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

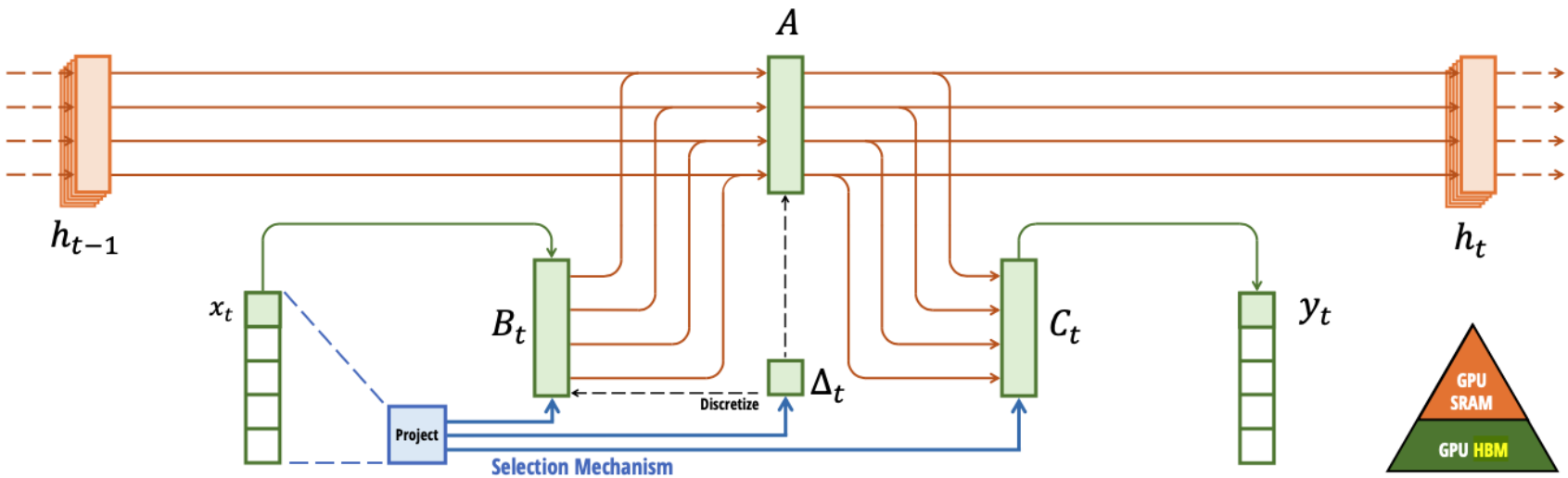

Figure 1: (**Overview.**) Structured SSMs independently map each channel (e.g. $D = 5$) of an input x to output y through a higher dimensional latent state h (e.g. $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate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