On the State of State Space Models

By Philipp Nazari

1. What are State-Space Models?

Dynamical Systems to Linear RNNs

- Continuous dynamical system:

$$\dot{\boldsymbol{h}}(t) = \boldsymbol{A}\boldsymbol{h}(t) + \boldsymbol{B}\boldsymbol{x}(t)$$
$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{h}(t)$$

- Discretization yields a **Linear RNN**:

$$\boldsymbol{h}_t = \tilde{\boldsymbol{A}}\boldsymbol{h}_{t-1} + \tilde{\boldsymbol{B}}\boldsymbol{x}_t$$

- where $\tilde{\boldsymbol{A}} = \exp(\Delta \boldsymbol{A})$ and $\tilde{\boldsymbol{B}} = \boldsymbol{A}^{-1}\left(\tilde{\boldsymbol{A}} - \boldsymbol{I}\right)\boldsymbol{B}$.

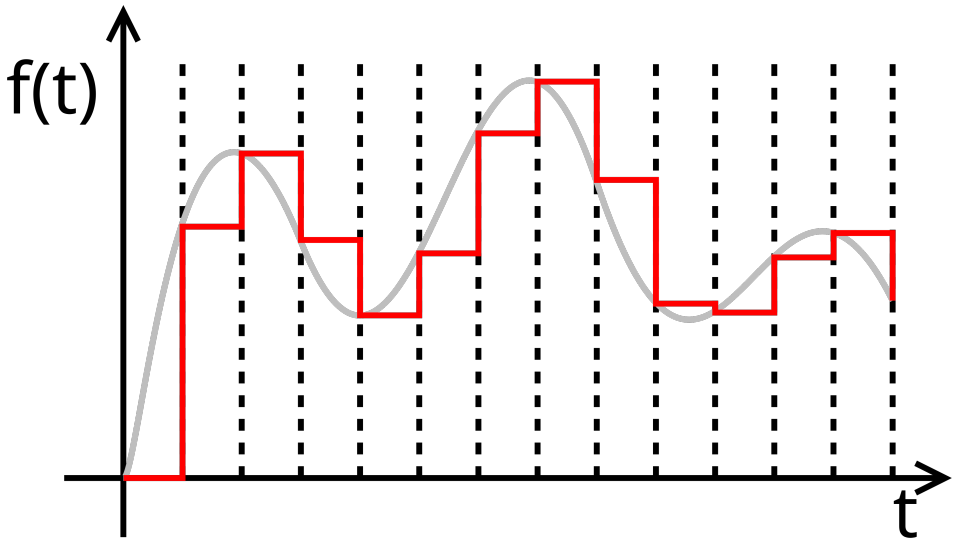

Figure 1: Zero-Order Hold discretization

The Case of LRU

LRU[1]:

$$\boldsymbol{h}_t = \mathrm{diag}\left(e^{i\theta_k+\lambda_k}\right)\boldsymbol{h}_{t-1} + \gamma \boldsymbol{B}\boldsymbol{x}_t \in \mathbb{C}^d$$

$\boldsymbol{h}_t$ acts as a fixed-size **summary of the past**

- **Questions**:
  1. How well does model **utilize** this fixed-size hidden state?
  2. Can we make it more **efficient**?
- We want **large states** for better memory, but **small states** for efficiency.

[1] A. Orvieto et al. Resurrecting Recurrent Neural Networks for Long Sequences.

3. CompreSSM: In-Training Compression

Controllability & Observability Gramians

- **identify** and **truncate** unused state dimensions during training[2].
- Solve the discrete Lyapunov equations for the **Gramians**:

$$\tilde{A}P\tilde{A}^\top + \tilde{B}\tilde{B}^\top = P, \qquad \tilde{A}^\top Q\tilde{A} + C^\top C = Q$$

- $P =$ **Controllability** (input - hidden state)
- $Q =$ **Observability** (hidden state - output)

A state dimension that is hard to reach **and** hard to observe is essentially dead weight.

---

[2] Makram Chahine, **Philipp Nazari**, Daniela Rus, T. Konstantin Rusch. *CompreSSM: The Curious Case of In-Training Compression of State Space Models*. arXiv preprint arXiv:2510.02823 (2025).

Hankel Singular Values & Balanced Truncation

- **Hankel Singular Values (HSVs)** combine both perspectives into a single importance ranking:

$$\sigma_i = \sqrt{\lambda_i(\boldsymbol{PQ})}, \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$$

- **Balanced realization**: transform coordinates. States with small σ_i are both hard to control and hard to observe.
- **Truncation**: discard the tail of the balanced system.

The CompreSSM Algorithm

- **The Pipeline**. During the **first 10%** of training:
  1. Extract matrices and compute HSVs periodically.
  2. Sort and truncate the state dimension r based on an energy threshold.
  3. Replace weights and continue training.

CompreSSM Results (CIFAR10)

- **performance close to a full-size model** at the **speed and memory footprint** of a much smaller one.

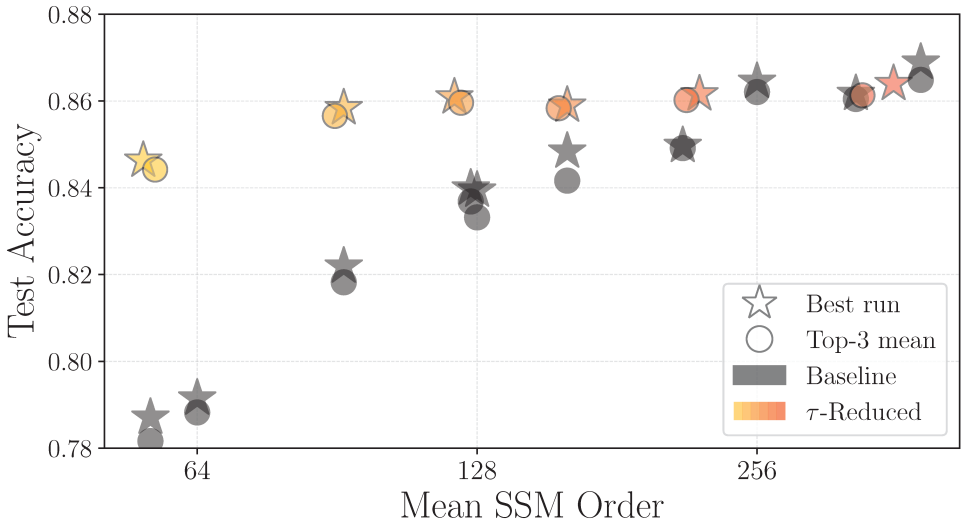

Figure 2: Test Acc vs. Final Dim

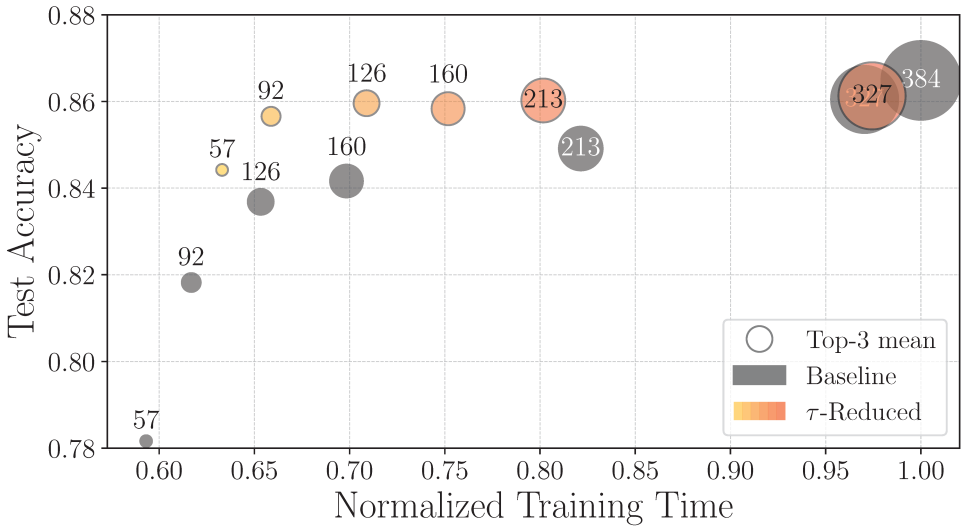

Figure 3: Test Acc vs. Train Time

5. The Key to State Reduction in Linear Attention

Linear Attention as an SSM

- **Drop the softmax.** The output at position t becomes:

$$\boldsymbol{o}_t = \sum_{s=1}^{t} (\boldsymbol{q}_t^\top \boldsymbol{k}_s) \boldsymbol{v}_s = \boldsymbol{q}_t^\top \underbrace{\sum_{s=1}^{t} \boldsymbol{k}_s \boldsymbol{v}_s^\top}_{\boldsymbol{S}_t \in \mathbb{R}^{d_k \times d_v}}$$

- The past is compressed into a **fixed-size state** S_t:

$$\boldsymbol{S}_t = \boldsymbol{S}_{t-1} + \boldsymbol{k}_t \boldsymbol{v}_t^\top, \qquad \boldsymbol{o}_t = \boldsymbol{q}_t^\top \boldsymbol{S}_t.$$

Rank Collapse in Linear Attention

- **Observation**: $S_t \in \mathbb{R}^{d_k \times d_v}$ exhibit a distinct **low-rank structure**[3].
- **Post-training reduction**: faster and more memory-efficient models.
- **Method**: structured pruning method that **jointly selects query- and key-channels**.

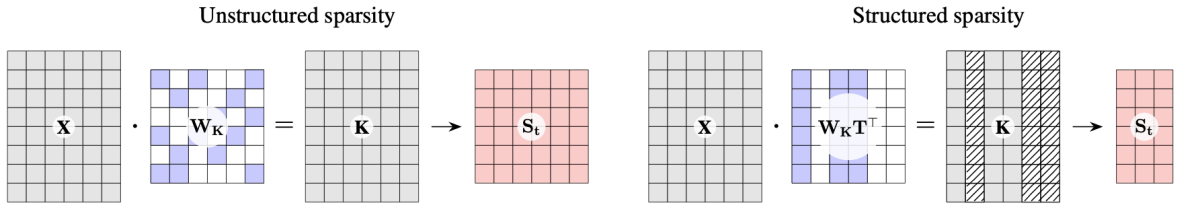

Figure 1. **Left:** Unstructured pruning yields a sparse weight matrix $\mathbf{W_K}$ yet preserves the column dimension of $\mathbf{K}$, leaving the domain of the state matrix $\mathbf{S}_t \in \mathbb{R}^{d_v, d_k}$ invariant. **Right:** Structured pruning eliminates basis vectors, mapping keys to a lower-dimensional space $\mathbb{R}^{d'_k}$ where $d'_k < d_k$. This results in a compressed state $\mathbf{S}_t \in \mathbb{R}^{d_v, d'_k}$. This reduction strictly decreases the FLOP count required to compute the recurrence. This figure is inspired by Ashkboos et al. (2024a, Figure 1).

[3] **Philipp Nazari**, T. Konstantin Rusch. *The Key to State Reduction in Linear Attention: A Rank-based Perspective*. arXiv preprint arXiv:2602.04852 (2026).

Empirical Results: 50% Reduction

Can remove ≈ **50% of the query and key channels**

Table 1. Comprehensive post-RFT evaluation of Gated DeltaNet 1.3B. We report Wikitext and Lambada perplexities, the average zero-shot generation accuracy (**ZS**, Gao et al. (2024)) as well as average retrieval accuracy (**Ret**, Arora et al. (2024)) across varying compression ratios (best results in **bold**, second best underlined).

	75% Compression				50% Compression				40% Compression				30% Compression			
Method	Wiki ↓	LMB ↓	ZS ↑	Ret ↑	Wiki ↓	LMB ↓	ZS ↑	Ret ↑	Wiki ↓	LMB ↓	ZS ↑	Ret ↑	Wiki ↓	LMB ↓	ZS ↑	Ret ↑
Rand	19.3	18.1	54.6	23.9	16.9	11.5	57.8	32.1	16.5	10.7	58.3	34.6	16.1	10.2	58.7	37.3
L1	18.6	16.8	55.3	25.7	16.8	13.0	57.2	32.9	16.5	11.9	57.8	34.6	16.3	11.3	58.3	35.4
S-Wanda	18.2	15.1	55.4	26.2	16.6	11.7	57.8	33.2	16.4	11.5	58.1	34.4	16.1	11.0	58.6	37.2
Grad	**17.8**	**12.5**	**56.8**	**26.9**	16.4	**10.5**	**58.3**	**34.3**	**16.1**	**10.1**	58.6	**35.8**	**15.9**	**9.9**	**59.0**	**38.0**
DRRQR	17.9	14.7	56.2	26.1	**16.3**	12.0	58.0	33.0	**16.1**	11.0	**58.7**	35.0	**15.9**	10.7	58.8	36.5
Baseline	16.8	9.7	59.4	40.3												

Figure 4: Perplexity and zero-shot performance

Thank You! Questions?