Graph Attention Networks for New Product Sales Forecasting in E-Commerce

Chuanyu Xu[1], Xiuchong Wang[1], Binbin Hu[2], Da Zhou[3(✉)], Yu Dong[1], Chengfu Huo[1], and Weijun Ren[1]

[1] Alibaba Group
[2] Ant Group
[3] Xiamen University
{tracy.xcy, xiuchong.wxc, dongyu.dy, chengfu.huocf,afei}@alibaba-inc.com *, bin.hbb@antfin.com, zhouda@xmu.edu.cn

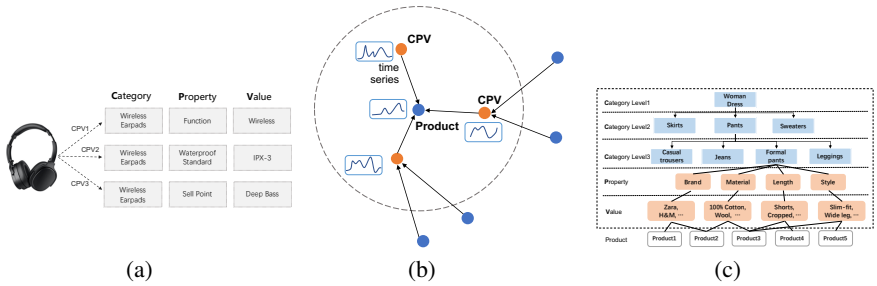
Abstract. Aiming to discover competitive new products, sales forecasting has been playing an increasingly important role in real-world E-Commerce systems. Current methods either only utilize historical sales records with time series based models, or train powerful classifiers (*e.g.,* DNN and GBDT) with subtle feature engineering. Despite effectiveness, they have limited abilities to make prediction for new products due to the sparsity of product-related features. With the observation on real-world data, we find that some additional time series features (*e.g.,* brand and category) implying product characteristics also play vital roles in new product sales forecasting. Hence, we organize them as a new kind of dense feature called CPV (Category-Property-Value) and propose a Time Series aware Heterogeneous Graph (TSHG) to integrate CPVs and products based time series into a unified framework for fine-grained interaction. Furthermore, we propose a novel Graph Attention Networks based new product Sales Forecasting model (GASF) that jointly exploits high-order structure and time series features derived from THSG for new product sales forecasting with graph attention networks. Moreover, a multi trend attention (MTA) mechanism is also proposed to solve temporal shifting and spatial inconsistency between the time series of products and CPVs. Extensive experiments on an industrial dataset and online system demonstrate the effectiveness of our proposed approaches.

Keywords: Time Series aware Heterogeneous Graph · Graph Attention Networks · Multi trend attention.

1 Introduction

With the development of E-Commerce, the user scale on E-Commerce platform is constantly expanding and the consumer demand is increasingly diversified. Discovering and supplying high-quality new products has been becoming the core technology to meet diversified needs of customers. According to the statistical data in Alibaba [4], millions of new products are released everyday and they contribute 40% of GMV (Gross

* First Author and Second Author contribute equally to this work.
[4] https://www.alibaba.com/



(a) (b) (c)

Fig. 1. (a) Illustration of CPVs for wireless headphone. (b) A toy example of TSHG in our scenario. (c) Illustration of CPV structure.

Merchandise Volume) in the platform. Therefore, new product sales forecasting, which aims to discover competitive new products, has played a fundamental role in enhancing the productivity of E-Commerce and satisfying user experience.

Intuitively, the new product sales forecasting in E-Commerce can be formulated as a time series prediction problem, which is well explored in numerous studies. Naturally, conventional solutions propose to adopt time series models (*e.g.,* Auto-Regressive Integrated Moving Average (ARIMA) [2] and Long Short Term Memory network (LSTM) [7]) for prediction, which only utilize historical sales records. However, the sparsity and instability of historical records of new products may harm the performance of these models. Besides, a series of feature based methods are proposed to perform subtle feature engineering for each product, and then train a powerful classifier (*e.g.,* Gradient Boosting Decision Tree (GBDT) [4] and Deep Neural Networks [3]) for prediction. Due to the powerful ability of feature learning, these methods have achieved a considerable improvement on sales forcasting task. Nevertheless, we argue that the existing methods have three major limitations for new product sales forecasting.

- **L1**: They commonly utilize product-related features (*e.g.,* user behavior features and product static information) to make prediction for product sales in future. Unfortunately, these features are sparse or even absent for new products in real-world application, which seriously hinders the forecasting performance.
- **L2**: These approaches treat new products and its additional features (*e.g.,* category and brand) separately, which ignores the interaction between them, resulting in sub-optimal performance in the complex scenario.
- **L3**: With the analysis of real-world data, we find the temporal shifting and spatial inconsistency (we will revisit temporal shifting and spatial inconsistency in latter sections) between the time series of products and additional features in our business scenario, which is poorly explored in the existing approaches.

To address these issues, we aim to comprehensively explore and exploit abundant product-related features and time series features in a more proper way, and propose a novel Graph Attention network based Sales Forecasting approach, called GASF shortly.

Inspired from daily business experience that the sales of new products are mainly determined by whether their characteristics meet the market trend, besides original

time series features, we propose to construct *Category - Property - Value* (called CPV shortly) features to characterize the trends of new products more comprehensively. Figure 1(a) shows an example. Wireless headphone is described by the CPV "*Wireless Earpads - Function - Wireless*", which denotes that it belongs to the "*Wireless Earpads*" category and has the "*Wireless*" "*Function*". Note that a product can be described by multiple CPVs and a CPV can also be used to describe multiple products. In contrast with historical sales records, the proposed CPV features capture the sales trends in the macro level, especially for new products with limited features or interactions (**L1**). Therefore, it is quite likely to take full advantage of time series features for improving the performance on sales forecasting task by integrating above two aspects of information together.

On the other hand, graph has been proposed as a general approach to model various types of objects. In order to jointly consider products and extracted CPVs together, we propose **T**ime **S**eries aware **H**eterogeneous **G**raph (TSHG for short) to effectively capture underlying specialities of new products for sales forecasting. As shown in Figure 1(b), products and their CPVs are connected and objects (products or CPVs) in TSHG contains time series. With the help of recently emerging graph neural networks [15], high-order structure derived from TSHG and time series features in products and CPVs can be naturally explored in an unified framework (**L2**).

With the observation of real data in our E-commerce scenario , we find the temporal shifting and spatial inconsistency between the time series of products and CPVs, which means the response speed (*i.e.,* temporal) and intensity (*i.e.,* spatial) of products and CPVs are quite different for the hot spot in the market. To fill this gap, we proposed a novel Multi Trend Attention (MTA) mechanism in GASF, which (1) shifts the trend of the product over multiple time units on the time axis to get multiple distances of the product and CPV trends (spatial inconsistency), and (2) gets the time series trend by taking first-order derivative and ensures the trend in the same space (spatial inconsistency) (**L3**). With MTA, our model is expected to learn fine-grained interaction of time series between products and CPVs beyond topological structure.

To sum up, we make the following contributions:

- Inspired from daily business experience, we construct a new kind of heterogeneous feature called CPV in E-Commerce to overcome the sparsity of new products. Moreover, we propose to frame the new product sales forecasting problem in the setting of TSHG, which integrates the products, CPVs and time series in an unified framework.
- We propose GASF, an end-to-end approach to simultaneously extract time series and structural information in TSHG. To our best knowledge, it is the first attempt to introduce deep graph learning with attention mechanism for sales forecasting task, which provides a new perspective to capture fine-grained interaction between products and other objects in real-world E-commerce scenarios.
- With the analysis of real data, the temporal shifting and spatial inconsistency between the time series of products and CPVs is uncovered and a novel multi trend attention mechanism is designed in GASF model to solve it.
- We perform extensive experiments on an Alibaba dataset for product sales forecasting. The results demonstrate that our model consistently and significantly out-

perform various state-of-the-arts. Moreover, our model also achieves significant performance improvement on online system.

2 Preliminaries

In real-world E-Commerce systems, a new product is associated with a series of basic information (*i.e.,* category, property and value) when it is released. In order to comprehensively characterize new products for sales forecasting, we proposed to extract category-property-value (CPV) features with normalization as follows:

- **Category**: we apply the clustering technology and calculate frequencies for category names in each cluster, the name with highest frequency will be selected as the standard category and other category names are mapped to it.
- **Property**: We summarize properties on category and normalize them via Word2vec [11, 12]. After representing each property as embedding, we follow the same process flow of category to obtain standard properties with clustering and mapping.
- **Value**: We summarize values on category-property. These values can be normalized by synonyms and Word2vec algorithm.

We show an example of well-established CPV structure in Figure 1(c). CPV is defined as a kind of property and value under a category in E-Commerce. For example, we can observe that "Product2" can be describe as "Woman Dress - Brand - Zara" or "Woman Dress - Material - Wool" in Figure 1(c).

In this paper, we aim at leveraging above CPV features and time series to effectively capture underlying specialities of new products for sales forecasting. Hence, we frame our task in the setting of time series aware heterogeneous graph, which considers the CPV features and time series in an unified framework.

Definition 1. ***Time series aware heterogeneous graph (TSHG)*** *A TSHG is defined as a directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, \mathcal{T}\}$, where $\mathcal{V} = \mathcal{V}^{CPV} \bigcup \mathcal{V}^{Pro}$ consists of the CPV node set $\mathcal{V}^{CPV}$ and the product node set $\mathcal{V}^{Pro}$ and $\mathcal{E}$ contains edges connecting products and corresponding CPVs. $\mathcal{T} = \{t_v | \in \mathcal{V}\}$ is the set of times series on nodes (products or CPVs). Moreover, $t_v = \{\tau^1, ..., \tau^M\}$, where τ^i is a fixed-length time series and indicates that t_v is a M-channel time series composed of M different single-dimensional time series.*

The TSHG provides a flexible way to model various complex interactions between products and CPVs in an unified framework, which could be used to enhance sales forecasting. Given the above preliminaries, we are ready to formulate our task.

Definition 2. ***TSHG enhanced sales forecasting*** *Given an time series aware heterogeneous graph $\mathcal{G}$, for each product $p \in \mathcal{P}$, we aim to learn prediction functions $\mathcal{F}^C(p|\mathcal{G}; \Theta^C)$ and $\mathcal{F}^R(p|\mathcal{G}; \Theta^R)$ to estimate whether product p will be sold in the future (classification) and its total sales over a time period (regression), respectively. Here Θ^C and Θ^R represent the parameters of the prediction function $\mathcal{F}^C$ and $\mathcal{F}^R$, respectively.*

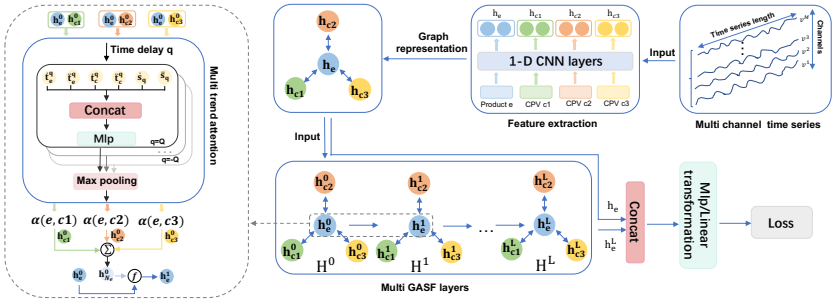


Fig. 2. The overall architecture of the proposed GASF approach.

3 Proposed GASF Model

In this section, we present GASF, an unified model to leverage CPV features and time series for new products sales forecasting with graph attention network. We show the framework of our proposed model in Figure 2.

3.1 Feature Extraction

Since the original time series for products are multi-channel time series, we firstly present how to extract features to characterize products and CPVs, as show in Figure 2. Following the well-established technology in previous work [13, 18], we set up multiple convolution neural network (CNN) layers for dimension reduction and feature extraction. Specifically, we adopt two 1-D CNN layers with "VALID" padding and the number of filters of second CNN layer is set to be 1. Hence, each node $v \in \mathcal{V}$ can be represented as a fixed-length vector $\mathbf{h}_v$. It is worthwhile to note that node v can be a product or a CPV in TSHG.

3.2 GASF layer

As mentioned above, we propose a TSHG to flexibly model various complex interactions between products and CPVs in an unified framework. We now build upon the architecture of graph attention network [15] to recursively capture structural information in TSHG. Distinct from previous works [15, 8, 9], we propose a multi trend attention (MTA): $\mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ to generate attentive weights between nodes, which overcomes the temporal shifting and spatial inconsistency between products and CPVs in our scenario. Here we start with the description of a single GASF layer, consisting of information propagation and information aggregation, followed by the stack of multiple GASF layers.

Information Propagation Intuitively, a certain node (a product or a CPV) in TSHG can be easily influenced by its neighbors. In order to capture such fine-grained interactions, we perform information propagation between a target node and its neighbors.

Formally, given a node v, we use $\mathcal{N}_v$ to denote its neighbor set. We use linear weighed combination to characterize the local structural information for node v:

$$\mathbf{h}_{\mathcal{N}_v} = \sum_{c \in \mathcal{N}_v} \alpha(v, c)\mathbf{h}_c, \tag{1}$$

where $\alpha(v, c)$ weighs the importance of each propagation on edge $v \leftarrow c$. We implement $\alpha(v, c)$ via multi trend attention, which will be introduce later.

Information Aggregation Next, we aggregate the node representation $\mathbf{h}_v$ and its neighborhood representation $\mathbf{h}_{\mathcal{N}_v}$ to enhance the expressive ability. We simply takes summation of target node and its neighbor representation as follows:

$$f_{Sum} = \text{ReLU}(\mathbf{h}_v + \mathbf{h}_{\mathcal{N}_v}). \tag{2}$$

High-order propagation Since a single GASF may be inadequate in capturing complex interactions between products and CPVs in TSHG, we further stack multiple GASF layers to explore the information propagated from high-order neighbors. Formally, given a node v, we recursively obtain its representation through information propagation and aggregation in the l-th step as follows:

$$\mathbf{h}_v^{(l)} = f(\mathbf{h}^{(l-1)}, \sum_{c \in \mathcal{N}_v} \alpha(v, c)\mathbf{h}_c^{(l-1)}). \tag{3}$$

Here, we recursively propagate the representations from a target node's neighbors to refine the node's representation in THSG. Moreover, we set $\mathbf{h}_v^{(0)} = \mathbf{h}_v$ at the initial information propagation iteration.

3.3 Multi Trend Attention

The key idea of attention mechanism is to learn a weighted representation across target node and its neighbors, which aims to propagate more informative features from neighbors to target node. Hence, attention mechanism is naturally implemented to learn the similarity between time series of products and CPVs in our well-established TSHG. The more similar the time series trends of the two nodes are, the more relevant they are. This implementation is based on the assumption that products and CPVs have the similar response to the hot spots in market, but temporal shifting and spatial inconsistency between the time series of products and CPVs are widely existed in our scenario.

– **Temporal shifting** means products and CPVs have different response times to market trends, which may lead to a gap between the time series of products and CPVs on the temporal view. In Figure 3(a) we observe that the trend of product and CPV2 are more similar, which indicates that they are more related to each other. However, CPV2 responds to the market more slowly than the product, resulting in the temporal shifting of time series between between them. This phenomenon is very common in E-Commerce. For example, the release of "*Iphone*" may subsequently lead to the growth of CPV "*Iphone protective cover - Style- Cartoon*" over a period of time.

– **Spatial inconsistency** indicates the different response intensity of time series of products and CPVs for hot spots in market. As shown in Figure 3(a), it is clear that the euclidean distance of product between CPV1 is smaller than that of product between CPV2, even though the time series trend of CPV2 is more similar to the product. In Figure 3(b) we notice that these time series trends are comparable in a space where the euclidean distance of the product and CPV2 becomes smaller and more reasonable. It shows that time series trend reflects the similarity between time series in a better way.

Temporal shifting and spatial inconsistency between the time series of products and CPVs reveal that products and CPVs have different response speed and intensity for hot spots in market, which cannot be captured by traditional attention mechanism. Hence, we propose a novel multi trend attention mechanism to overcome this issue, aiming to calculate the relevance of products and CPVs. For convenience, we denote the target product and CPV node in TSHG as e and c, respectively. For each product, we move q times unit for the time series of product e, where $q > 0$ means move forward and $q < 0$ means move backward. Note that we retain the time series where product e and CPV c overlap on the time axis, and thus missing values before and after the retained time series are filled with the first and last value, respectively. Subsequently, we can get fixed-length vectors for product e and CPV c, and denote them as $\mathbf{h}_e^q$ and $\mathbf{h}_c^q$. Now we can get their trends $\dot{\mathbf{t}}_e^q$ and $\dot{\mathbf{t}}_c^q$ as follow:

$$\dot{\mathbf{t}}_e^q = \frac{\mathbf{h}_e^q[1:d] - \mathbf{h}_e^q[0:d-1]}{\mathbf{h}_e^q[0:d-1] + \lambda}, \dot{\mathbf{t}}_c^q = \frac{\mathbf{h}_c^q[1:d] - \mathbf{h}_c^q[0:d-1]}{\mathbf{h}_c^q[0:d-1] + \lambda}, \tag{4}$$

where $\mathbf{h}_e^q[0:d-1]$ and $\mathbf{h}_e^q[1:d]$ denotes the first and last (d-1) -dimension features of $\mathbf{h}_e^q$, respectively ($\mathbf{h}_c^q[0:d-1]$ and $\mathbf{h}_c^q[1:d]$ are similar). λ is a smoothing parameter. In additon, we apply Eq. (4) to $\dot{\mathbf{t}}_e^q$ and $\dot{\mathbf{t}}_c^q$ to get their trend $\ddot{\mathbf{t}}_e^q$ and $\ddot{\mathbf{t}}_c^q$, which can reveal the speed of time series trend change.

Next, we calculate the similarity between the trend of the product e and the CPV c (*i.e.,* $\dot{\mathbf{t}}_e^q$ and $\dot{\mathbf{t}}_c^q$) as well as the speed of their trend changes with q time units interval, which is defined as:

$$\dot{s}^q(e,c) = g(\dot{\mathbf{t}}_e^q,\ \dot{\mathbf{t}}_c^q), \ddot{s}^q(e,c) = g(\ddot{\mathbf{t}}_e^q,\ \ddot{\mathbf{t}}_c^q), \tag{5}$$

where $g(\cdot,\cdot)$ measures the similarity of two vectors, which is set as the inverse of euclidean distance in our paper.

By integrating above information together, we are ready to formulate the attention score of product e amd CPV c with q time units interval as follows:

$$\alpha^q(e,c) = \text{ReLU}(\mathbf{W}[\dot{\mathbf{t}}_e^q \,||\, \dot{\mathbf{t}}_c^q \,||\, \ddot{\mathbf{t}}_e^q \,||\, \ddot{\mathbf{t}}_c^q \,||\, \dot{s}^q(e,c) \,||\, \ddot{s}^q(e,c)] + b), \tag{6}$$

where $\mathbf{W}$ and b is the weight matrix and bias, respectively. And $||$ is the concatenation operation.

Since the time unit interval for each product-CPV pair is different from each other, we choose the maximum interval of Q time units to obtain the final attention score as follows:

$$\alpha(e,c) = \max_{-Q \le q \le Q} \alpha^q(e,c) \tag{7}$$

(a) Time Series (b) Time Series Trend

Fig. 3. (a) represents sales volume of product, cpv1 and cpv2. (b) represents the approximate 1st-order right partial derivative of these time series. Temporal shifting represents the time difference between the reaction of the product and CPV to the market. Spatial inconsistency represents magnitude of the reaction of the product and CPV to the market, i.e. there exist dimensional inconsistency between the time series of the product and CPV.

3.4 Model Learning

After L-th propagation, we denote $h_v^{(L)}$ as the final representation, which captures both time series and structural information in TSHG. Inspired by [19, 6], we concatenate the initial vector (*i.e.,* $\mathbf{h}_v$) and high-order representation (*i.e.,* $h_v^{(L)}$) for later prediction.

$$\mathbf{h}_v^f = \tanh(\mathbf{W}_f[\mathbf{h}_v^0||\mathbf{h}_v^{(L)}] + \mathbf{b}_f), \tag{8}$$

where $\mathbf{W}_f$ and $\mathbf{b}_f$ is the weight matrix and bias vector, respectively. And $||$ denotes the concatenation operation.

As mentioned above, we aim to predict whether a product will be sold in the future (classification) and its total sales over a time period (regression), respectively. In our work, we feed $\mathbf{h}_f$ into MLP module for classification in order to implement a nonlinear function for feature interaction, while a linear transformation is adopted for regression. To guide the learning progress, we choose the cross entropy function with negative sampling for classification [14] and mean squared error function for regression [1].

4 Experiments

In this section, we conduct comprehensive experimental studies to verify the effectiveness of our method by answering the following three questions:

RQ1 Does our proposed GASF model outperform other state-of-the-art methods on both classification and regression tasks?
RQ2 How does the proposed GASF perform for new products sales forecasting at different released times ?
RQ3 How sensitive is the proposed GASF model to the hyper-parameters ?

4.1 Experimental Settings

Datasets. To demonstrate the effectiveness of the proposed approach, we conduct experiments on an Alibaba [5] real dataset. The dataset contains 8428378 new products and

[5] https://www.1688.com

1765293 CPVs, new products are released over the time period from 01-05-2019 to 30-10-2019. Also, products and CPVs are 7-channel time series composed of 7 different single-dimensional time series. These time series are extracted from online traffic logs and represent seven different user behaviors such as exposure page views (PV), exposure unique visitor (UV), click PV, click UV, add to cart UV, pay UV and order UV. The length of each behavior time series is 60, representing the number of such behavior in the past 60 days. The labels of each product are whether it would be sold in the next 30 days (classification) and its total sales in the next 30 days (regression). In our experiment, the training set is taken over 08-21-2019 to 08-30-2019. The testing set is taken in 09-30-2019.

Evaluation Protocol. We evaluate the performance of the proposed model on two main tasks, namely binary classification and regression. Two metrics are used here for binary classification evaluation: 1) Area Under Curve (AUC) to evaluate the model's ranking performance; and 2) Precision at Top-N (P@N) to evaluate the model's ability of distinguishing top products. Regression task aims to predict the total sales for a product. We introduce two classical metrics [13] for the performance evaluation: weighted Mean Absolute Percentage Error (wMAPE) and Mean Absolute Error (MAE).

Baselines. We compare our model with the following methods:**Historical Average (HA)** is a heuristic-based baseline. **Lasso** takes the historical sales records as input for Logistic Regression/Linear Regression with L_1 regularization. **Gradient Boosting Decision Tree (GBDT)** is a common used technique for both classification and forecasting regression problem in industry, we carefully design 70 features from these 9 log indicators. **DNN** is a simple neural network architecture with 3 fully connection layers and a linear regression layer. **GBDT-CPV** adds CPV features on the basis of GBDT. we apply a pooling (average, maximum, median) to all neighbor heterogeneous features to improve the acquirement of information. **CNN-WD** [20] is a convolutional neural network based model for sales forecasting in E-Commerce.

Parameter Settings. For all approaches, we tune the model parameters by grid search and report the performance on the testing dataset. For GBDT models, we take these parameters: num_rounds=200, max_depth = 6, subsample = 0.8 and learning rate = 0.1. For DNN model, the dimensions for fully connected layers are [256,128], a dropout with p = 0.2 is applied to the output of last fully connected layer and learning rate is set to 0.001. For GASF model, the kernel sizes and number of filters for 1-D CNNs are [3, 7] and [20, 1], the dimensions for fully connected layers are [256,128] and λ=0.00001, we use Adam as optimizer [10] with a learning rate 0.0001.

4.2 Performance Comparison (RQ1)

Now we compare the performance of our GASF with the baselines. The comparison results are shown in Table. 1. The main observations are summarized as follows:

1. GASF achieves the best performance on both classification and regression tasks. One-sample paired t-test shows that all the improvements are statistically significant ($p < 0.005$). We think that the outperformance of our GASF would benefit

	Binary Classification				Regression	
Methods	AUC	P@30K	P@300K	P@1M	MAE	wMAPE
HA	0.6834	0.5597	0.1044	0.0541	4.0135	0.8891
Lasso	0.7854	0.5785	0.1045	0.0539	2.9324	0.6495
GBDT	0.8286	0.6492	0.1529	0.0589	2.6822	0.5941
GBDT-CPV	0.8591	0.6566	0.1566	0.0591	2.6746	0.5914
DNN	0.8441	0.6531	0.1584	0.0578	2.5137	0.5570
CNN-WD	0.8475	0.5825	0.1577	0.0581	2.5006	0.5541
GASF	0.8669	**0.6827**	0.1650	0.0592	2.4354	0.5396
GASF-MTA	**0.8750**	0.6821	**0.1661**	**0.0594**	**2.4056**	**0.5329**

Table 1. Overall performance comparison. The best performance of each setting is highlighted as bold font.

	AUC			
Methods	1 day	2-10 days	11-30 days	31-90 days
HA	0.5237	0.5967	0.7114	0.7913
GBDT	0.7403	0.7963	0.8547	0.8860
GBDT-CPV	0.7618	0.8117	0.8664	0.8914
CNN-WD	0.6745	0.7815	0.8611	0.8977
GASF-MTA	**0.7807**	**0.8324**	**0.8732**	**0.8995**

Table 2. Performance comparison over different released period. The best performance of each setting is highlighted as bold font.

from the design of GATs in capturing the spatial and temporal features from the input jointly.

2. Among all the baselines, HA and ARIMA typically underperform machine learning and deep learning based models, mainly because they only rely on historical sales records of new products. GBDT-CPV model outperforms original GBDT by absolute 3 points in AUC, indicating the significance of CPVs.
3. Thanks to the Multi Trend Attention (MTA), GASF-MTA outperforms GASF for both tasks. This shows the effectiveness of the proposed MTA and also indicates the importance of taking temporal shifting and spatial inconsistency of the time series into account.

4.3 Performance for different released times (RQ2)

In real business, the released period of new products ranges from 1 day to 90 days. The shorter the released period of a new product, the less information the forecasting model can obtain. Additional heterogeneous features are particular useful to alleviate such a "cold-start" problem in new product forecasting. Here, we study the forecasting performance *w.r.t.* the released periods, which varies in the set of $\{$1 day, 2 - 10 days, 10 - 30 days, 30 - 90 days $\}$. We show the AUC comparison results on the classification task in Table. 2. The main observations are summarized as follows:

1. With the increase of the released time, all models has achieved a better performance. This is due to the enhancement of new product historical information.
2. For all released period, our proposed GASF-MTA model shows a significant outperformance than other models. Improvements increase as release time decrease, this shows the effectiveness of our proposed model for new product sales forecasting.
3. Among all the GBDT approaches, GBDT-CPV outperforms GBDT for all released period. We should note that the improvements stem from the CPV. This shows the effectiveness of our proposed CPV for new product sales forecasting.

(a) L (b) Q

Fig. 4. The impact of key hyper-parameters for GASF (L: the number of the GASF layers; Q: the maximum interval of time units).

4.4 Hyper-parameter Study (RQ3)

In this section, we explore the impact of key hyper-parameters for GASF and how L and Q empirically influence the learning effect of GASF.

Figure 4(a) shows the performance of GASF with respect to L. We can see that $L = 2$ is better than $L = 1$; however, increasing L beyond 2 gives marginal returns in performance. This makes sense because larger L includes information of further nodes and thus needs deeper GASF to learn, which is more difficult to optimize. In this case, it may be easily over-fitting with more layers [5].

Figure 4(b) shows the performance of GASF with respect to Q. We can see that with the increasing of Q, the performance is improved to a maximum, and then decrease. This indicates the positive effect of using a larger interval for temporal shifting. However, too large interval may introduce noise and compromise the performance.

4.5 Comparison with Online System

We have successfully deployed our proposed GASF in our real promotion business scenario of Alibaba, and compare it with the best online baseline model(*i.e.,* GBDT-CPV) on Dec 2020. The experimental results shows that has a relative gain by **3.4%** and **7.6%** than on GMV and number of deduplicated buyers (BYR) respectively. This observation demonstrates the effectiveness and business value of our proposed approach in E-Commerce.

5 Conclusion and Future Work

In this paper, a novel GAT architecture model is presented for new product sales forecasting in E-Commerce, named GASF. GASF models products and their CPVs by a general graph-structured time series and extracts spatial and temporal features simultaneously. The experiments on two real-world tasks and online system demonstrate a significant outperformance of our proposed model. To the best of our knowledge, this is the first time to apply GATs for sales forecasting. For further improvements, we will pursue two directions. The first is to explore more heterogeneous relations such as completing relation between substitutable products, which may enhance the representation

ability of our model. The second is to incorporate side information [17, 16] into our multi trend attention, which can provide more flexibility to learn the weights of different product-CPV pair.

References

1. Chai, T., Draxler, R.R.: Root mean square error (rmse) or mean absolute error (mae)?–arguments against avoiding rmse in the literature. Geoscientific model development **7**(3), 1247–1250 (2014)
2. Chatfield, C.: The analysis of time series: an introduction (2003)
3. Chen, C., Liu, Z., Zhou, J., Li, X., Qi, Y., Jiao, Y., Zhong, X.: How much can a retailer sell? sales forecasting on tmall. In: PAKDD. pp. 204–216. Springer (2019)
4. Friedman, J.H.: Greedy function approximation: a gradient boosting machine. Annals of statistics pp. 1189–1232 (2001)
5. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: AISTATS. pp. 249–256 (2010)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
7. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
8. Hu, B., Shi, C., Zhao, W.X., Yu, P.S.: Leveraging meta-path based context for top-n recommendation with a neural co-attention model. In: SIGKDD. pp. 1531–1540 (2018)
9. Hu, B., Zhang, Z., Shi, C., Zhou, J., Li, X., Qi, Y.: Cash-out user detection based on attributed heterogeneous information network with a hierarchical attention mechanism. In: AAAI. pp. 946–953 (2019)
10. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR Workshop (2013)
12. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS. pp. 3111–3119 (2013)
13. Qi, Y., Li, C., Deng, H., Cai, M., Qi, Y., Deng, Y.: A deep neural framework for sales forecasting in e-commerce. In: CIKM. pp. 299–308 (2019)
14. Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., Mei, Q.: Line: Large-scale information network embedding. In: WWW. pp. 1067–1077 (2015)
15. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., Bengio, Y.: Graph attention networks. ICLR (2017)
16. Wang, H., Zhang, F., Hou, M., Xie, X., Guo, M., Liu, Q.: Shine: Signed heterogeneous information network embedding for sentiment link prediction. In: WSDM. pp. 592–600 (2018)
17. Wang, H., Zhang, F., Zhang, M., Leskovec, J., Zhao, M., Li, W., Wang, Z.: Knowledge-aware graph neural networks with label smoothness regularization for recommender systems. In: SIGKDD. pp. 968–977 (2019)
18. Wang, J., Sun, T., Liu, B., Cao, Y., Zhu, H.: Clvsa: A convolutional lstm based variational sequence-to-sequence model with attention for predicting trends of financial markets. In: IJCAI. pp. 3705–3711 (2019)
19. Xu, K., Li, C., Tian, Y., Sonobe, T., Kawarabayashi, K.i., Jegelka, S.: Representation learning on graphs with jumping knowledge networks. arXiv preprint arXiv:1806.03536 (2018)
20. Zhao, K., Wang, C.: Sales forecast in e-commerce using convolutional neural network. arXiv preprint arXiv:1708.07946 (2017)