FMA

fazad

August 2025

1 FM-based Multi-head Attention

We define the **FM-approximated multi-head attention** as a function

$$\tilde{M}(x) : \mathbb{R}^{n \times d} \to \mathbb{R}^n$$

which approximates standard self-attention using factorized projections and low-rank interactions.

This function is defined as:

$$\tilde{M}(x) = \tilde{U}(\tilde{L}(x))$$

where:

- $\tilde{L}(x) : \mathbb{R}^{n \times d} \to \mathbb{R}^{n \times d_h \times h}$ is a learned linear transformation mapping input x to a tensor structured for h heads,
- $\tilde{U}(y) : \mathbb{R}^{n \times d_h \times h} \to \mathbb{R}^n$ applies head-wise scoring using a shared projection.

For each head $k \in \{1, \ldots, h\}$, we compute:

$$y_k = \tilde{L}_k(x) \in \mathbb{R}^{n \times d_h}, \qquad \alpha_k = \text{softmax}(y_k W_k + b_k) \in \mathbb{R}^n$$

where $W_k \in \mathbb{R}^{d_h \times 1}$ and $b_k \in \mathbb{R}$ are parameters of a shallow scoring layer.

The attention scores are aggregated by averaging across heads:

$$\tilde{M}(x) = \frac{1}{h} \sum_{k=1}^{h} \alpha_k$$

The final attended representation is computed as:

$$z = \sum_{i=1}^{n} \tilde{M}(x)_i \cdot x_i \in \mathbb{R}^d$$

and broadcast across all positions to produce a uniform contextual representation:

$$Z = \text{repeat}(z, n) \in \mathbb{R}^{n \times d}$$

Time Complexity Comparison

Standard Multi-head Self-Attention

Let $x \in \mathbb{R}^{n \times d}$ be the input. The standard multi-head self-attention involves:

- Query/Key/Value projections: $O(nd^2)$
- Attention score matrix computation: $O(n^2 d)$
- Output projection: $O(n^2 d)$

Total time complexity:

$$\boxed{O(nd^2 + n^2 d)}$$

FM-Based Attention

Given the same input $x \in \mathbb{R}^{n \times d}$, the FM-inspired attention involves:

- Linear projection $L(x)$: $O(nd^2)$
- Head-wise scoring and softmax: $O(nd)$
- Final aggregation and reweighting: $O(nd)$

Total time complexity:

$$\boxed{O(nd^2 + nd)}$$

Summary Table

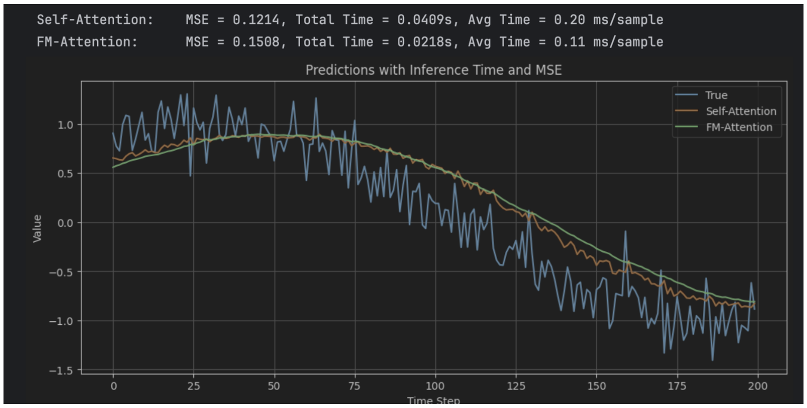
Method	**Time Complexity**	**Dominant Term (when $d \gg n$)**
Standard Attention	$O(nd^2 + n^2 d)$	$O(n^2 d)$
FM-Based Attention	$O(nd^2 + nd)$	$O(nd^2)$

Table 1: Time complexity comparison between standard and FM-based attention

2 Preliminary Results

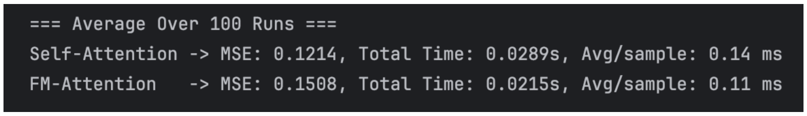
For the initial results, the values of the sequence length, n were set to 50, 100 and 500. The forecast horizon is set to 200. The value of dimension, d is set to 16. The findings are as follows:

2.1 n = 50, d = 16



```
Self-Attention:    MSE = 0.1214, Total Time = 0.0409s, Avg Time = 0.20 ms/sample
FM-Attention:      MSE = 0.1508, Total Time = 0.0218s, Avg Time = 0.11 ms/sample
```

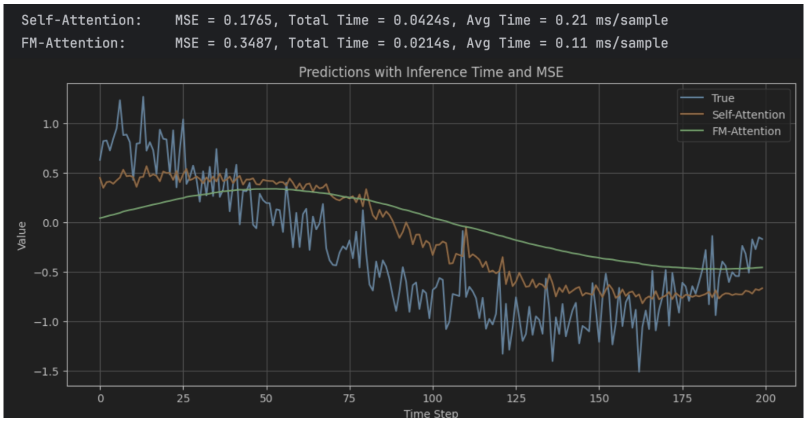
Figure 1: n=50, single execution



```
=== Average Over 100 Runs ===
Self-Attention -> MSE: 0.1214, Total Time: 0.0289s, Avg/sample: 0.14 ms
FM-Attention   -> MSE: 0.1508, Total Time: 0.0215s, Avg/sample: 0.11 ms
```

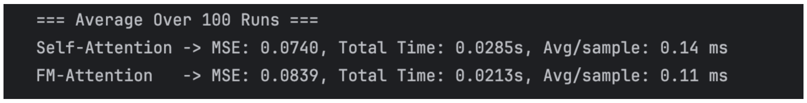
Figure 2: n=50, average of 100 executions

2.2 n = 100, d = 16



```
Self-Attention:    MSE = 0.1765, Total Time = 0.0424s, Avg Time = 0.21 ms/sample
FM-Attention:      MSE = 0.3487, Total Time = 0.0214s, Avg Time = 0.11 ms/sample
```

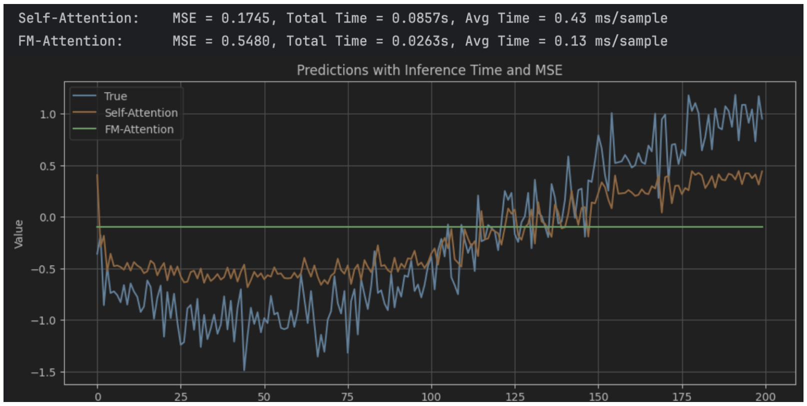
Figure 3: n=100, single execution



```
=== Average Over 100 Runs ===
Self-Attention -> MSE: 0.0740, Total Time: 0.0285s, Avg/sample: 0.14 ms
FM-Attention   -> MSE: 0.0839, Total Time: 0.0213s, Avg/sample: 0.11 ms
```

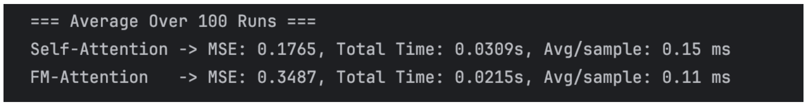
Figure 4: n=100, average of 100 executions

2.3 n = 500, d = 16



```
Self-Attention:    MSE = 0.1745, Total Time = 0.0857s, Avg Time = 0.43 ms/sample
FM-Attention:      MSE = 0.5480, Total Time = 0.0263s, Avg Time = 0.13 ms/sample
```

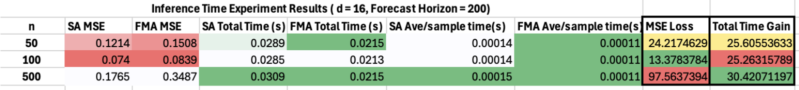
Figure 5: n=500, single execution



```
=== Average Over 100 Runs ===
Self-Attention -> MSE: 0.1765, Total Time: 0.0309s, Avg/sample: 0.15 ms
FM-Attention   -> MSE: 0.3487, Total Time: 0.0215s, Avg/sample: 0.11 ms
```

Figure 6: n=500, average of 100 executions

The source code is available with Github link[1]



Inference Time Experiment Results (d = 16, Forecast Horizon = 200)

n	SA MSE	FMA MSE	SA Total Time (s)	FMA Total Time (s)	SA Ave/sample time(s)	FMA Ave/sample time(s)	MSE Loss	Total Time Gain
50	0.1214	0.1508	0.0289	0.0215	0.00014	0.00011	24.2174629	25.60553633
100	0.074	0.0839	0.0285	0.0213	0.00014	0.00011	13.3783784	25.26315789
500	0.1765	0.3487	0.0309	0.0215	0.00015	0.00011	97.5637394	30.42071197

Figure 7: Initial Results with MSE and Execution Time

The results show that there is a tradeoff between accuracy and execution time. These initial results support the Problem Statement that if n is sufficiently large, there is significant gain is execution time. However, it comes at the cost of accuracy dropping significantly. The number of epochs were set to 20 for all the experiments.

The initial results are promising and extensive experiments with varying sequence lengths, time horizons, dimensions with a larger epochs having early stopping condition will give a clear idea of

[1] `https://github.com/ftazad/FMA_AI_Innovation_Challenge_Summer_2025`

the performance. The authors of this work [1] demonstrate that intra-variate dependencies are the primary contributors to prediction performance on benchmarks, while inter-variate dependencies have a minor impact. This can be taken into account to investigate how it affects the results during the validation phase of the experiments.

References

[1] Yu Chen, Nathalia Céspedes, and Payam Barnaghi. A closer look at transformers for time series forecasting: Understanding why they work and where they struggle. In *Forty-second International Conference on Machine Learning*.