Efficient Sequence Modeling: Spring 2025 PURA

Yubin Kim
Georgia Institute of Technology
ykim3041@gatech.edu

1 Executive Summary

This research, conducted under the Spring 2025 Pennington Undergraduate Research Award (PURA), explores the development of efficient transformers, essential for sequence modeling tasks. Sequence modeling is a foundational component of machine learning research applicable to many fields, ranging from natural language processing to time series forecasting and image analysis. While traditional models like recurrent neural networks (RNNs) and their improved successors (LSTM, GRU) enabled deep learning models to incorporate contextual data, they suffered from limited scalability due to the lack of parallelization capabilities. This paradigm shifted in 2017 when transformer architectures were introduced, enabling parallel training and state-of-the-art performance across diverse domains. However, their quadratic time complexity has spurred interest in developing more computationally efficient alternatives.

In this project, we focused on proposing novel sequence modeling architectures with reduced computational complexity, while preserving or improving modeling accuracy. We developed mathematically principled models and validated them through extensive testing on downstream tasks, including Long Range Arena for general sequence tasks, WikiText for language modeling, and SC10 for speech classification. Such downstream tasks are closely tied to real-life applications, ensuring the high applicability of the created models.

Among the models evaluated, SAMoVAR stood out by achieving state-of-the-art performance in time series analysis tasks. By integrating autoregressive structures with efficient transformer design principles, SAMoVAR consistently outperformed competitive baselines on multivariate datasets such as ETT, traffic, solar, and electricity.

The findings of this research highlight promising directions for building scalable and effective sequence models, balancing performance and efficiency. Future work will extend this methodology to new architectures and applications, with the aim of publishing impactful results and advancing practical AI deployment.

2 Research Motivation (Related Works)

Throughout the history of AI and machine learning, sequence modeling has been a topic of high interest. Starting in the 1980s, recurrent neural networks kickstarted the development of deep learning methods specialized for sequence modeling [5]. Using a hidden state that carried information from the previous data points, recurrent neural network models were able to account for context, although drawbacks, including gradient vanishing and gradient explosion, were apparent. Subsequent research improved the performance of such recurrent neural networks, creating models like LSTM and GRU, which included gating mechanisms, improving the model's performance [4, 1]. Although this mostly solved the gradient explosion and vanishing problem of RNNs, a big roadblock was still present in the fact that the training process wasn't parallelizable, meaning that training these models was a very time and resource-consuming task, and scaling was pretty much impossible.

Beginning in 2017 the field of sequence modeling was revolutionized through the invention of transformers. [9] Training query, key, and value for each timestep, transformers were able to learn detailed contextual dependencies. Most importantly, transformers could be trained in parallel, without

having to go step by step in the temporal direction, enabling the training of extremely large models. This revolutionary breakthrough led to a boom in the AI industry, which can be felt in everyday life through large language models like ChatGPT.

Despite the upsides highlighted in the paragraph above, transformers are still criticized for their need for computing power, with larger models requiring the extensive use of countless GPUs for training. The high computational costs of transformers are caused by the fact that their training scales in quadratic time ($\mathcal{O}(n^2)$), and to address this issue, researchers have been developing sequence modeling architectures with lower time complexities. Linear transformers started the development of the field by suggesting a linear alternative to transformers, which was done through the elimination of the sigmoid activation function in context calculation [6]. Researchers have since been creating different efficient sequence modeling structures, including efficient transformers and state-of-the-art models, which include models like S4 and Mamba, attempting to reach the perfect balance between efficiency and performance [3, 2]. Through my research with the Pennington Undergraduate Research Award, I aim to contribute to the advancement of state-of-the-art AI architectures, particularly in improving the efficiency of sequence modeling models.

3 Research Objective

The research objectives of the Spring 2025 PURA award period were the following:

1. Propose novel sequence modeling architectures capable of effectively modeling sequential data with lower computational complexity. The effectiveness of the proposed models should be backed up mathematically.
2. Verify the performance of the developed models by performing the following downstream tasks:
    - Natural Language Processing
    - Time Series Analysis
    - Image Analysis
    - Speech Classification
    - Etc.

4 Research Methodology

After identifying a potential breakthrough in the field of efficient transformers, the first step for validating our model was to create mathematical representations. These mathematical representations were supplemented with visual diagrams to improve intuition.

Given our limited computing resources, it was natural to prioritize smaller datasets for experiments. We tested the models in the following tasks using the following datasets:

- **Long Range Arena (LRA)** [8]: A dataset we used extensively is the famous Long Range Arena. Long Range Arena consists of ListOps, text classification, document retrieval, image classification, pathfinder, and pathfinder-x tasks. ListOps tests a model's hierarchical reasoning capabilities, text and document retrieval tests a model's ability to analyze natural language data, image classification tests the model's capability of handling images, and pathfinder tasks measure long-distance spatial reasoning capabilities. Most of the tasks laid out in the research objective section were performed utilizing the Long Range Arena dataset.- A dataset we used extensively is the famous Long Range Arena. (cite) Long Range Arena consists of ListOps, text classification, document retrieval, image classification, pathfinder, and pathfinder-x tasks. ListOps tests a model's hierarchical reasoning capabilities, text and document retrieval tests a model's ability to analyze natural language data, image classification tests the model's capability of handling images, and pathfinder tasks measure long-distance spatial reasoning capabilities. Most of the tasks laid out in the research objective section were performed utilizing the Long Range Arena dataset.
- **WikiText**: To further test our models on natural language processing tasks, we utilized the WikiText dataset. The WikiText dataset is a collection of over 100 million tokens extracted

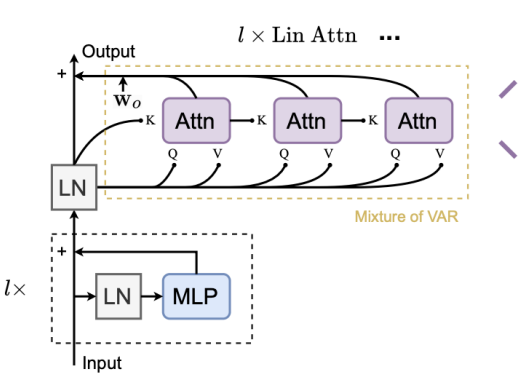
from selected articles on Wikipedia. Being composed of full articles, the WikiText dataset effectively tests the ability to model long-term dependencies.

- **SC10**: For speech classification, we used the Speech Commands dataset, SC10. SC10 is an audio dataset of spoken commands in a 1D raw waveform sequence sampled at 16 kHz. Due to the extremely long input length of 16000 tokens, this dataset effectively tests a model's ability to model extremely long-range continuous signals.

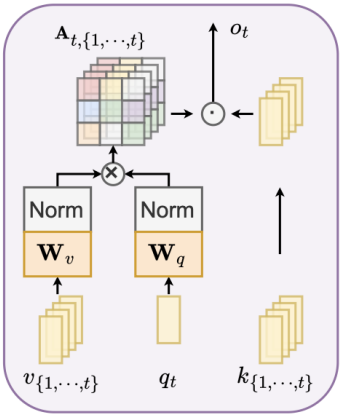
- Depending on the characteristics of each model, potential applications in other downstream tasks, such as DNA modeling, were tested.

5 Results

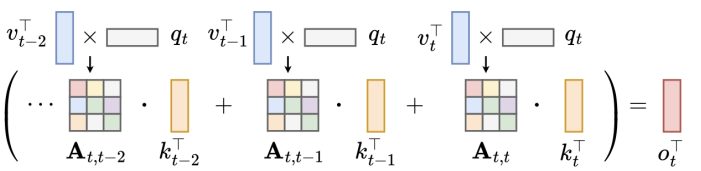
Through extensive experiments, we verified that many of our models achieved state-of-the-art or near-state-of-the-art performance on many downstream tasks, including natural language processing, image analysis, time series analysis, and DNA analysis. It is important to note that not all models behaved similarly in those downstream tasks. While there were models that worked in any general sequence, some models only showed exceptional performance on certain tasks.



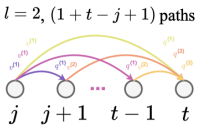
(d) Structural Aligned Mixture of VAR (SAMoVAR) Transformer for Autoregressive Forecasting



(e) Conceptual Illustration of VAR Linear Attention Block on Step t



(a) Illustration of the VAR Structure for Step t in a Single Linear Attention Layer



(b) Traversal and Combination of All Possible Temporal Influence Paths as the Final Mixture of VAR Weights

For an example, I will use the results for the SAMoVAR model submitted for ICML and is currently available via preprint [7]. For this paper, we make alterations to the vector autoregressive model, aligning it with linear transformers, particularly in time series analysis use cases. The two figures above visualize the structure of SAMoVAR. Since we were focusing on time series data, we tested our model on time series multivariate time series datasets, including weather, solar, electricity, electricity transformer temperature (ETT), and traffic datasets. The table below shows the average MSE on each dataset after having tested on varying input lengths, compared to other state-of-the-art models, with the best-performing model bolded and the runner-up underlined.

Model	SAMoVAR	LinTrans	FixedVAR	CATS	iTransformer	FITS	PatchTST	Dlinear	EncFormer
Weather	**0.214**	0.217	0.247	0.216	0.232	0.222	0.221	0.233	0.251
Solar	**0.184**	0.189	0.430	0.206	0.219	0.209	0.202	0.216	0.212
ETTh1	**0.401**	0.419	0.564	0.408	0.454	0.440	0.413	0.422	0.906
ETTh2	0.324	0.346	0.391	**0.320**	0.374	0.354	0.330	0.426	0.877
ETTm1	**0.339**	0.346	0.519	0.345	0.373	0.354	0.346	0.347	0.735
ETTm2	**0.240**	0.243	0.278	0.243	0.265	0.247	0.247	0.252	0.576
ECL	**0.151**	0.166	0.345	**0.151**	0.170	0.167	0.159	0.165	0.664
Traffic	0.391	0.438	0.717	**0.385**	0.414	0.418	0.391	0.431	0.824
PEMS03	**0.150**	0.188	0.375	0.225	0.212	0.234	0.230	0.254	0.443
PEMS04	**0.102**	0.136	0.404	0.184	0.171	0.256	0.222	0.246	0.377
PEMS08	**0.234**	0.261	0.674	0.359	0.271	0.296	0.290	0.357	0.681
AvgRank	**1.41**	3.41	8.16	2.86	5.43	5.20	4.00	5.82	8.30
#Top1	**29**	3	0	9	1	0	2	1	0

As you can see from the table above, our proposed model, SAMoVAR, beats other state-of-the-art models such as Linear Transformer, CATS, PatchTST, and Dlinear on time series analysis tasks. This suggests that our proposed model does bring a significant improvement in modeling long-term dependencies with real-world implications.

6 Conclusions and Future Works

By iteratively improving and testing novel models, we identified architectures like SAMoVAR that can achieve state-of-the-art results on sequence modeling tasks. Such models advance the field of efficient transformers, bringing real-life implications in improving the effectiveness and efficiency of natural language processing, time series analysis, and other sequence analysis models that are tightly knit to our everyday lives.

We will continue repeating and improving the process of identifying efficient transformers for sequence modeling tasks. The models identified as effective will be published through relevant academic conferences, encouraging the utilization of our breakthroughs in making advancements in various fields.

References

[1] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *EMNLP*, pages 1724–1734, 2014.

[2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[3] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *Advances in Neural Information Processing Systems*, 34:12874–12886, 2021.

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[5] John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, 79(8):2554–2558, 1982.

[6] Angelos Katharopoulos, Apoorv Vyas, Nicolas Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, pages 5156–5165. PMLR, 2020.

[7] Jiecheng Lu and Shihao Yang. Linear transformers as var models: Aligning autoregressive attention mechanisms with autoregressive forecasting. *arXiv preprint arXiv:2502.07244*, 2025.

[8] Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. Long range arena: A benchmark for efficient transformers. *arXiv preprint arXiv:2011.04006*, 2020.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.