Transformer with Sparse Attention Mechanism for Industrial Time Series Forecasting

Yang Li\textsuperscript{1*}, Chunhua Tian\textsuperscript{2*}, Yuyan Lan\textsuperscript{1}, Chentao Yu\textsuperscript{3} and Keqiang Xie\textsuperscript{4}

\textsuperscript{1} School of Microelectronics, University of Chinese Academy of Sciences, Beijing 100190, China
\textsuperscript{2} Beijing Innovation Center for Industrial Big Data, Beijing 100083, China
\textsuperscript{3} Lenovo Company, Beijing 100085, China
\textsuperscript{4} China Electronic Product Reliability and Environmental Testing Research Institute, Guangzhou 511370, China
Email: \textsuperscript{4}tch04@hotmail.com; \textsuperscript{*}These authors contributed equally to this work.

Abstract. Long sequence time-series forecasting has become a very important task in industrial data analysis. Recently, the Transformer model has been widely used in sequence processing tasks. However, because industrial time series data are generally long and mixed with abnormal data, conventional Transformer model may extract irrelevant information in the context, resulting in poor forecasting. In this paper, we present Transformer with a Sparse Attention Mechanism (SAM) which can ensure local context be better integrated into attention mechanism. Inspired by the gating mechanism of LSTM, the most interesting part of sequence information are retained and the rest of the unimportant information are filtered. More attention can be focused on the factors that contribute most to the forecasting value of the sequence through this method. This method can efficiently capture long-range dependency between output and input. Furthermore, we leverage STL (Seasonal and Trend decomposition using Loess) model and IQR (Interquartile Range) method to address the outlier data. By applying this model to real-world datasets, our method achieves significant performance improvements over other methods.

Keywords. Time-series; transformer; sparse attention mechanism.

1. Introduction

Time series forecasting is widely used in industrial fields, such as wind energy forecasting and electricity demand forecasting. This task leverages past behaviour to make a prediction. In addition, the industrial time series data are generally long and there are abnormal conditions, which bring some challenges of forecasting.

Recently, attention mechanism and Transformer [1] framework of BERT [2] have shown excellent performance in the task of NLP (natural language processing). As a sequence generation task, time series forecasting task can also use the Transformer and achieve better results than Long Short-Term Memory [3]. In each time series window, however, the conventional Transformer assigns credit to all components of the sequence when it is applied directly, causing a lack of focus. In addition, the strong correlation information between windows may be weakened, which makes it infeasible to process long sequences.

To solve this problem, retaining the most interesting part of sequence information is a new perspective. Therefore, a novel Transformer with a Sparse Attention Mechanism (SAM) is proposed...
as shown in figure 1, which can be used to process long sequence. It is based on Informer [4] model from AAAI 2021 best paper. When each sequential window processed by attention mechanism arrives, the SAM pays attention to top-\(N\) positions in the sequence. The advantages of the SAM are as follows: When model meets a trend, it will be easier to move towards it; When the peak value of the sequence comes, it will be easier to gradually attenuate; When the low point of the sequence comes, it will be easier to gradually rise. Thus, the SAM can perform more concentrated attention than traditional transformer. In addition, due to the existence of abnormal working condition data in industrial data, we leverage STL (Seasonal and Trend decomposition using Loess) [6] model and IQR (Interquartile Range) method to tackle the outlier data. In conclusion, main contributions of this paper are as follows:

- We present Transformer with Sparse Attention Mechanism for time series forecasting, which can ensure local context be better integrated into attention mechanism.
- We combine STL (Seasonal and Trend decomposition using Loess) model and IQR (Interquartile Range) method to eliminate abnormal data.
- We verify the new model on three Industrial data, the results show that our method can achieve more accurate forecasting results.

2. Related Works
With the increasing application of forecasting, many methods have been put forward. These methods can be divided into classical methods and deep learning methods.

The classical methods generally use statistical learning theory. Seeger et al. [7] propose a method based on Bayesian that can focus more attention mechanism on target statistics. Box et al. [8] introduce the establishment and prediction of stochastic model in detail. However, due to the limited scalability of these methods, they are not suitable for large-scale forecasting tasks.

The deep learning methods generally adopt encoder-decoder frameworks to achieve accurate prediction. In recent years, these methods mainly use RNN and its variants for sequence prediction [9], and achieve some good performance. But with the emergence of Transformer based on self-attention proposed by Vaswani et al. [1], the task of sequence generation has achieved great success. After that, some works gradually apply Transformer model to long time series forecasting. Last year, Zhou et al. [5] present an Informer model which sharply reduces the total space complexity. However, because the industrial time series data are generally long, the conventional Transformer model may extract
irrelevant information in the context resulting in poor forecasting. Therefore, we present a novel Transformer model to deal with this issue.

3. Method
In this section, we first give the definition of time-series forecasting and formulate it as a supervised machine learning task. Then, we introduce our new Transformer model in detail.

3.1. Problem Description
Consistent with the baseline model, the size of each sliding window is fixed. The input time-series data of time $t$ can be easily obtained as follows: $X_t = \{x'_1, ..., x'_d \mid x'_i \in \mathbb{R}^d\}$. The corresponding prediction results of the input can be expressed as follows: $Y_t = \{y'_1, ..., y'_d \mid y'_i \in \mathbb{R}^d\}$. Each data point in the above formula represents a vector containing features.

3.2. Transformer Model with SAM
The conventional Transformer model may extract irrelevant information in the context resulting in poor forecasting because the industrial time-series data are too long. In order to solve it, we present Transformer with a Sparse Attention Mechanism, which makes attention only focus on top-$N$ elements. This new model can ensure local context be better integrated into the attention mechanism. Compared with general attention mechanism, this method does not need to integrate some irrelevant information into the model calculation. The model framework we proposed is shown as figure 2.

![Figure 2](image-url)

**Figure 2.** Method framework diagram. The data processed by Query matrix and Key matrix is fed into the SAM. Then the SAM computes the Score matrix and the Mask matrix by top-$N$ selection. Finally, the SAM computes a sequence of vector outputs.

Our model is mainly based on the Informer framework. In the self-attention layer, the input of each layer has three distinct components according to the Transformer architecture: query matrices $Q = W_q z$, key matrices $K = W_k z$, and value matrices $V = W_v z$. Then we can easily get the attention scores as the score matrix $s = QK^T d^{-1/2}$.

At this time, we should choose the attention scores with larger demonstrate higher relevance. In our model, the $N$-th score of each row $i$ in $s$ is chosen as the threshold $\hat{S}_i$. As a result, we can get a total of $lq$ values. Then we compare the value of each row in $s$ with the corresponding $S$ value of the row, and set all elements less than $\hat{S}$ value to 0. The results of attention selection can be obtained by multiplying the obtained matrix $M$ with the previous $s$ matrix. After that, we can get the final output representation.
of self-attention $O$ according to the following steps of Transformer $O = \text{soft} \max(s \cdot M)V$. Consistent with the conventional transformer model, we set all the 0 elements to $-\infty$ for later processing when we filter the attention.

Afterwards, all outputs are concatenated and linearly projected again. Through such top-$N$ attention method, our model Transformer with SAM can get more focused attention to ensure the preservation of important components.

4. Experiments

In this section, we will briefly introduce the datasets and evaluation metrics used in our experiment. Then we will give the comparison results and analysis with other methods.

4.1. Real-World Datasets and Evaluation Metrics

In our experiment, we used three real-world datasets: The Electricity Transformer Temperature datasets named ETTm1 (for 15-minute-level), ETTh1 (for 1-hour-level) and ETTh2 (for 1-hour-level) are collected by the Informer [5]. We also divide each dataset into three parts: 60% training set, 20% verification set, and 20% testing set.

In this paper, two regression algorithm evaluation metrics are used to compare the results, including MSE and MAE. The MSE metric represents the Mean Square Error, and the formula is

$$MSE = \frac{1}{L} \sum_{i=1}^{L} (y_i - \hat{y}_i)^2.$$ And the MAE metric represents the Mean Absolute Error, and the formula is

$$MAE = \frac{1}{L} \sum_{i=1}^{L} \lvert y_i - \hat{y}_i \rvert.$$  

4.2. Implementation Details

In the data processing stage, we use STL method to decompose the time-series data and get the residual part. Then the IQR method is used to screen the outliers, and the ratio of outliers is 5%. During the model training, we use the ASGD (Adam stochastic gradient descent) with $\alpha = 0.9$ and $\beta = 0.999$. The learning rate starts from 5e-4 and decays twice every epoch. The batch size is 32 and the comparison methods are set as recommended.

Before the comparison of experiments, we should first determine the optimal $N$. We compare the effects of different $N$ with our model on ETTh1 dataset as shown in table 1. We prolong the prediction horizon (we set $L$ to 24 steps, 168 steps and 720 steps) as recommended to achieve a higher prediction performance. The best results have been shown in bold. The results show that setting the value of $N$ to 8 can achieve better results for $N$ in (2, 4, 8, 16).

| $N$   | 2   | 4   | 8   | 16  |
|-------|-----|-----|-----|-----|
| Metric | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE | MSE | MAE |
| ETTh1 | 24  | 0.157 | 0.276 | 0.128 | 0.247 | 0.088 | 0.215 | 0.095 | 0.237 |
|       | 168 | 0.235 | 0.399 | 0.204 | 0.362 | 0.172 | 0.320 | 0.181 | 0.341 |
|       | 720 | 0.632 | 0.717 | 0.297 | 0.481 | 0.233 | 0.398 | 0.262 | 0.421 |

4.3. Results of Experiments

In this part, we show the results of our method and other methods on the above three real-world datasets. As for our proposed Transformer model with a Sparse Attention Mechanism (hereinafter referred to as SAM), it is listed separately as a novel model. We also use several advanced models (hereinafter referred to as Informer, LogTrans [10], Reformer [11], DeepAR [12], ARIMA [13][3] for comparison. Each method is predicted in the form of single variable series. We prolong the prediction...
horizon (24 steps, 168 steps and 720 steps) as recommended to achieve a higher prediction performance. The best results have been shown in bold.

Table 2. Univariate forecasting results of three datasets.

| Methods | Metric | SAM | Informer | LogTrans | Reformer | DeepAR | ARIMA |
|---------|--------|-----|----------|----------|----------|--------|-------|
| ETTh1   |        |     |          |          |          |        |       |
| 24      | MSE    | 0.088 | 0.215    | 0.091    | 0.246    | 0.103  | 0.259 |
|         | MAE    | 0.246 | 0.389    | 0.355    | 0.270    | 0.375  | 0.359 |
| 168     | MSE    | 0.176 | 0.320    | 0.187    | 0.355    | 0.207  | 0.375 |
|         | MAE    | 0.273 | 0.437    | 0.301    | 0.270    | 0.375  | 0.570 |
| 720     | MSE    | 0.233 | 0.398    | 0.257    | 0.421    | 0.273  | 0.463 |
|         | MAE    | 0.273 | 0.463    | 0.298    | 0.421    | 0.273  | 0.463 |
| ETTh2   |        |     |          |          |          |        |       |
| 24      | MSE    | 0.090 | 0.235    | 0.099    | 0.241    | 0.102  | 0.255 |
|         | MAE    | 0.102 | 0.437    | 0.304    | 0.385    | 0.246  | 0.422 |
| 168     | MSE    | 0.229 | 0.382    | 0.235    | 0.390    | 0.246  | 0.422 |
|         | MAE    | 0.246 | 0.570    | 0.298    | 0.422    | 0.246  | 0.422 |
| 720     | MSE    | 0.287 | 0.440    | 0.442    | 0.442    | 0.303  | 0.493 |
|         | MAE    | 0.298 | 1.044    | 1.044    | 1.044    | 0.493  | 1.044 |
| ETTm1   |        |     |          |          |          |        |       |
| 24      | MSE    | 0.028 | 0.152    | 0.034    | 0.160    | 0.066  | 0.202 |
|         | MAE    | 0.073 | 0.228    | 0.091    | 0.160    | 0.066  | 0.202 |
| 168     | MSE    | 0.180 | 0.382    | 0.187    | 0.384    | 0.199  | 0.386 |
|         | MAE    | 0.246 | 0.570    | 0.298    | 0.570    | 0.246  | 0.570 |
| 720     | MSE    | 0.486 | 0.639    | 0.519    | 0.665    | 0.598  | 0.702 |
|         | MAE    | 0.598 | 2.112    | 1.528    | 2.112    | 1.528  | 2.112 |

4.4. Results Analysis

The results of different methods are shown in table 2. Obviously, the scores of MSE and MAE of our model are decreased. It shows that our proposed model significantly improves the inference performance. The reason is that our model can enhance the concentration of the Transformers attention through top-N selection. Other methods will gradually produce some irrelevant information in the context resulting in poor forecasting. In addition, although the baseline model “Informer” performs better than other methods, our model can still get better results than it. This is also a better indication that the SAM model can indeed produce more accurate predictions. There is also a phenomenon in table 2: Our model “SAM” does not improve the forecasting results much compared with the baseline model “Informer” in the ETTh2 dataset. The reason is that the period of the data points in the fixed size window is short and violent, and there is no significant difference in each attention score, which leads to the failure of our method to play the most important role. This is also the direction we will continue to study in the future. And the forecasting results about part of the three datasets are shown as figure 3.

Figure 3. Qualitative results of SAM. (a), (b) and (c) are the results of datasets ETTh1, ETTh2 and ETTm1, respectively.

5. Conclusions

In this paper, we propose a novel transformer model with sparse attention mechanism for time series forecasting task. This new model can make the attention in conventional transformer pay more attention to the components which contribute the most. In addition, we use STL (Seasonal and Trend decomposition using Loess) model and IQR (Interquartile Range) method to tackle the outlier data. We generate a more accurate forecasting result. However, our model will slightly lose some details in the final result. We will adopt the way of multi-layer information fusion to add some details in the future. In conclusion, the SAM can provide a new way to improve the forecasting effect of time series data.
Acknowledgments
This work is partially supported by National Key R&D Program of China (2018YFB1700605) and Open Fund of MIIT Key Laboratory of Industrial Quality Big Data (2021 IEQBD-01).

References
[1] Vaswani A, Shazeer N, Parmar N, et al. 2017 Attention is all you need Advances in Neural Information Processing Systems 5998-6008.
[2] Devlin J, Chang M W, Lee K and Toutanova K. 2019 BERT: Pre-training of deep bidirectional transformers for language understanding NAACL HLT 1 4171-86.
[3] Zhu X, Fu B, Yang Y, et al. 2019 Attention-based recurrent neural network for influenza epidemic prediction BMC Bioinformatics 20 1-10.
[4] Zhou H, Zhang S, Peng J, Zhang S, Li J, Xiong H and Zhang W 2020 Informer: Beyond efficient transformer for long sequence time-series forecasting Preprint gr-qc/1207436.
[5] Beltagy I, Peters M E and Cohan A 2020 Longformer: The long-document transformer Preprint gr-qc/0405150.
[6] Cleveland R B, Cleveland W S, McRae J E, et al. 1990 STL: A seasonal-trend decomposition Journal of Official Statistics 6 2203-73.
[7] Seeger M, Salinas D and Flunkert V 2016 Bayesian intermittent demand forecasting for large inventories Adv. Neural Inf. Process. Syst. 4653-61.
[8] Geurts M, Box G E P and Jenkins G M 1977 Time series analysis: Forecasting and control J. Mark. Res. 14 269.
[9] Kondo K, Ishikawa A and Kimura M 2019 Sequence to sequence with attention for influenza prevalence prediction arXiv.
[10] Li S, Jin X, Xuan Y, et al. 2019 Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting NIPS 32 5243-53.
[11] Kitaev N, Kaiser Ł and Levskaya A 2020 Reformer: The efficient transformer Preprint gr-qc/0104451.
[12] Salinas D, Flunkert V, Gasthaus J and Januschowski T 2020 DeepAR: Probabilistic forecasting with autoregressive recurrent networks Int. J. Forecast. 36 1181-91.
[13] Adebiyi A A, Adewumi A O and Ayo C K 2014 Stock price prediction using the ARIMA model Proc. - UKSim-AMSS 16th Int. Conf. Comput. Model. Simulation pp 106-12.