Paper Review: From Known to Unknown: Knowledge-guided Transformer for Time-Series Sales Forecasting in Alibaba

Abdullah Mamun

October 11, 2021

About this paper

From Known to Unknown: Knowledge-guided Transformer for Time-Series Sales Forecasting in Alibaba

Xinyuan Qi, Hou Kai, Tong Liu, Zhongzhong Yu, Sihao Hu, Wenwu Ou

Alibaba Group

{qishui.qxy,houkai.hk,sihao.hsh}@alibaba-inc.com, {yingmu,santong.oww}@taobao.com, yuzzhongcs@163.com

Abstract

Time series forecasting (TSF) is fundamentally required in many real-world applications, such as electricity consumption planning and sales forecasting. In e-commerce, accurate time-series sales forecasting (TSSF) can significantly increase economic benefits. TSSF in e-commerce aims to predict future sales of millions of products. The trend and seasonality of products vary a lot, and the promotion activity heavily influences sales. Besides the above difficulties, we can know some future knowledge in advance except for the historical statistics. Such future knowledge may reflect the influence of the future promotion activity on current sales and help achieve better accuracy. However, most existing TSF methods only predict the future based on historical information. In this work, we make up for the omissions of future knowledge. Except for introducing future knowledge for prediction, we propose Aliformer based on the bidirectional Transformer, which can utilize the historical information, current factor, and future knowledge to predict future cooperation operations, including factory production, merchant stocking, and marketing strategy planning, cannot be achieved alone without the time-series sales forecasting algorithm.

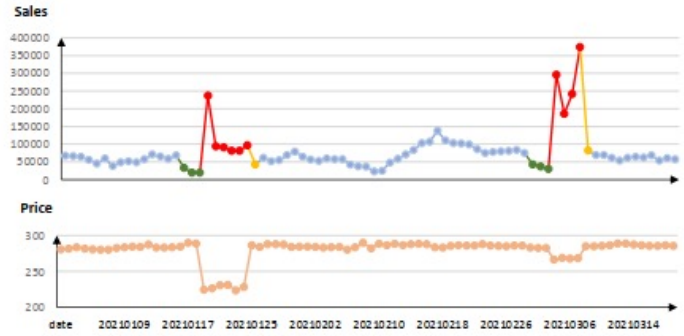

Figure 1: The time series of sales and price of an eye-shadow in Tmall experienced two promotion activities: one range from Jan 20 to Jan 25 and another from Mar 5 to Mar 8.

About this paper

- Published on arxiv in September 2021.
- Proposes a modified transformer named Aliformer

Introduction

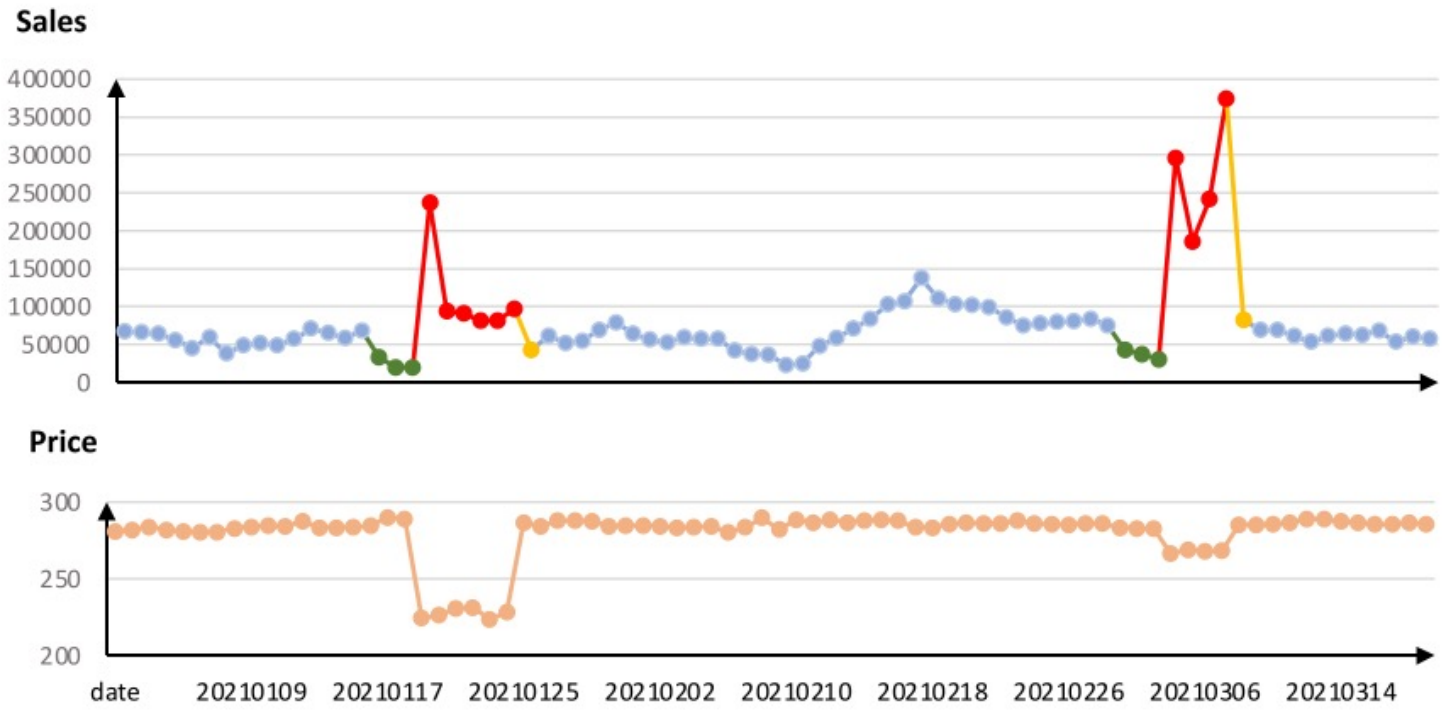

Figure 1: The time series of sales and price of an eye-shadow in Tmall experienced two promotion activities: one range from Jan 20 to Jan 25 and another from Mar 5 to Mar 8.

Method

Problem Formulation

Given a sequence of historical statistics and knowledge information, the time series sale forecasting task aims to forecast the statistics in a future period. Let n denotes a product, its historical statistics and knowledge information can be represented as a chronological sequence:

$$\mathbb{S} = \left\{ s_1^{(n)}, s_2^{(n)}, ..., s_T^{(n)} \right\}$$

$$\mathbb{K} = \left\{ k_1^{(n)}, k_2^{(n)}, ..., k_T^{(n)} \right\}$$

where the statistics $s_t^{(n)}$ represents the historical statistics at t-th time (e.g. historical payed amount *et al.*), $k_t^{(n)}$ denotes the t-th time knowledge information (e.g. price at t-th time, which may vary at each time).

For time t, input $x_t^{(n)}$ can be represented with an embedding layer Emb:

$$x_t^{(n)} = \text{Emb}\left(\left\{ s_t^{(n)}, k_t^{(n)} \right\} \right) \quad 1 \leq t \leq T$$

where Emb means an FC layer for numerical features and a lookup table for id features, which map the features to $\mathbb{R}^{d_x}$. The sum of numerical and id features make up the input x_t^n.

Given the historical information of product n, the time series sale forecasting method predict the future sales can be formulated as:

$$\left\{ y_{T+1}^{(n)}, ..., y_{T+L}^{(n)} \right\} = f_h \left(x_1^{(n)}, ..., x_T^{(n)} \right)$$

Method

certain discount). That is $k_t^{(n)}$ $(T+1 \leq t \leq T+L)$ can be known and:

$$\mathbb{K} = \left\{ k_1^{(n)}, k_2^{(n)}, ..., k_{T+L}^{(n)} \right\}$$

With the future knowledge involved, the input $x_t^{(n)}$ can be represented as:

$$x_t^{(n)} = \begin{cases} \text{Emb}\left(\left\{ s_t^{(n)}, k_t^{(n)} \right\} \right) & 1 \leq t \leq T \\ \text{Emb}\left(\left\{ u_t^{(n)}, k_t^{(n)} \right\} \right) & T+1 \leq t \leq T+L \end{cases}$$

the time series sale forecasting be formulated as:

$$\left\{ y_{T+1}^{(n)}, ..., y_{T+L}^{(n)} \right\} = f_b \left(x_1^{(n)}, ..., x_{T+L}^{(n)} \right)$$

where f_b denotes predicting the future based on both the historical information and future knowledge, $u_i^{(n)}$ means a default value or learnable token.

Method

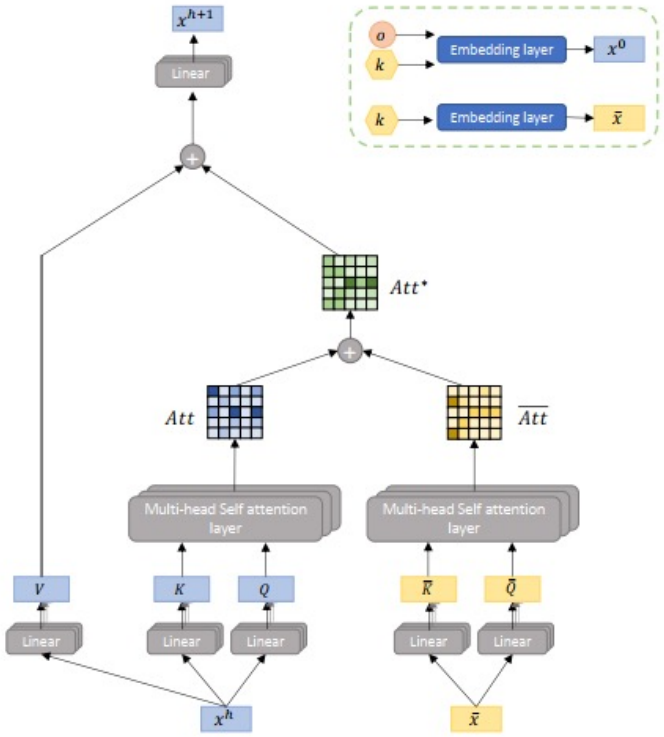

Figure 3: AliAttention layer. The knowledge-guided branch utilizes the pure knowledge information and outputs the knowledge-guided attention $\overline{\text{Att}}$ to revise the final attention map Att^*

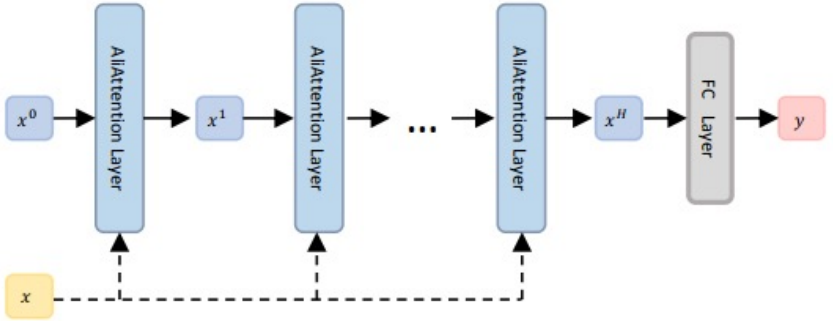

Figure 4: The overview of Aliformer. Original knowledge information is provided to each AliAttention layer.

based on the historical statistics and knowledge with the vanilla self-attention (VSA):

$$\text{Att}\,(i,j) = \frac{\left(x_i^h W_Q\right)\left(x_j^h W_K\right)^T}{\sqrt{d}}$$

$$x_i^{h+1} = \sum_j \text{Softmax}(\text{Att}\,(i,j))\left(x_i^h W_V\right) W$$

where x_i^h is $x^{(n)}$ in h-th VSA layer at time i, d is a scale factor, W_Q, W_K, W_V, W represent linear layer to compute Q, K, V in vanilla self-attention and x_i^{h+1} respectively. With

(rest of the presentation was continued with the paper)

Thank You!

Contact: abdullahal.mamun1@wsu.edu