

Table 1: Univariate short-term forecasting results in the M4 dataset. All forecasting horizons are in [6, 48]. The best results are in **bold** and the second best are underlined.

| Models    | CATS  | TimeMixer     | Timesnet      | PatchTST | MICN   | FiLM   | DLinear | Autoformer | Informer |
|-----------|-------|---------------|---------------|----------|--------|--------|---------|------------|----------|
| Quarterly | SMAPE | <b>9.979</b>  | <u>9.996</u>  | 10.100   | 10.644 | 15.214 | 12.925  | 12.145     | 11.360   |
|           | MASE  | <b>1.164</b>  | <u>1.166</u>  | 1.182    | 1.278  | 1.963  | 1.664   | 1.520      | 1.401    |
|           | OWA   | <u>0.878</u>  | <b>0.825</b>  | 0.890    | 0.949  | 1.407  | 1.193   | 1.106      | 1.027    |
| Monthly   | SMAPE | <b>12.557</b> | <u>12.605</u> | 12.670   | 13.399 | 16.943 | 15.407  | 13.514     | 14.062   |
|           | MASE  | <b>0.916</b>  | <u>0.919</u>  | 0.933    | 1.031  | 1.442  | 1.298   | 1.037      | 1.141    |
|           | OWA   | <b>0.866</b>  | <u>0.869</u>  | 0.878    | 0.949  | 1.265  | 1.144   | 0.956      | 1.024    |
| Yearly    | SMAPE | <u>13.263</u> | <b>13.206</b> | 13.387   | 16.463 | 25.022 | 17.431  | 16.965     | 14.727   |
|           | MASE  | <u>2.967</u>  | <b>2.916</b>  | 2.996    | 3.967  | 7.162  | 4.043   | 4.283      | 3.418    |
|           | OWA   | <u>0.779</u>  | <b>0.776</b>  | 0.786    | 1.003  | 1.667  | 1.042   | 1.058      | 0.881    |
| Others    | SMAPE | <b>4.560</b>  | <u>4.564</u>  | 4.891    | 6.558  | 41.985 | 7.134   | 6.709      | 24.460   |
|           | MASE  | <b>3.107</b>  | <u>3.115</u>  | 3.302    | 4.511  | 62.734 | 5.09    | 4.953      | 20.960   |
|           | OWA   | <b>0.970</b>  | <u>0.982</u>  | 1.035    | 1.401  | 14.313 | 1.553   | 1.487      | 5.879    |
| Average   | SMAPE | <b>11.701</b> | <u>11.723</u> | 11.829   | 13.152 | 19.638 | 14.863  | 13.639     | 14.086   |
|           | MASE  | <b>1.557</b>  | <u>1.559</u>  | 1.585    | 1.945  | 5.947  | 2.207   | 2.095      | 1.771    |
|           | OWA   | <b>0.838</b>  | <u>0.840</u>  | 0.851    | 0.998  | 2.279  | 1.125   | 1.051      | 1.230    |

Table 2: Detailed descriptions and experimental settings of M4 Dataset.

| Dataset      | Forecasting Horizon | Dataset Size      | Layers | Embedding Size | Batch Size | Epoch | Patience | Learning Rate |
|--------------|---------------------|-------------------|--------|----------------|------------|-------|----------|---------------|
| M4-Quarterly | 8                   | (24000, 0, 24000) | 3      | 64             | 32         | 30    | 10       | $10^{-3}$     |
| M4-Monthly   | 18                  | (48000, 0, 48000) | 3      | 64             | 32         | 30    | 10       | $10^{-3}$     |
| M4-Yearly    | 6                   | (23000, 0, 23000) | 3      | 64             | 32         | 30    | 10       | $10^{-3}$     |
| M4-Weekly    | 13                  | (359, 0, 359)     | 3      | 64             | 32         | 30    | 10       | $10^{-3}$     |
| M4-Daily     | 14                  | (4227, 0, 4227)   | 3      | 64             | 32         | 30    | 10       | $10^{-3}$     |
| M4-Hourly    | 48                  | (414, 0, 414)     | 3      | 128            | 32         | 30    | 10       | $10^{-3}$     |

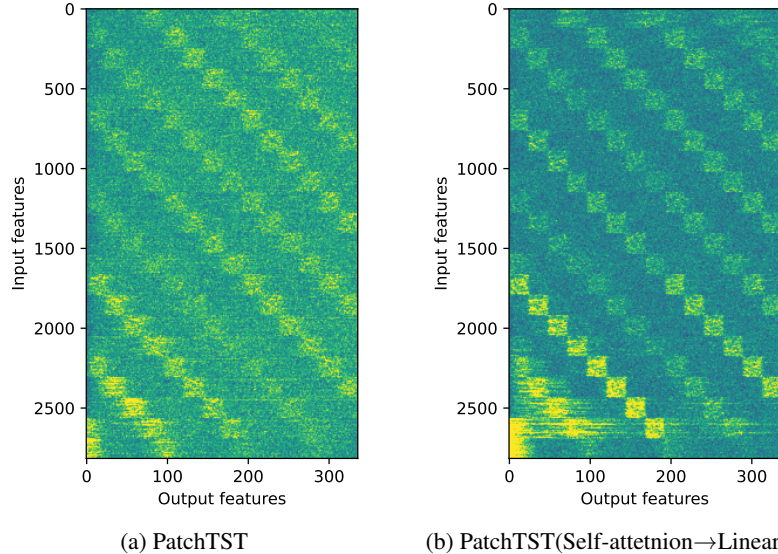


Figure 1: Absolute values of the weights in the final linear layer. PatchTST(Self-attention→Linear) appears to capture temporal information more effectively. The temporal information loss caused by self-attention results in the final linear weights of the original PatchTST model appearing as a blurred version of the PatchTST(Self-attention→Linear)’s linear weights. This demonstrates the potential drawback of self-attention beyond the complexity problem.