lrnn-lib

SAiDL team
saidl.root@gmail.com

Abstract

Linear RNN Models have recently gained a lot of traction in Seq2Seq modeling. Several models have been developed along with an extensive theory detailing how to train them stably. These architectures are, however, distributed across software frameworks and some even require modifying CUDA-kernels for baseline performance; a few have no public implementations. This limits their general purpose usage drastically, this letter proposes `lrnn-lib`, a simple library that exposes multiple levels of interface for several current models to aid community adoption.

1 How did we get to Linear RNNs?

Typically, a non-linear RNN is the "conventional" form of an RNN, described by

$$\begin{aligned} x_k &= \alpha(W_{xx}x_{k-1} + W_{xu}u_k) \\ y_k &= \beta(W_{yx}x_k) \end{aligned} \tag{1}$$

Where α and β are non-linear functions. These non-linearities are actually responsible for most of the claims about RNN-expressivity (e.g. Turing Completeness (Chung & Siegelmann, 2021)). Equation (1) suffers from two major problems which impede scale.

- **The vanishing / exploding gradient problem** (Hochreiter & Schmidhuber, 1997), which affects the model's ability to learn long-range dependencies,
- **Sequential computation during training**, this prevents utilization of modern hardware developments like GPUs, TPUs, etc.

Despite these problems, an attractive feature of RNNs is the $\mathcal{O}(1)$ time-complexity during inference, opposed to $\mathcal{O}(n^2)$ for transformers (Vaswani et al., 2017) (where n is the input sequence length).

Linear RNNs aim to solve both of these problems while staying performant on downstream tasks and maintain the $\mathcal{O}(1)$ time complexity during inference.

1.1 Applications

Linear RNNs have been used as sequence modelers in a variety of domains, from Audio: Text-to-speech (Goel et al., 2022), ASR (Pei, 2025), Enhancement (Pei, 2025; Pei et al., 2025), RNA modeling (Ramesh et al., 2025), Vision (Liu et al., 2024), Event-streams (Schöne et al., 2024b) and even Point-clouds (Han et al., 2024). In addition, they have set new benchmarks on synthetic tasks in the long-range-arena (LRA) (Tay et al., 2021). Typically, transformers are neither efficient, nor effective for long sequences ($\geq 2^{10}$) (Yu et al., 2024) which is where these models prove extremely useful.

1.2 Why `lrnn-lib`?

Apart from the framework problem mentioned before, there are interesting aspects of these models which make different models applicable for different tasks. An example of this is also described in Centaurus (Pei, 2025), they show that non-LTI SSMs (like Mamba) are actually *difficult* to train for continuous modalities like Audio, however, these models are

very performant on discrete modalities like Text (Gu & Dao, 2024). Additionally, the scale of data available and the hardware requirements for a problem may also impact the Linear RNN to use.

2 `lrnn-lib` Models

We plan to implement the models listed in Table 1, detailed descriptions can be found in individual papers (models listed in chronological order).

Model	SISO	LTI	Public Implementation	Framework
S4 (Gu et al., 2022)	✓	✓	✓	PyTorch
S5 (Smith et al., 2023)	✗	✓	✓	JAX
LRU (Orvieto et al., 2023)	✗	✓	✗	N/A
Event-SSM (Schöne et al., 2024b)	✗	✓	✓	JAX
S6 (Gu & Dao, 2024)	✓	✗	✓	PyTorch
STREAM (Schöne et al., 2024a)	✓	✗	✗	N/A
S7 (Soydan et al., 2024)	✗	✗	✗	N/A
Centaurus (Pei, 2025)	✗	✗	✓	PyTorch

Table 1: Comparison of Models (SISO: Single-Input Single-Output, LTI: Linear Time-Invariant).

The core recurrence still boils down to eq. (1), these models analyze it from a signal processing perspective, which allows them to talk about stability constraints on model parameters.

If this project excites you please fill the form below by **31st May EOD IST (strict deadline)**:

`https://forms.gle/MVGve9hMnhZHgmHr9`

References

Stephen Chung and Hava Siegelmann. Turing completeness of bounded-precision recurrent neural networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 28431–28441. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/ef452c63f81d0105dd4486f775adec81-Paper.pdf.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Re. It's raw! Audio generation with state-space models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7616–7633. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/goel22a.html.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces, 2024. URL https://arxiv.org/abs/2312.00752.

Albert Gu, Karan Goel, and Christopher Re. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=uYLFoz1vlAC.

Xu Han, Yuan Tang, Zhaoxuan Wang, and Xianzhi Li. Mamba3d: Enhancing local features for 3d point cloud analysis via state space model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 4995–5004, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681173. URL https://doi.org/10.1145/3664647.3681173.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9 (8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, and Yunfan Liu. Vmamba: Visual state space model. *arXiv preprint arXiv:2401.10166*, 2024.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences, 2023. URL https://arxiv.org/abs/2303.06349.

Yan Ru Pei. Let SSMs be convnets: State-space modeling with optimal tensor contractions. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PkpNRmBZ32.

Yan Ru Pei, Ritik Shrivastava, and FNU Sidharth. atennuate: Optimized real-time speech enhancement with deep ssms on raw audio, 2025. URL https://arxiv.org/abs/2409.03377.

Krithik Ramesh, Sameed M. Siddiqui, Albert Gu, Michael D. Mitzenmacher, and Pardis C. Sabeti. Lyra: An efficient and expressive subquadratic architecture for modeling biological sequences, 2025. URL https://arxiv.org/abs/2503.16351.

Mark Schöne, Yash Bhisikar, Karan Bania, Khaleelulla Khan Nazeer, Christian Mayr, Anand Subramoney, and David Kappel. Stream: A universal state-space model for sparse geometric data, 2024a. URL https://arxiv.org/abs/2411.12603.

Mark Schöne, Neeraj Mohan Sushma, Jingyue Zhuge, Christian Mayr, Anand Subramoney, and David Kappel. Scalable event-by-event processing of neuromorphic sensory signals with deep state-space models, 2024b. URL https://arxiv.org/abs/2404.18508.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Taylan Soydan, Nikola Zubić, Nico Messikommer, Siddhartha Mishra, and Davide Scaramuzza. S7: Selective and simplified state space layers for sequence modeling, 2024. URL https://arxiv.org/abs/2410.03464.

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena : A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qVyeW-grC2k.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan (Celine) Lin. Unveiling and harnessing hidden attention sinks: enhancing large language models without training through attention calibration. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.