models have benefited from years of intensive optimization, ecosystem support, and specialized hardware acceleration (e.g., NVIDIA GPU advancements).

The "Hybrid" Solution: Rather than a direct replacement, industry leaders are increasingly favoring hybrid models (like AI21's Jamba), which combine Mamba's efficient inference with Transformer's strong attention mechanism for complex reasoning.

Why have lead AI models not incorpotated mamba II

Hardware Ecosystem: Current AI infrastructure is heavily optimized for the parallel, quadratic attention mechanisms of Transformers, whereas Mamba requires different, highly optimized, and specialized hardware algorithms to realize its full potential.

Emergence of Competitors: The research landscape has shifted, with new alternatives like Gated Deltanet, RWKV7, and TTT (Time-Time-Time) potentially surpassing early Mamba implementations.