Transformers for Time Series Forecasting

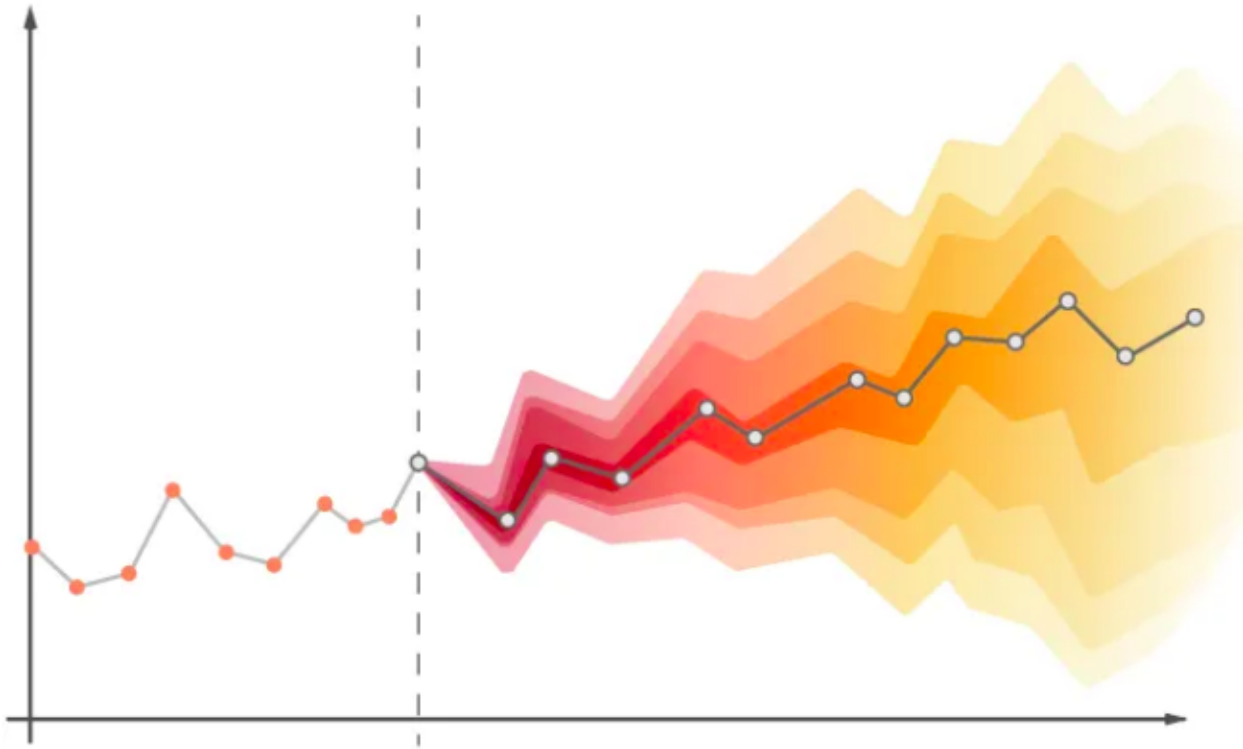
Marco Zanotti

University Milano-Bicocca

Contents

1. The TSF Problem

Time series forecasting (TSF) is the task of predicting future values of a given sequence based on previously observed values.



The TSF problem may be essentially identified by the following aspects:

- **Prediction objective**: point forecasting vs probabilistic forecasting
- **Forecast horizon**: short-term vs long-term forecasting
- **Input-Output dimension**: univariate vs multivariate forecasting
- **Forecasting task**: single-step vs multi-step forecasting

The TSF problem is usually faced with statistical models (ARIMA, ETS) or deep learning models (RNN, LSTM).

The main challenges of the TSF problem are:

- uncertainty increases as the forecast horizon increases
- difficulty in capturing multiple complex patterns over time
- difficulty in capturing long-term dependencies (critical for long-term forecasting)
- difficulty to handle long input sequences (critical for long-term forecasting)

2. Vanilla Transformer

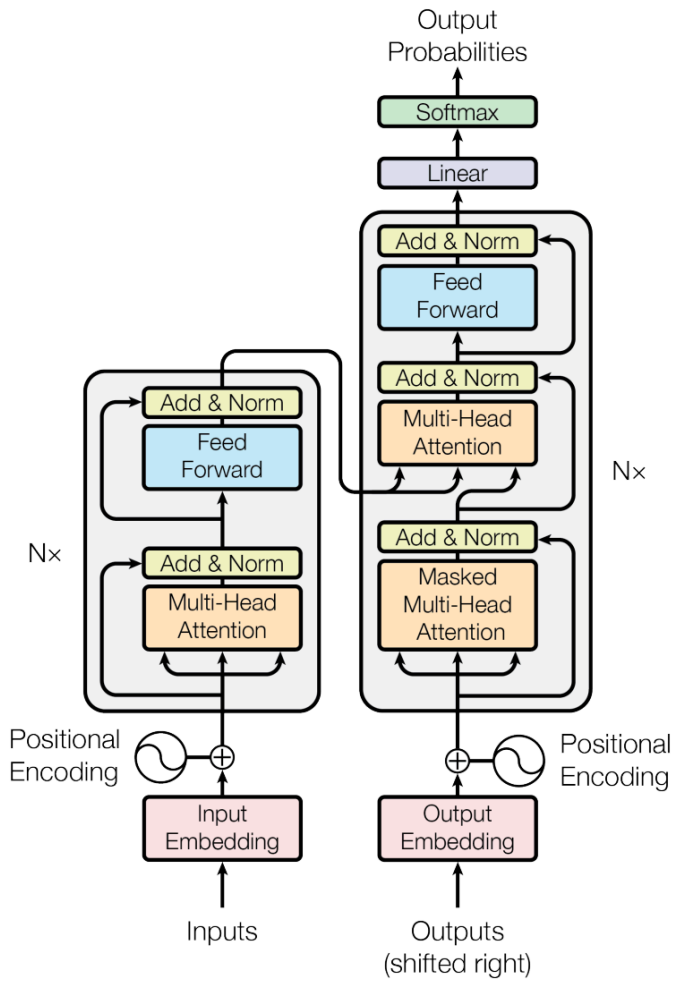
desc

The LSTM was seen to suffer from "short-term memory" over long sequences.

Transformers retain direct connections to all previous timestamps, allowing information to propagate over much longer sequences. However, this entails a new challenge: the model will be directly connected to an exploding amount of input. In order to filter the important from the unimportant, Transformers use an algorithm called self-attention.

a second challenge that needs to be addressed. The time series is not processed sequentially; thus, the Transformer will not inherently learn temporal dependencies. To combat this, the positional information for each token must be added to the input. In doing so, the self-attention block will have context of relative distance between a given time stamp and the current one, as an

Vanilla Transformer



aa

The residual connections are a commonly used technique for training deep neural networking systems and help train the model. It also helps it stabilize and learn.

Layer normalization is generally used in neural networks to process sequential data. It helps faster the convergence of training.

They have the powerful ability to learn long-range dependencies, which is crucial for accurate and true prediction.

They are difficult to train and demand time, consistency, and hardwork.

These models are often large and complex, which makes them tough to optimize.

Transformer models need data to learn the correlation between the input and output sequences.

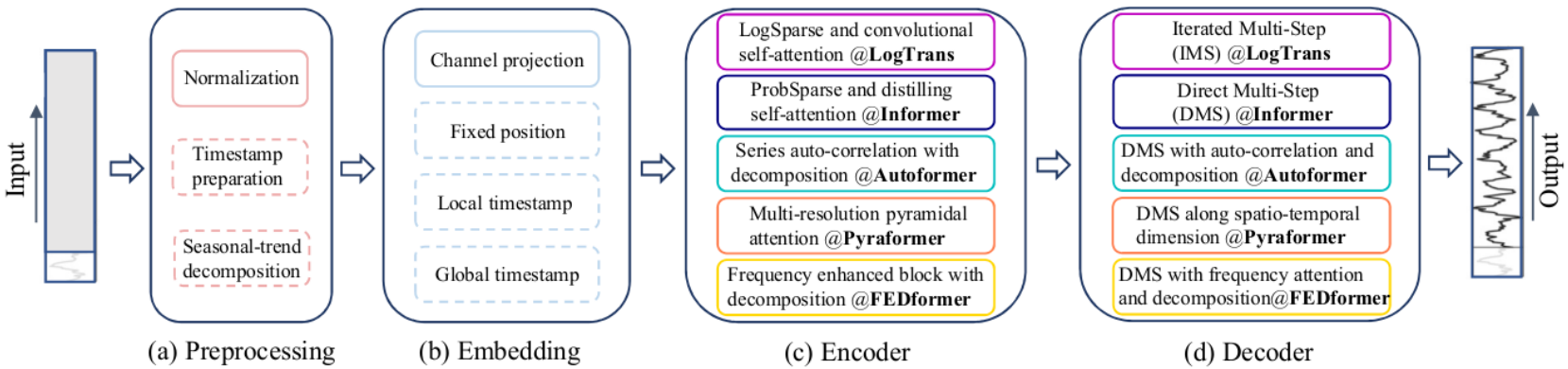
vanilla Transformers can be used for the TSF

Problems with vanilla Transformers

3. TSF Transformers

Classification

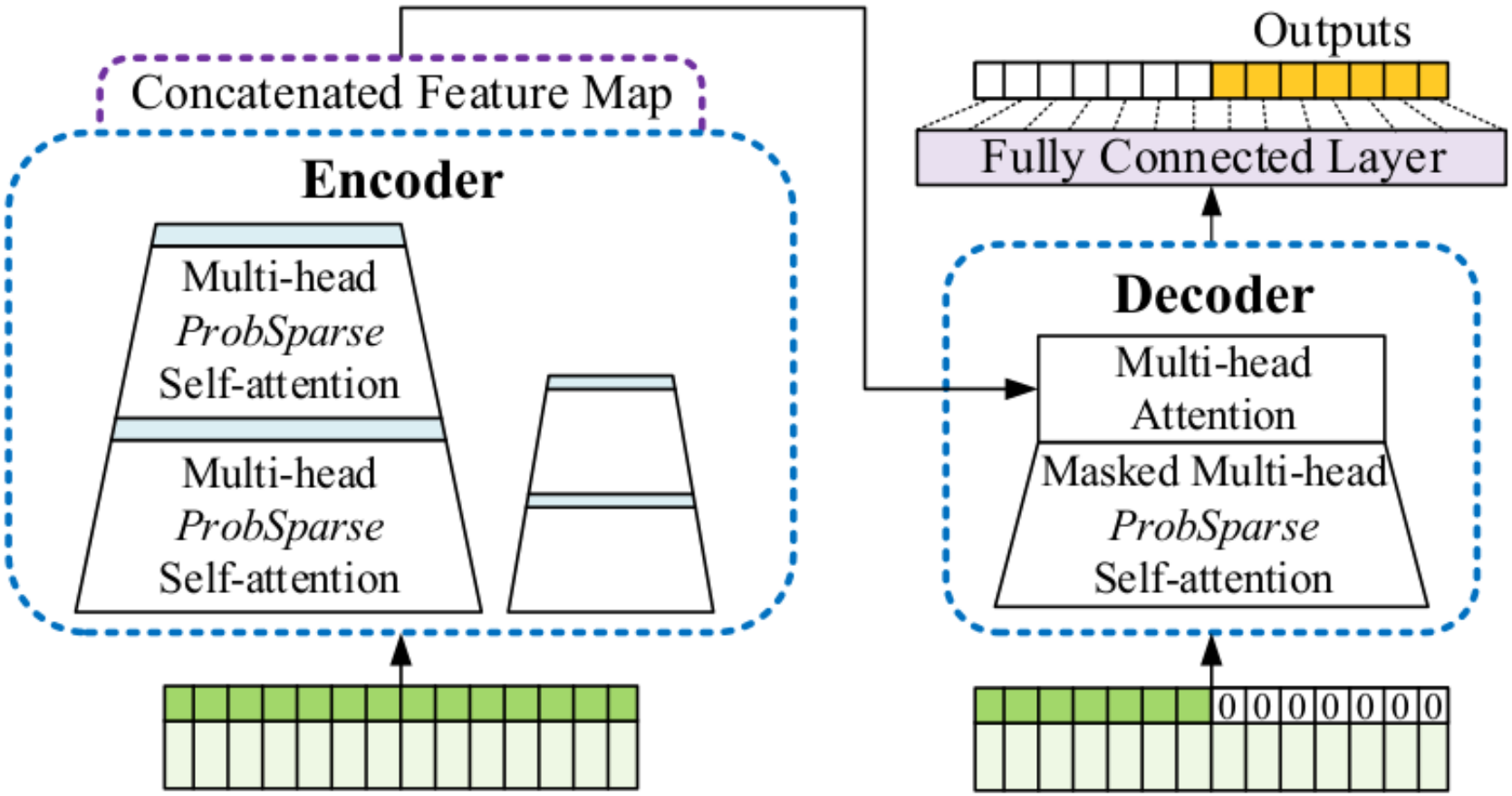
cambiarla con una più sensata con le modifiche su positional, architettura e moduli



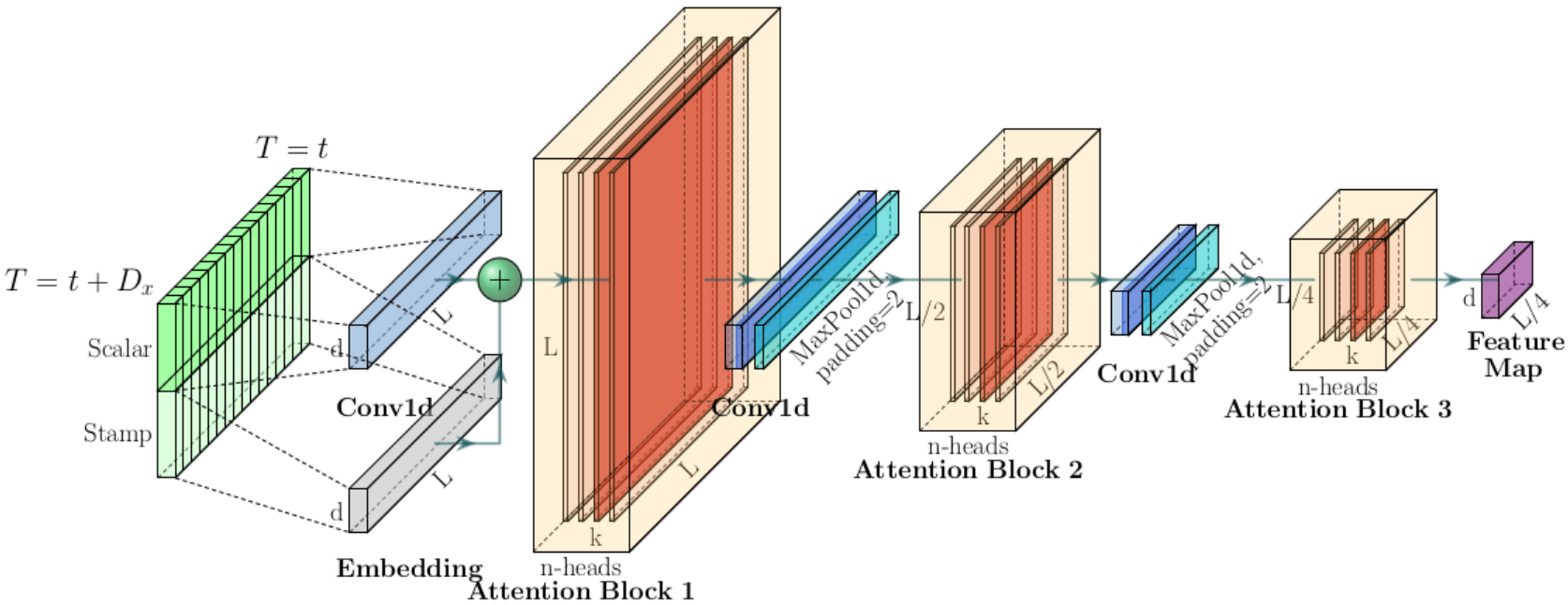
Informer

aa

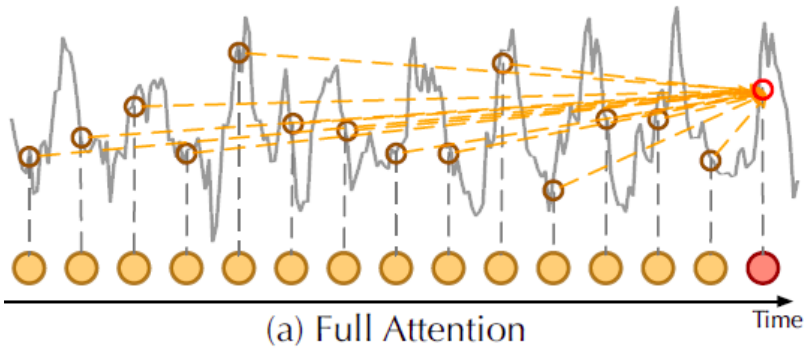
Informer - Architecture



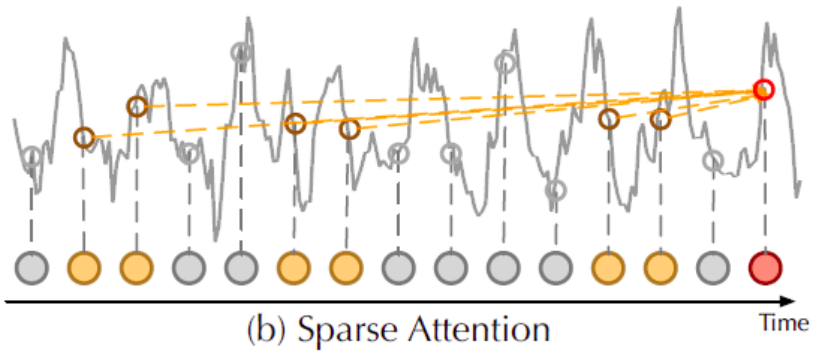
Informer - Causal Convolution Layers



Informer - ProbSparse Attention



(a) Full Attention



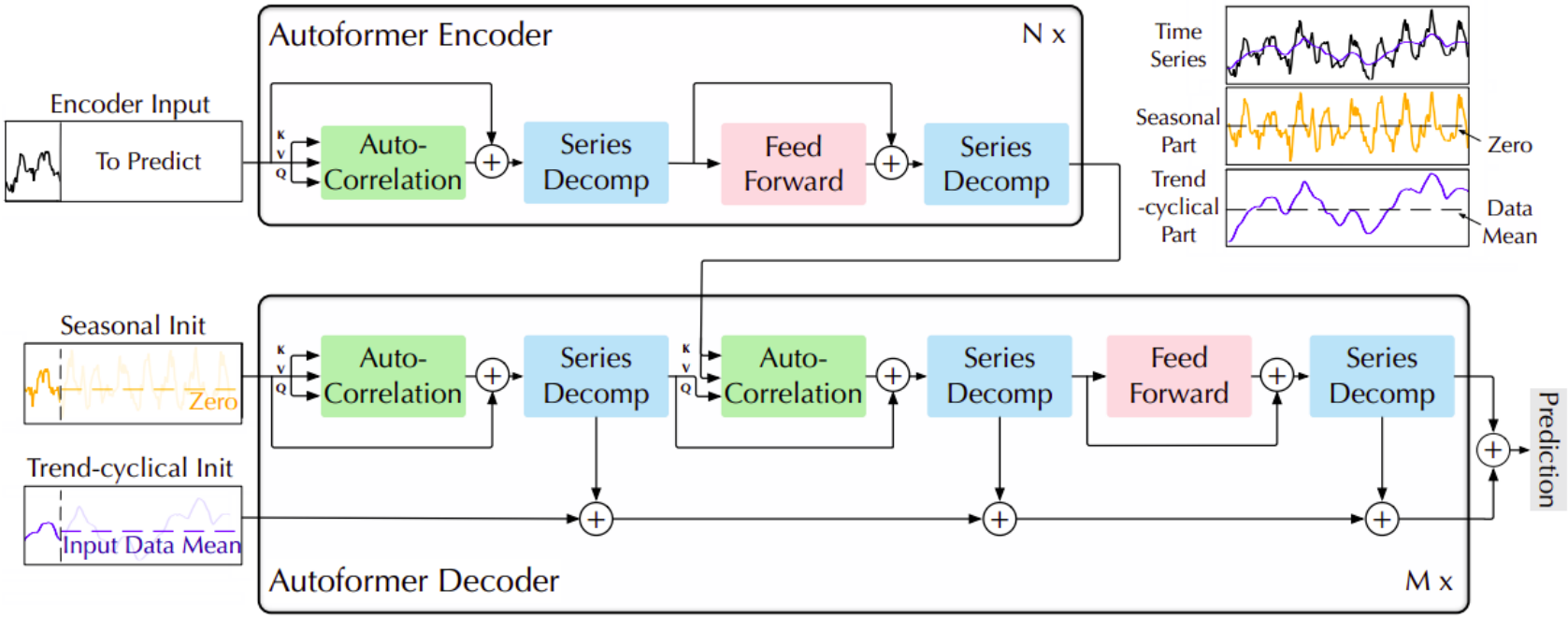
(b) Sparse Attention

Autoformer

Autoformer builds upon the traditional method of decomposing time series into seasonality and trend-cycle components. This is achieved through the incorporation of a Decomposition Layer, which enhances the model's ability to capture these components accurately. Moreover, Autoformer introduces an innovative auto-correlation mechanism that replaces the standard self-attention used in the vanilla transformer.

the two key contributions of Autoformer: the Decomposition Layer and the Attention (Autocorrelation) Mechanism.

Autoformer - Architecture



Autoformer - Decomposition Layer

Autoformer incorporates a decomposition block as an inner operation of the model, as presented in the Autoformer's architecture above. As can be seen, the encoder and decoder use a decomposition block to aggregate the trend-cyclical part and extract the seasonal part from the series progressively

For an input series X with length L, the decomposition layer returns Xtrend and Xseasonal

Autoformer - Decomposition Layer

aa

Autoformer - Attention Mechanism

In addition to the decomposition layer, Autoformer employs a novel auto-correlation mechanism which replaces the self-attention seamlessly. In the vanilla Time Series Transformer, attention weights are computed in the time domain and point-wise aggregated. On the other hand, as can be seen in the figure above, Autoformer computes them in the frequency domain (using fast fourier transform) and aggregates them by time delay.

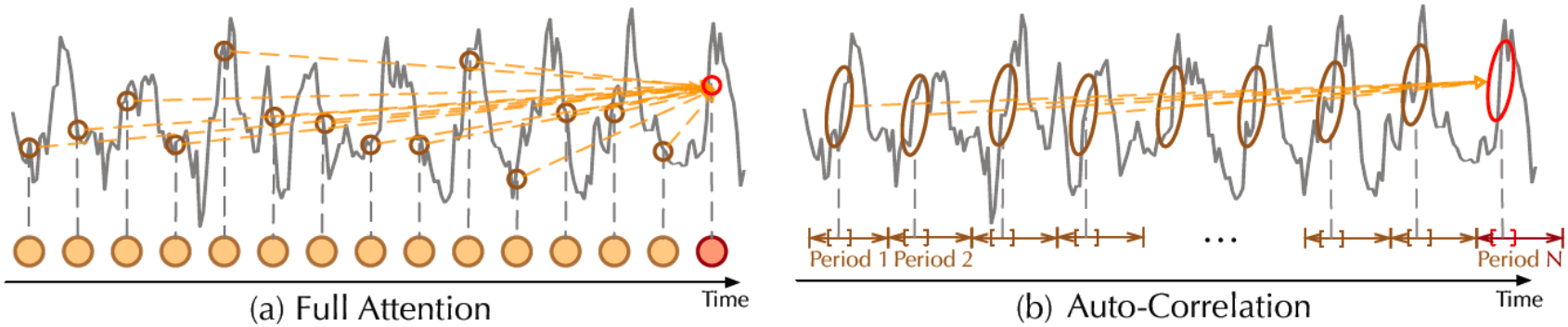
Autoformer - Attention Mechanism

Using autocorrelation, Autoformer extracts frequency-based dependencies from the queries and keys, instead of the standard dot-product between them. You can think about it as a replacement for the QK^T term in the self-attention.

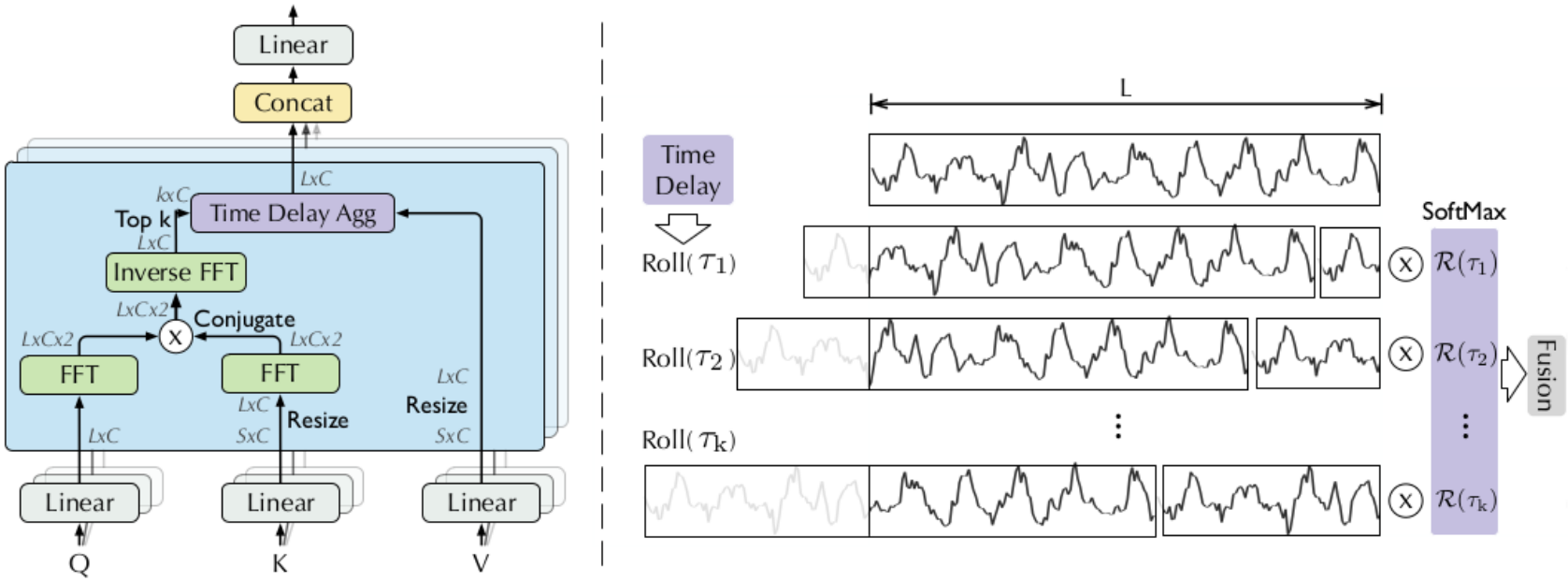
In practice, autocorrelation of the queries and keys for all lags is calculated at once by FFT. By doing so, the autocorrelation mechanism achieves O(LlogL) time complexity (where L is the input time length), similar to Informer's ProbSparse attention.

Next, we will see how to aggregate our attn_weights with the values by time delay, process which is termed as Time Delay Aggregation.

Autoformer - Autocorrelation Attention



Autoformer - Time Delay Aggregation



4. Conclusions

Conclusions

- aa

Bibliografy

Ailing Z., et al., 2023, 'Are Transformers Effective for Time Series Forecasting?', AAAI

Haixu W., et al., 2021, 'Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting', NeurIPS

Haoyi Z., et al., 2021, 'Informer: Beyond efficient transformer for long sequence time-series forecasting', AAAI

Lara-Benitez P., et al., 2021, 'Evaluation of the Transformer Architecture for Univariate Time Series Forecasting', Advances in Artificial Intelligence, CAEPIA

Qingsong W., et al., 2022, 'Transformers in Time Series: A Survey', AAAI

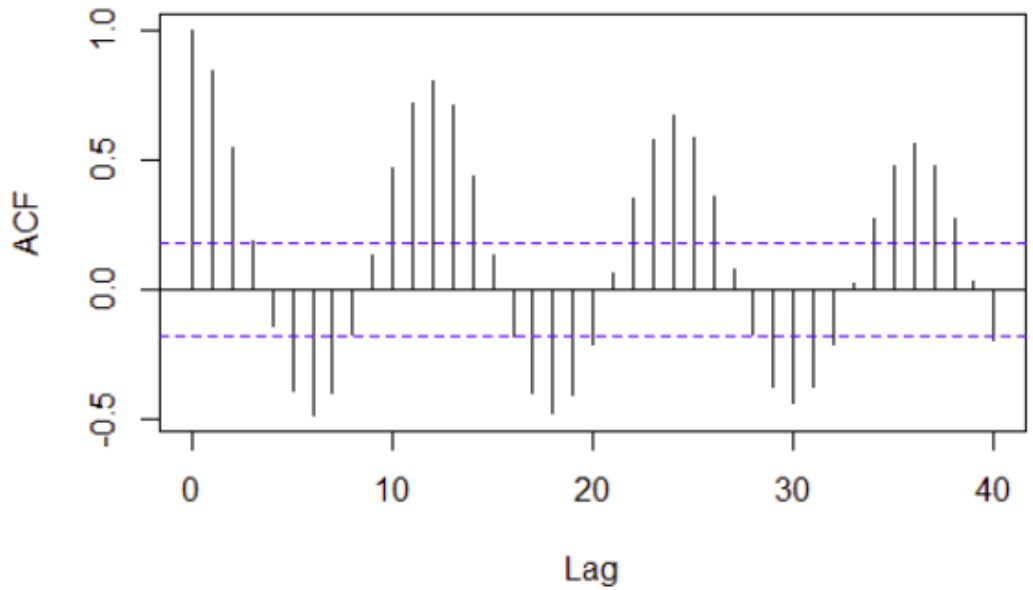
Thank you!

Appendix

Autocorrelation

In theory, given a time lag k, autocorrelation for a single discrete variable Y is used to measure the "relationship" (pearson correlation) between the variable's current value at time t to its past value at time $t - k$.

$$Autocorrelation(k) = Corr(Y_t, Y_{t-k})$$



Time Series Decomposition

In time series analysis, decomposition is a method of breaking down a time series into three systematic components: trend-cycle, seasonal variation, and random fluctuations.