Paper Review: Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting

Abdullah Mamun

May 10, 2021

About this paper

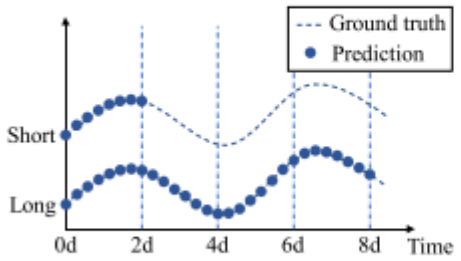
Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting

Haoyi Zhou, [1] **Shanghang Zhang,** [2] **Jieqi Peng,** [1] **Shuai Zhang,** [1] **Jianxin Li,** [1]
Hui Xiong, [3] **Wancai Zhang** [4]

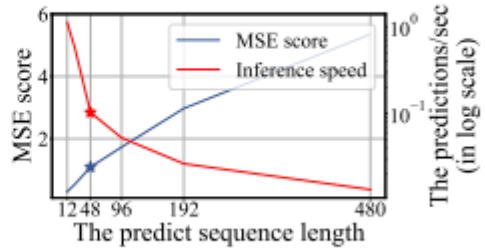
[1] Beihang University [2] UC Berkeley [3] Rutgers University [4] SEDD Company
{zhouhy, pengjq, zhangs, lijx}@act.buaa.edu.cn, shz@eecs.berkeley.edu, {xionghui,zhangwancaibuaa}@gmail.com

Abstract

Many real-world applications require the prediction of long sequence time-series, such as electricity consumption planning. Long sequence time-series forecasting (LSTF) demands a high prediction capacity of the model, which is the ability to capture precise long-range dependency coupling between output and input efficiently. Recent studies have shown the



(a) Sequence Forecasting.



(b) Run LSTM on sequences.

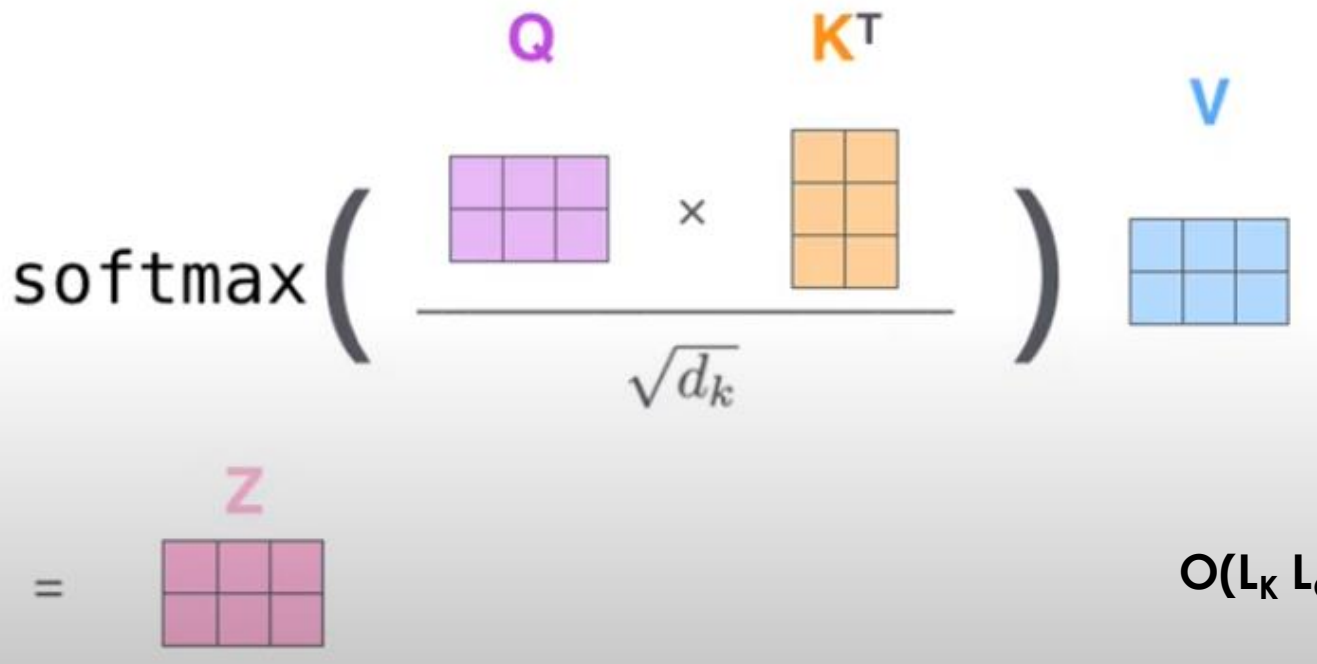
About this paper

- Published in the **35th AAAI** conference in 2021.
- Cited by **7** as of May 10, 2021
- Received the best paper award in 35th AAAI 2021.
- Main goal is building a **forecasting model** for multi-dimensional time series data.

Why Informer?

Transformer cell:

High 1. Space-2. Time complexity

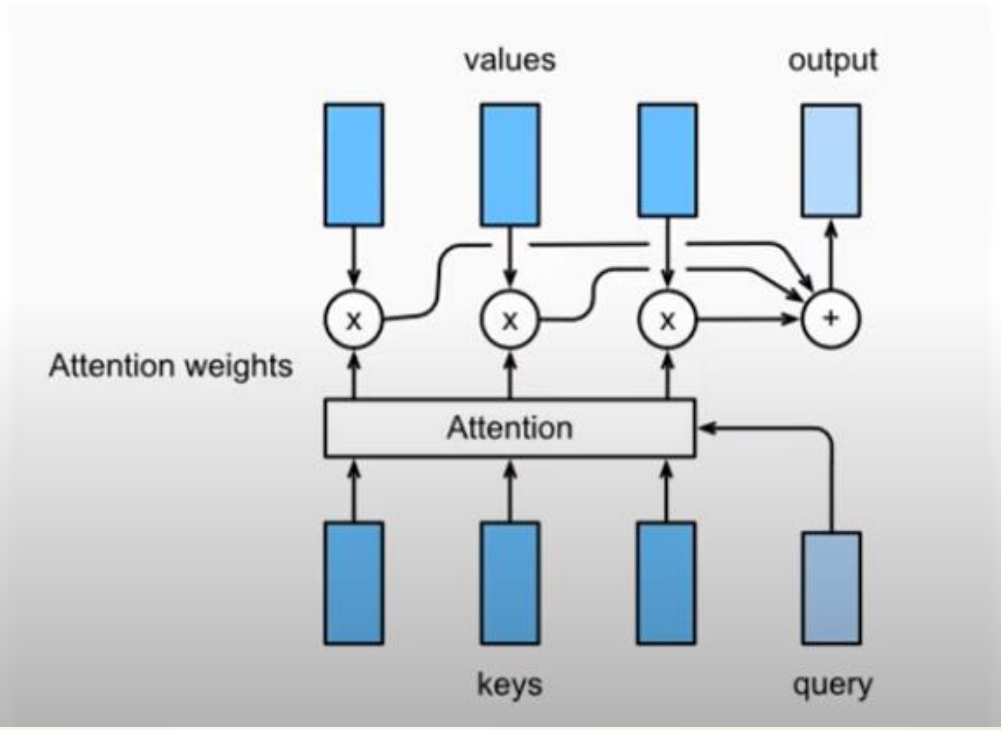


$O(L_K L_Q)$

Why Informer?

Attention mechanism:

3. Predicts one output at a time



Why Informer?

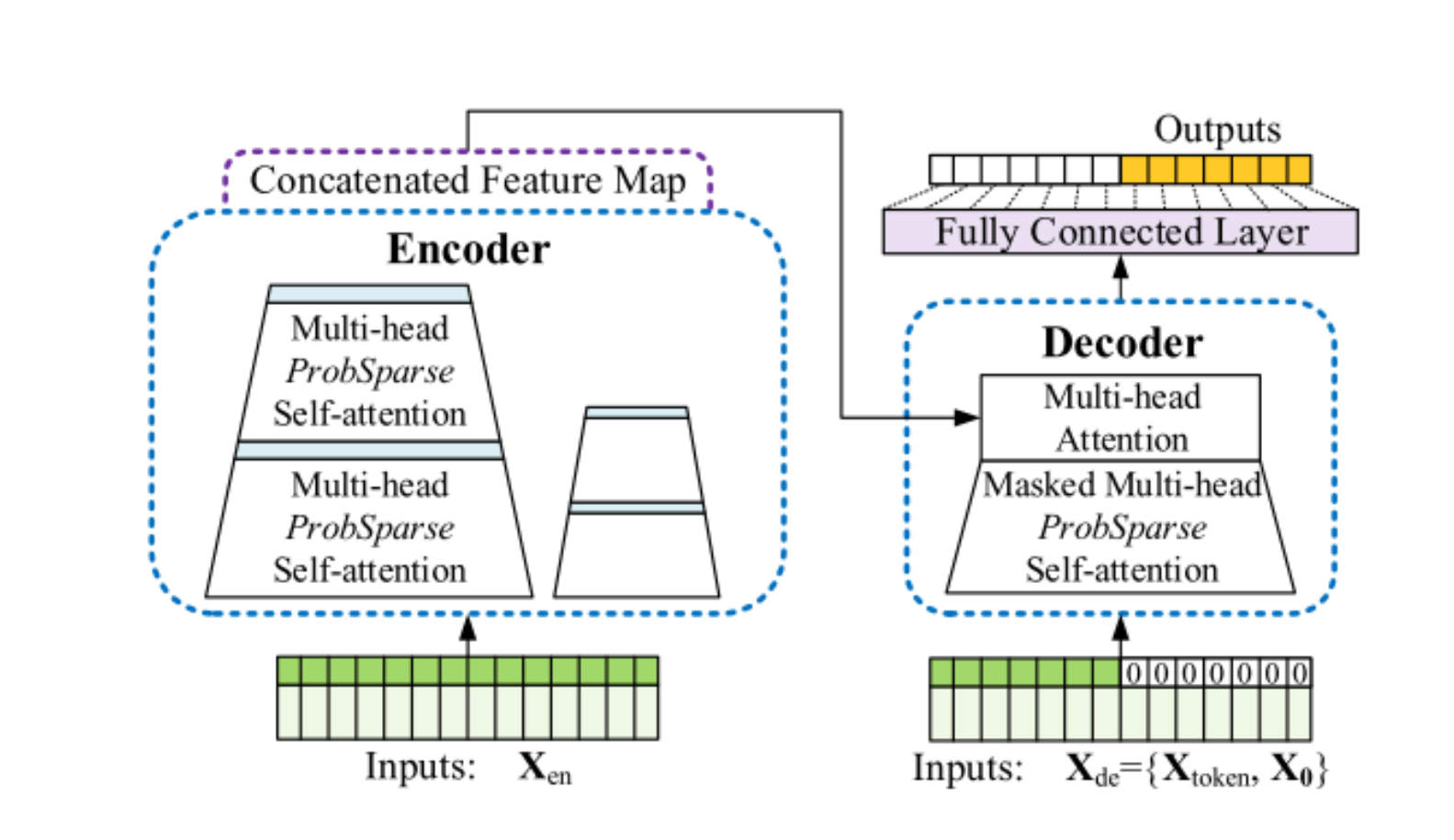
Informer:

1. **Query sparcity** Measurement lowers **Space** complexity

2. **ProbSparse** lowers **Time** Complexity

3. Predicts **sequence** in one batch (Generative Style Decoder)

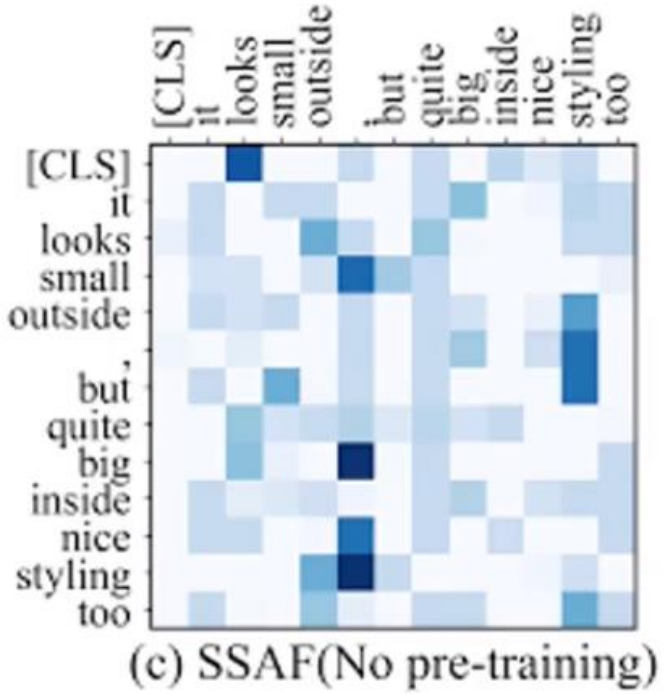
Informer Model Overview



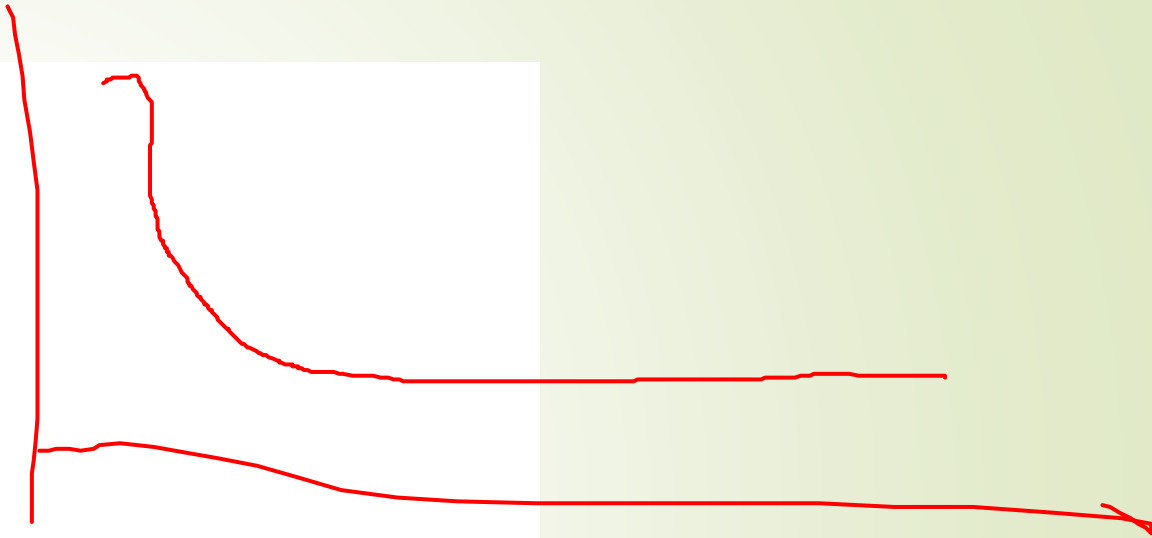
Main Improvements

1

Attention values:



(c) SSAF(No pre-training)



long-tail distributed

Therefore:

Prune needed.

Main Improvements

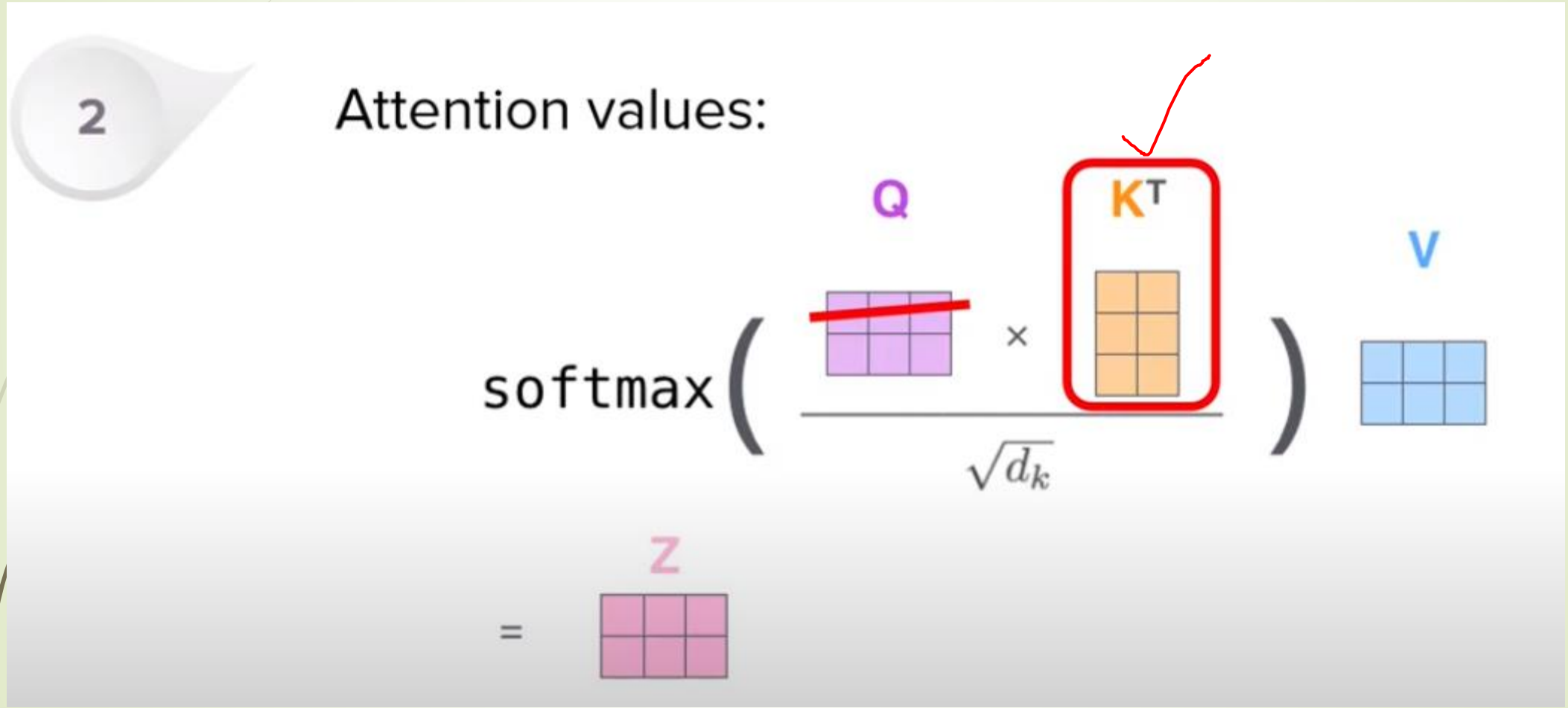
$$\mathcal{A}(\mathbf{q}_i, \mathbf{K}, \mathbf{V}) = \sum_j \frac{k(\mathbf{q}_i, \mathbf{k}_j)}{\sum_l k(\mathbf{q}_i, \mathbf{k}_l)} \mathbf{v}_j = \mathbb{E}_{p(\mathbf{k}_j|\mathbf{q}_i)}[\mathbf{v}_j]$$

1

Therefore: thresholding the queries, so that **queries** inducing top-u KLD-values are kept (rests set to zero)

u =c ln L_Q (not proven!)

Main Improvements



Main Improvements

However, the traversing of all the queries for the measurement $M(\mathbf{q}_i, \mathbf{K})$ requires calculating each dot-product pairs, i.e., quadratically $\mathcal{O}(L_Q L_K)$, besides the LSE operation has the potential numerical stability issue. Motivated by this, we propose an empirical approximation for the efficient acquisition of the query sparsity measurement.

Lemma 1. *For each query* $\mathbf{q}_i \in \mathbb{R}^d$ *and* $\mathbf{k}_j \in \mathbb{R}^d$ *in the keys set* $\mathbf{K}$*, we have the bound as* $\ln L_K \leq M(\mathbf{q}_i, \mathbf{K}) \leq \max_j\{\mathbf{q}_i\mathbf{k}_j^\top/\sqrt{d}\} - \frac{1}{L_K}\sum_{j=1}^{L_K}\{\mathbf{q}_i\mathbf{k}_j^\top/\sqrt{d}\} + \ln L_K$*. When* $\mathbf{q}_i \in \mathbf{K}$*, it also holds.*

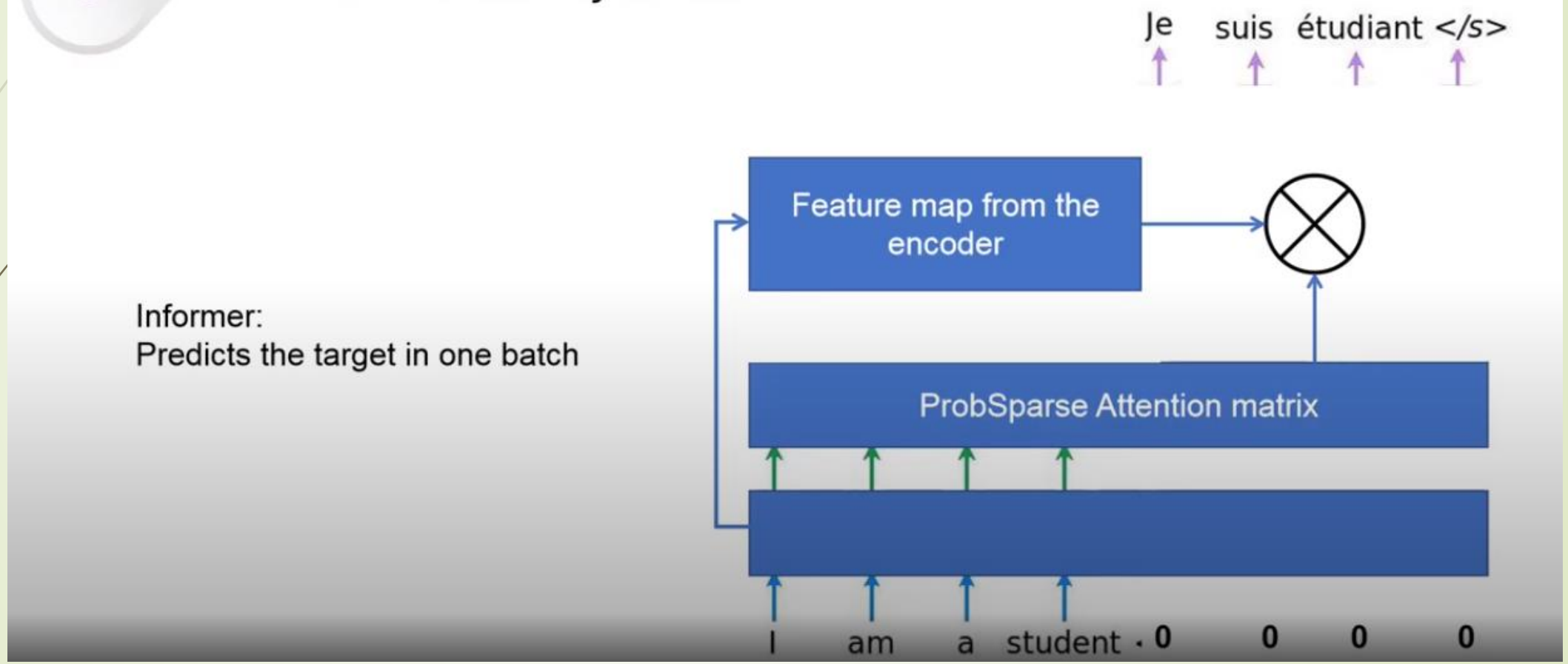
From the Lemma 1 (proof is given in Appendix D.1), we propose the max-mean measurement as

$$\overline{M}(\mathbf{q}_i, \mathbf{K}) = \max_j\{\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}}\} - \frac{1}{L_K}\sum_{j=1}^{L_K}\frac{\mathbf{q}_i\mathbf{k}_j^\top}{\sqrt{d}} \quad (4)$$

Main Improvements

3

Generative Style Decoder:



Informer:
Predicts the target in one batch

Experiments and Results

4 Experiment

Datasets

We extensively perform experiments on four datasets, including 2 collected real-world datasets for LSTF and 2 public benchmark datasets.

ETT (Electricity Transformer Temperature)[2]: The ETT is a crucial indicator in the electric power long-term deployment. We collected 2-year data from two separated counties in China. To explore the granularity on the LSTF problem, we create separate datasets as $\{\text{ETTh}_1, \text{ETTh}_2\}$ for 1-hour-level and ETTm_1 for 15-minute-level. Each data point consists of the target value "oil temperature" and 6 power load features. The train/val/test is 12/4/4 months.

ECL (Electricity Consuming Load)[3]: It collects the electricity consumption (Kwh) of 321 clients. Due to the missing data (Li et al. 2019), we convert the dataset into hourly consumption of 2 years and set 'MT_320' as the target value. The train/val/test is 15/3/4 months.

Weather[4]: This dataset contains local climatological data for nearly 1,600 U.S. locations, 4 years from 2010 to 2013, where data points are collected every 1 hour. Each data point

Experiments and Results

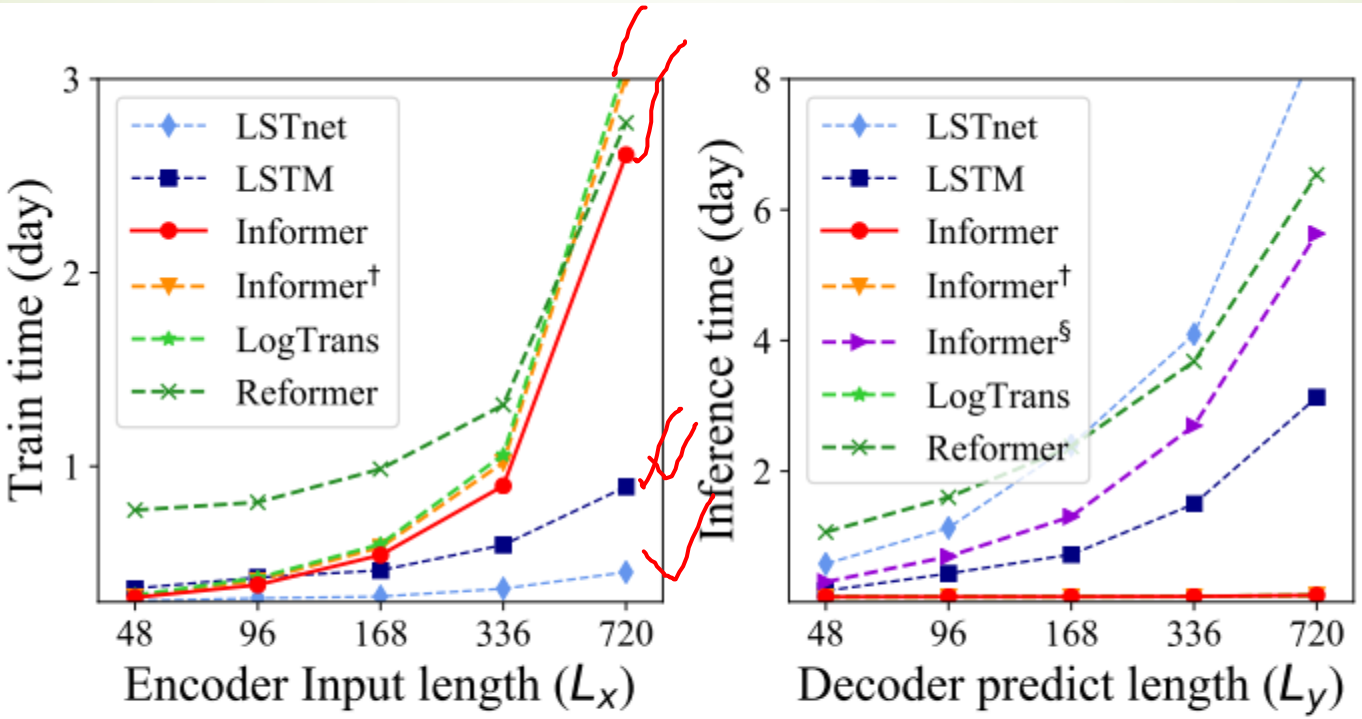
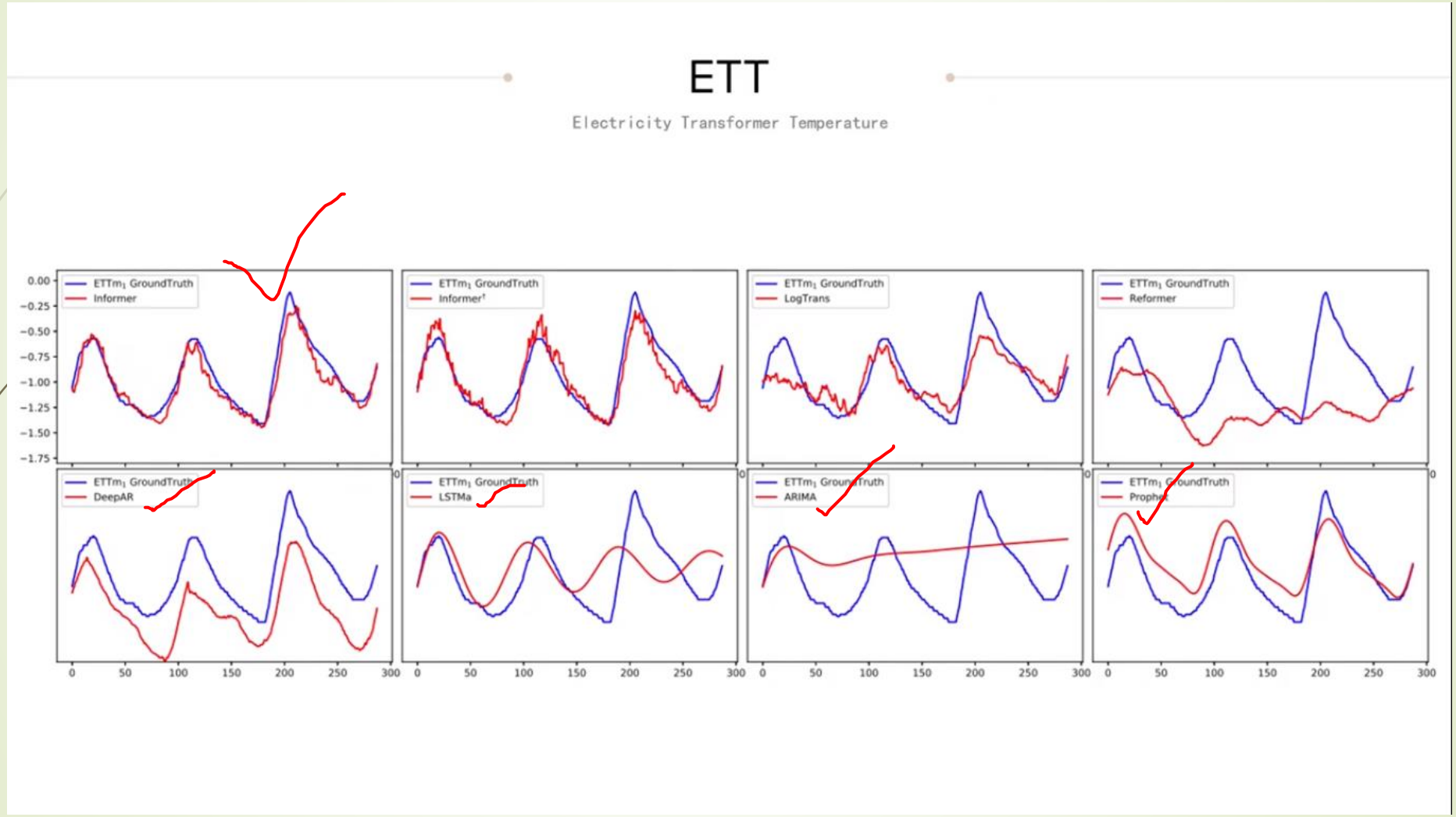


Figure 5: The total runtime of training/testing phase.

Experiments and Results



Limitations

1. assume all keys are IID following some Gaussian distributions **(basic assumption)**

2. It holds only for queries resulting in top-u KLDs **(conition)**.

3. The attention values are long-tail distributed **(verified, but on one dataset only)**.

4. The variance of query-key dot products also decreases along with the KLD **(Alert, comes from no where)**.

Thank You!

Contact: abdullahal.mamun1@wsu.edu