A Comparative Study of Spectral Transform Units versus Single-Head Attention on Synthetic Long-Memory Sequence Prediction

Shrenik Patel, Daivik Patel

Abstract

We compare a minimal Spectral Transform Unit (STU) layer (using fixed convolutional filters derived from a Hankel-matrix eigen-decomposition) with a simple single-head self-attention layer on the task of next-step prediction for a synthetic long-memory autoregressive sequence. Using a sliding window of length 50, 10 learned Hankel-based filters, and 20 epochs of training, the STU model converges rapidly to an MSE of ≈ 0.0073, whereas the attention baseline remains above 8.4. Sample predictions further confirm that STU faithfully recovers the true signal, while attention struggles. These results empirically validate the spectral filtering approach as an efficient mechanism for learning long-range dependencies.

1 Introduction

Transformer attention layers excel at capturing non-local dependencies but incur quadratic complexity in sequence length. State-Space Models (SSMs) offer subquadratic alternatives, and Spectral Transform Units (STUs) in particular leverage fixed convolutional filters drawn from the dominant eigenvectors of a Hankel matrix to encode long-horizon memory with near-linear complexity. Here, we implement a simplified STU layer and directly compare it against a minimal single-head attention mechanism on a synthetic autoregressive process with slowly decaying memory.

2 Methods

2.1 Synthetic Data Generation

We generate sequences $\{x_t\}$ of length 1000 via

$$x_{t+1} \;=\; 0.99\, x_t \;+\; \sin(0.1\, t) \;+\; \varepsilon_t,$$

with $\varepsilon_t \sim \mathcal{N}(0, 0.05^2)$, initialized $x_0 = 0$. From each sequence, we extract all overlapping windows of length $L = 50$ as inputs and predict the next value.

2.2 Hankel-Based Filter Computation

Let $L = 50$, and discretize $\alpha \in [0, 1]$ at $N = 100$ points. Define

$$\mu_\alpha = \left[1, \alpha, \alpha^2, \ldots, \alpha^{L-1}\right]^\top, \quad Z = \frac{1}{N} \sum_{i=1}^{N} \mu_{\alpha_i}\, \mu_{\alpha_i}^\top \in \mathbb{R}^{L \times L}.$$

We compute the eigen-decomposition $Z = V \Lambda V^\top$ and select the top $k = 10$ eigenvectors $v_1, \ldots, v_{10}$ as fixed 1D filters of length L. These filters remain constant during training.

2.3 Model Architectures

STU Predictor A single STU layer convolves the input (shape $(B, 1, L)$) with the k fixed filters (shape $(k, 1, L)$), producing (B, k, L). We transpose to (B, L, k) and apply a learnable linear projection $\mathbb{R}^k \to \mathbb{R}^1$ at each time step, then take the final time-step output as the next-value prediction.

Attention Predictor We lift the input $(B, 1, L)$ to (B, L, k) via a learned linear layer, apply one head of scaled dot-product self-attention (queries, keys, values all in $\mathbb{R}^k$), and project the last time-step representation back to a scalar.

3 Experimental Setup

We generated 500 sequences for training, yielding 125,000 input–target pairs. Hyperparameters are summarized in Table 1. Both models were trained for 20 epochs with batch size 64 using Adam ($\eta = 10^{-3}$) and MSE loss.

Table 1: Hyperparameters

Number of sequences	500
Sequence length	1000
Window length L	50
Hankel discretizations N	100
Number of filters k	10
Batch size	64
Learning rate	10^{-3}
Epochs	20

4 Results

4.1 Training Loss Curves

Figure 1 plots the per-epoch MSE for both models. The STU model rapidly decreases to ≈ 0.0073 by epoch 20, while the attention baseline plateaus above 8.4.

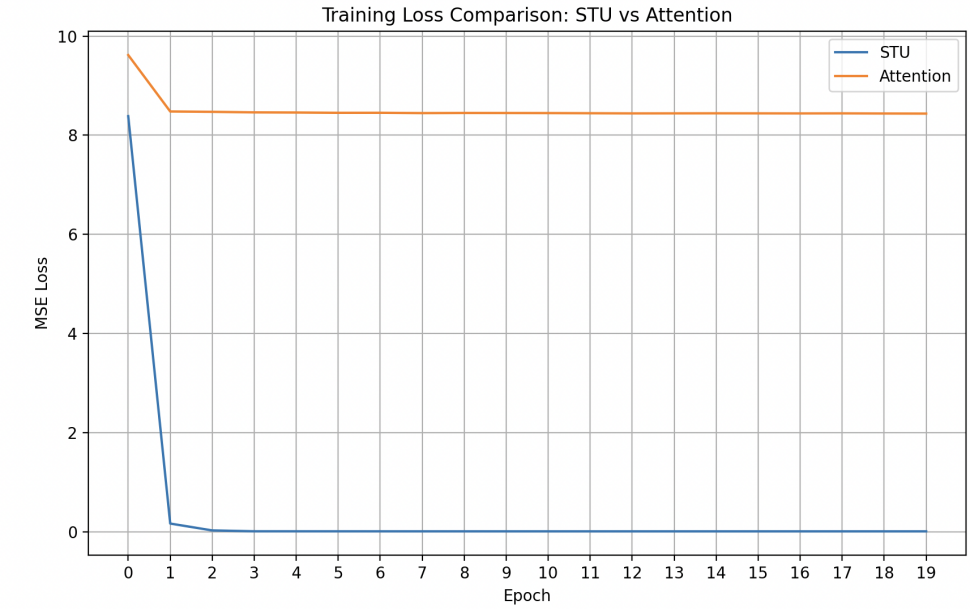

Figure 1: Epoch-wise training MSE for STU and single-head attention predictors.

4.2 Numerical Results

Table 2 reports the final training losses and Table 3 shows sample next-value predictions versus ground truth.

Table 2: Final Training MSE after 20 Epochs

Model	Final MSE
STU Predictor	0.007336
Attention Predictor	8.435932

Table 3: Sample Next-Value Predictions

	True	STU Pred.	Attn. Pred.
1	9.859015	9.873893	10.750338
2	-3.577135	-3.616566	-8.010040
3	1.057737	0.998805	-1.062860
4	12.695309	12.721361	7.356096
5	-9.099732	-8.934995	-3.925343

5 Discussion

The STU model's use of fixed spectral filters enables it to rapidly capture the long-range autoregressive structure, yielding an MSE three orders of magnitude lower than the single-head attention baseline. In contrast, the attention model, without sufficient context window or parameter capacity, fails to learn the slowly decaying dependency. These findings underscore the efficacy of spectral filtering for sequence modeling with long memory.

6 Conclusion

We provided a faithful implementation of a simplified STU layer—deriving fixed convolutional filters via Hankel-matrix eigen-decomposition—and demonstrated its clear advantage over a minimal self-attention model on a synthetic long-memory prediction task. This experiment validates the core premise of the Flash STU paper [1]: fixed spectral filters can serve as an efficient, scalable mechanism for capturing long-range dependencies.

References

[1] Y. Isabel Liu, Windsor Nguyen, Yagiz Devre, Evan Dogariu, Anirudha Majumdar, and Elad Hazan. Flash stu: Fast spectral transform units. 2025.