Relation-aware Meta-learning for Market Segment Demand Prediction with Limited Records

Jiatu Shi\textsuperscript{1*}, Huaxiu Yao\textsuperscript{2*}, Xian Wu\textsuperscript{3}, Tong Li\textsuperscript{1},
Zedong Lin\textsuperscript{1}, Tengfei Wang\textsuperscript{1}, Binqiang Zhao\textsuperscript{1}, Zhenhui Li\textsuperscript{2}
\textsuperscript{1}Alibaba Group, \textsuperscript{2}Pennsylvania State University, \textsuperscript{3}University of Notre Dame
\{jiatu.sjt, litong.lt, zedong.lzd, wenlin.wtf, binqiang.zhao\}@alibaba-inc.com
\{huaxiyao, JessieLj\}@psu.edu, xwu9@nd.edu

\textbf{ABSTRACT}

Recently, E-commerce platforms have extensive impacts on our human life. To provide an efficient platform, one of the most fundamental problems is how to balance the demand and supply in market segments. While conventional machine learning models have achieved great success on data-sufficient segments, it may fail in a large-portions of segments in E-commerce platforms, where there are not sufficient records to learn well-trained models. In this paper, we tackle this problem in the context of market segment demand prediction. The goal is to facilitate the learning process in the target segments even facing a shortage of related training data by leveraging the learned knowledge from data-sufficient source segments. Specifically, we propose a novel algorithm, RMLDP, to incorporate a multi-pattern fusion network (MPFN) with a meta-learning paradigm. The multi-pattern fusion network considers both local and global temporal patterns for segment demand prediction. In the meta-learning paradigm, the transferable knowledge is regarded as the model parameter initializations of MPFN, which are learned from diverse source segments. Furthermore, we capture the segment relations by combining data-driven segment representation and segment knowledge graph representation and tailor the segment-specific relations to customize transferable model parameter initializations. Thus, even with limited data, the target segment can quickly find the most relevant transferred knowledge and adapt to the optimal parameters. Extensive experiments are conducted on two large-scale industrial datasets. The results show that our RMLDP outperforms a set of state-of-the-art baselines. In addition, RMLDP has also been deployed in Taobao, a real-world E-commerce platform. The online A/B testing results further demonstrate the practicality of RMLDP.

\textbf{KEYWORDS}

Market Segment Demand, Meta-learning, Segment Relation

\section{1 INTRODUCTION}

Large-scale E-commerce platforms (e.g., Amazon, Taobao) have significantly changed our life. To build a more efficient E-commerce platform, one of the most fundamental problems is how to balance the demand and supply in market, which requires an accurate demand prediction model for every market segment (e.g., wallet, belt). An accurate demand prediction model benefits the platform from three aspects: 1) pre-allocate resources to meet the market demand; 2) reduce the backlog of commodities; 3) optimize the allocation strategies of traffic source. In addition, due to the lags between upstream and downstream of the supply chain, real-time segment demand prediction (e.g., predict the next day’s demand) may be impractical. Instead, given the historical demand records, we study the market segment demand prediction problem as predicting the demand value of a future target time period (e.g., one month) several weeks in advance (as illustrated in Figure 1).

To predict market segment demand, traditional ensemble models (e.g., XGBoost [5]) and advanced deep learning methods (e.g., LSTM [12], GRU [6]) are capable of capturing time-varying sequential patterns (e.g., seasonal trend) and making accurate predictions. The superiority of these methods relies on large-scale labeled training data. Unfortunately, as illustrated in Figure 2, a large portion of market segments is long-tail and only has limited records, which leads to unsatisfactory prediction performance. Such unsatisfactory prediction performances, in turn, affect the efficiency of the platform. The reasons are two-fold: 1) Usually, market segments with more records are more likely to be exhibited in the platform. For the data-sufficient market segments, lack of exhibition opportunities causes the difficulties in collecting new records. The process finally forms a vicious circle, resulting in the homogenization of the platform; 2) The data-sufficient market segments is not equivalent to high-quality segments. Due to the limited resources (e.g., the number of exhibitions), for data-insufficient market segments, purely rely on the support of platform managers leaves the performance of the mainstream segments. Therefore, how to improve the prediction performance for market segments with limited data remains a non-trivial but necessary problem.

To tackle this “small data” problem, recently, knowledge transfer (e.g., transfer learning, meta-learning) [7, 27] has achieved great success in a series of applications, such as computer vision [20, 34], natural language processing [11, 17, 18], social goods [39]. To facilitate the learning process of target tasks with limited labeled data, knowledge transfer leverages the prior knowledge learned
from relevant source tasks. In the segment demand prediction, simply applying conventional knowledge transfer al-gorithms to improve the performance of data-insufficient segments faces the following two major challenges:

- **C1: How to improve the stability of knowledge transfer?** Usually, the performance of knowledge transfer relies on the similarity of distributions between source and target tasks. Significant difference between the distributions may cause unstable transfer or even worse prediction performance. Therefore, a sufficiently generalized knowledge transfer framework is required, which covers comprehensive and diverse temporal patterns of market segments.

- **C2: How to incorporate the complex relations among market segments?** It is non-trivial to capture the complex segment relations by using traditional knowledge transfer methods (e.g., fine-tuning), where the transferable knowledge are globally shared across source segments. However, in E-commerce platforms, the differences between segments can not be overlooked and thus the globally shared transferable knowledge may not be robust enough to all scenarios. For example, the demands of down jackets are probably similar to the demands of coats, while dissimilar to t-shirts. Thus, segment relations are necessary to be incorporated in knowledge transfer framework.

Hence, to address the above challenges, in this paper, we propose a novel framework RMLDP for data-insufficient market segment demand prediction. The goal for RMLDP is to build a customized meta-learning paradigm upon a market demand prediction model. Specifically, we first construct a multi-pattern fusion network (MPFN) for market segment demand prediction, which jointly captures both local and seasonal temporal patterns by two Gated recurrent units (GRUs). Regrading the MPFN as base model, next, the first challenge is solved by learning and transferring the model parameter initializations of the MPFN under the meta-learning paradigm. Here, various source segments sampled from diverse categories (e.g., food, clothing) are used for initialization learning. Finally, a data-driven segment representation and a segment knowledge graph representation are introduced to capture the complex segment relations. For each segment, the captured relational information are further used to modulate the model parameter initializations.

In summary, our major contributions are three-fold:

- To the best of our knowledge, we are the first to study the problem of market segment demand prediction with limited data by transferring the knowledge from mainstream segments.
- We develop a novel framework, RMLDP, to solve the market segment demand prediction task. RMLDP incorporates a multi-pattern fusion network with the meta-learning paradigm. The segment relations are further distilled to customize the model parameter initializations in meta-learning paradigm. Furthermore, we deploy the proposed method into the online platform.
- We collect the market demand records from two large-scale E-commerce platforms: Juhuasuan and Tiantiantemai. Comparing with baseline methods, the superior performance of RMLDP demonstrates the effectiveness of our framework under both offline and online scenarios.

## 2 RELATED WORK

In this section, we briefly discuss two categories of related work: time series prediction and knowledge transfer.

### 2.1 Time Series Prediction

Traditional approaches (e.g., ARIMA [26], Kalman filtering [19]) have been widely used in time series applications. These methods fail to capture complex non-linear temporal correlations due to the limited expressive capability. With stronger expressive power, deep learning methods, especially recurrent neural network based approaches (e.g., GRU [6] and LSTM [12]), have achieved great success in time series modeling [14, 15, 22, 29, 30, 32, 36]. To further improve the prediction performance, recently, more information have been incorporated in the basic recurrent neural network structures by applying attention mechanism [23, 29] or multi-resolution modeling [13, 37, 38]. However, all these methods rely on large-scale training data. In contrast, the goal for our work is to improve the prediction of data-insufficient target segments by knowledge transfer. Besides, those methods focus on the prediction for next step/a few steps. In this work we focus on the early prediction for a future time interval under the real-world E-commerce scenario.

### 2.2 Knowledge Transfer

To benefit the learning process on task with limited data, transferring knowledge from its related tasks has achieved great success in recent years [27]. Conventional transfer learning methods learn transferable latent factors between one source domain and one target domain. The latent factors are captured by a series of techniques, such as matrix factorization [21], manifold learning [10] and deep learning [20, 34]. Recently, meta-learning (a.k.a., learning to learn) provides a more stable and flexible way for knowledge transfer. The goal for meta-learning is to generalize the knowledge from various of tasks and then to adapt these knowledge to unseen tasks. In meta-learning, the transferable knowledge are regarded as model parameter initializations [7–9, 16, 41, 42], metric mapping function [24, 33, 35, 43], or meta-optimizer [4, 31], etc. In the time series related problems, Oreshkin et al. briefly discusses the relation between the neural time series prediction and meta-learning
meta-learning [25]. Yao et al. incorporates the gradient-based meta-
learning with a region functionality based memory [39] for few-shot
spatiotemporal prediction. However, this method highly relies on the
spatial semantic correlations between tasks, which limits its applica-
bility in our problem. To the best of our knowledge, we are the first
to study market segment demand prediction with limited records by
borrowing relation-aware knowledge from other segments.

3 PRELIMINARIES
In this section, we define some concepts and notations and then for-
mingly define our problem. Assuming the whole market is split into
I market segments \{s_1, \ldots, s_I\}, each market segment \(s_i\) represents
one category of products (e.g., sweaters, orange juice).

**Definition 1 (Market Demand Value)** For each segment \(s_i\) at time
step \(t_i\), the market demand value \(x_{s_i, t_i}\) is defined as the number of
purchasing requests in a fixed time window \([t_i, t_i + 1]\). In this paper,
the length fixed time interval \(t\) is defined as one day (i.e., \(t = 1\)).

**Definition 2 (Target Demand Value)** As illustrated in Figure 1,
we aim to predict the market demand for a future target time interval
\(T_t\) several weeks in advance. Supposing the current time stamp is \(t_c\)
and the time lag between the current time and the further target time
is \(T_g\), we define the target demand value \(y_{i, t_c}\) as the total market
demand value of \(s_i\) between time interval \([t_c + T_g, t_c + T_g + T_t]\) (i.e.,
\(y_{i, t_c} = \sum_{t_c+T_g}^{t_c+T_g+T_t} x_{s_i,t} \)).

**Problem: Market Segment Demand Prediction with Limited
Records** Assuming that we have a set of diverse source segments
\(s_1, \ldots, s_I\) and a target segment \(s_t\) with limited records, we aim to
predict the target demand value \(y_{t, t_c}\) in the testing dataset of
the target segment. Additionally, for each segment \(s_i\) at time stamp
\(t_c\), we further introduce several statistical features \(e_{i, t_c}\) (e.g., \# of
items, sellers, brands) and customers’ action features (e.g., click,
collect, add to cart and take order).

We denote the concatenation of market demand value \(x_{s_i, t_c} \in \mathbb{R}^1\)
and external features \(e_{i, t_c} \in \mathbb{R}^{-1}\) as \(x_{s_i, t_c} = x_{s_i, t_c} \oplus e_{i, t_c} \in \mathbb{R}^n\). The
market segment demand prediction model (a.k.a., base model) is
defined as \(f\) with the learnable parameters \(\theta\). Formally, our problem is
formulated as:

\[
y_{i, t_c} = \arg \max_{\hat{y}_{i, t_c}} p(y_{i, t_c} | \theta_{\hat{y}_i}, \{x_{s_1, t_c}, \ldots, x_{s_I, t_c}\})
\]

where \(\theta_{\hat{y}_i}\) denotes the segment-specific initializations, which are
transferred from all source segments using the target segment informa-
tion. Detailed discussions about customized model initializations are
in Section 4.2 and 4.3. We name the process of learning trans-
ferred knowledge (i.e., all grey blocks in Figure 3) from data-sufficient source segments. In particular, the base model \(f\) is first designed as a multi-pattern fusion net-
work (MPFN), where both local and seasonal temporal patterns are con-
considered. Then, RMLDP incorporates the base model \(f\) and the

4 METHODOLOGY
In this section, we introduce our proposed framework: RMLDP
(Relation-aware Meta-Learning for Demand Prediction). The whole
framework is shown in Figure 3. The goal for RMLDP is to facilitate
the learning process of data-insufficient target segment demand pre-
diction by adapting the transferred knowledge (i.e., all grey blocks in
Figure 3) from data-sufficient source segments. In particular, the
base model \(f\) is first designed as a multi-pattern fusion net-
work (MPFN), where both local and seasonal temporal patterns are con-
considered. Then, RMLDP incorporates the base model \(f\) and the
meta-learning paradigm, where the model parameter initializations are
regarded as transferable knowledge. To further modulate the
parameter initializations, we distill knowledge from segment repres-
resentations, including a data-driven segment representation and a
segment knowledge graph representation. In the following subsections,
we detail three key components: multi-pattern fusion network,
knowledge transfer and adaptation, relation-aware modulation.

4.1 Multi-pattern Fusion Network
In this subsection, we propose a multi-pattern fusion network
(MPFN) for market segment demand prediction. The framework is
illustrated in Figure 4. The goal for MPFN is to predict the target
demand value by capturing the temporal patterns from the his-
torical records. To achieve this goal, we adopt a GRU network to
capture non-linear relations among historical records. Concretely,
for predicting the target demand value \(y_{i, t_c}\) of segment \(s_i\), the most
recent \(|T_g|\) demand values (i.e., \(\{x_{s_i,t_1}, \ldots, x_{s_i,t_{|T_g|}}\}\)) are fed into
the GRU, which is formulated as:

\[
h_{t_c, t_c} = \text{GRU}^\theta(h_{t_c-1, t_c}; x_{s_i, t_c}).
\]

The temporal representation \(h_{t_c, t_c}\) encodes the local temporal pat-
terns from the closest records.

As mentioned in Section 3, different from real-time demand prediction,
there exists a time lag \(T_g\) between current time and the target prediction time. Thus, the temporal patterns captured from
closest demand records are probably insufficient to achieve satisfac-
tory performance. Fortunately, seasonal temporal patterns provide
us with useful periodic information. For example, the demand trend for
winter coat in this December is similar to the trend in the last
December. However, as suggested in [40], it is non-trivial to train a
single GRU network for handling long-term seasonal patterns due
to the risk of gradient vanishing. Instead, another GRU network is
introduced to model the seasonal patterns as:

\[
h_{t_c, t_c} = \text{GRU}^\theta(h_{t_c-1, t_c}; x_{s_i, t_c}),
\]

where \(t_1 = t_c + T_g - 365\) represents the corresponding historical
time of the target demand value (i.e., same day in the last year). The
sequence \(\{x_{s_i,t_1}, \ldots, x_{s_i,t_{|T_g|}}\}\) are fed into \(\text{GRU}^\theta\).

By fusing the hidden representations \(h_{t_c, t_c}\) and \(h_{t_c, t_c}\) as \(h_{t_c, t_c} =
\text{GRU}^\theta(h_{t_c, t_c} \oplus h_{t_c, t_c})\), both local and seasonal temporal patterns are captured.
Then, we use one fully connected layer for prediction as:

\[
y_{i, t_c} = W_f \hat{y}_{i, t_c} + b_f,
\]

where \(W_f\) and \(b_f\) are learnable parameters. In this paper, mean
square error (MSE) is used as loss function as:

\[
\mathcal{L} = \sum_{t_c}(y_{i, t_c} - \hat{y}_{i, t_c})^2.
\]

As mentioned in Section 3, the MPFN is regarded as base model \(f\)
with all learnable parameters are denoted as \(\theta\).

4.2 Knowledge Transfer and Adaptation
After constructing the base model MPFN, we discuss the meta-
learning paradigm, which transfers knowledge from source seg-
ments to the target segment with limited data. To increase the
robustness of knowledge transfer, the transferable knowledge are ex-
pected to be general enough and contain comprehensive relations
between market segments and their historical temporal patterns.
where the empirical risk is defined as mean square error in Eqn. (5). $\mathcal{D}_{te}^{i} = \{X_{i,te}, y_{i,te}\}_{i=1}^{N_{te}}$ is the training sets sampled from segment $s_i$, where $N_{te}$ denotes the number of training samples and $X_{i,te} = \{x_{i,te-|C|+1}, \cdots, x_{i,te-|C|+1}, \cdots, x_{i,te+|C|+1}, \cdots, x_{i,te+|C|+1}\}$ represents the used demand sequence in MPFN.

After getting the segment-specific parameter $\theta_i$, we sample the testing dataset $\mathcal{D}_{te}^{s_j} = \{X_{s_j,te}, y_{s_j,te}\}_{i=1}^{N_{te}}$ from $s_j$ to update the model parameter initializations $\theta_0$ by minimizing the empirical risk as:

$$\theta_0 \leftarrow \min_{\theta_0} \frac{1}{|I|} \sum_{i=1}^{|I|} \mathcal{L}(\mathcal{D}_{te}^{s_j}; \theta_0)$$  

(7)

where $|I|$ denotes the number of source segments. At the end of meta-training process, we get $\theta_0$ as the learned optimal model parameter initializations.

Given a target segment $s_j$, the segment-specific parameter $\theta_i$ is achieve by performing gradient descent starting from the learned initializations $\theta_0^*$ with the training data $\mathcal{D}_{te}^{i}$, i.e.,

$$\theta_i = \theta_0^* - \alpha \nabla_{\theta} \mathcal{L}(\mathcal{D}_{te}^{i}; \theta).$$  

(8)

We finally evaluate the performance by the testing set $\mathcal{D}_{te}^{s_j}$ of segment $s_j$ using adapted parameter $\theta_i$.

### 4.3 Relation-aware Customization

The above knowledge transfer and adaptation framework regards the transferable knowledge as the globally shared model parameter initializations $\theta_0^*$ across all source segments. However, the globally shared knowledge may incapable of well-capturing underlying complex segment relations. For example, suppose we need to predict the market demand of Men’s clothing, both local and seasonal temporal trends are similar to clothing from other groups (e.g., women, children), while the temporal trends are probably dissimilar to the electric appliances. Thus, in this section, we tailor the segment-specific relations to modulate the model parameter initializations. Specifically, we consider two types of segment relational representations: data-driven segment representation and segment knowledge graph representation. The data-driven segment representation implicitly encrypt the segment relations. Generated by users’ purchase records, the segment knowledge graph further explicitly models the relations among different segments. We detail these two types of representations in the following subsections.

#### 4.3.1 Data-driven Segment Representation

For data-driven segment representation, we encode the segment-specific information into one representation vector. The relations among segments are implicitly included in the representations since similar segments have similar representations. As suggested in [41], learning the representation of each segment $s_i$ is equal to aggregate the training data $\mathcal{D}_{te}^{i}$ to a representation vector. Here, we introduce one MPFN as aggregator denoted as MPFN*. The aggregator first encodes each
data sample into one vector and then a sample-level mean pooling layer is applied on the top of encoder. Formally, the aggregation process is formulated as:

$$q^d_i = \frac{1}{N_T} \sum_{t_c} MPFN^\theta_t (X^t_{i, t_c}).$$  \hspace{1cm} (9)

Empirically, only using the loss signal defined in Eqn. (7) to guide the segment representation learning is difficult. To increase the stability of segment representation learning, we introduce the reconstruction loss with a decoder $MPFN^\theta_{dec}$, which is defined as:

$$L_{rec} = \frac{1}{N_T} \sum_{t_c} ||X^t_{i, t_c} - MPFN^\theta_{dec} (MPFN^\theta_t (X^t_{i, t_c}))||^2_F$$  \hspace{1cm} (10)

where $|| \cdot ||_F$ is defined as Frobenius norm.

4.3.2 Segment Knowledge Graph Representation. In real-world E-commerce platforms, the relations between segments can further be reflected by users’ purchasing records. For each pair of segments, their similarity is proportional to the frequency of co-occurrence in the same order. For example, women usually purchase sweater and skirt together. But it is unlikely to purchase shampoo and refrigerator at the same time. Given users’ purchasing records, we build a segment knowledge graph $G_i$, where each node $n_i$ in the knowledge graph represents one segment. For each pair of nodes $n_u$ and $n_v$, the link weight $\omega_{uv}$ is calculated by the co-occurrence frequency in the same order. We further set a threshold to filter some low similarity links.

Then, to map each segment into a fixed low dimensional space and maintain their relational structure, we adopt Deepwalk [28] on the constructed knowledge graph $G_i$. The representation of each segment in the knowledge graph is denoted as $q^d_i$. Note that the data-driven segment representation and the segment knowledge graph representation are mutually complementary. In data-driven segment representation, the similarity of segments mainly reflects their temporal patterns. By contrast, the similarity of segments in this knowledge graph are guided by users’ purchasing records.

4.3.3 Relation Fusion and Knowledge Modulation. After generating the data-driven segment representation $q^d_i$ and the knowledge graph representation $q^g_i$, we then fuse these two types of representations and get the final segment-specific representation as: $q_i = q^d_i \odot q^g_i$. To customize the globally shared model parameter initialization $\theta_0$, we introduce a modulating function $M(\cdot)$, which consists of a mapping layer with an activation function. The modulating function is defined as:

$$M(q_i) = \sigma (W_m q_i + b_m).$$  \hspace{1cm} (11)

where $W_m$ and $b_m$ are trainable parameters. By using the modulating function, the segment representation is mapped to the same space of the model parameter initializations $\theta_0$. Then, the customization process is formulated as:

$$\theta_i = M(q_i) \odot \theta_0.$$  \hspace{1cm} (12)

Here $\theta_i$ represents the task specific parameter initializations. Then, for segment $i$, we perform the gradient steps starting from the customized initializations $\theta_i$ rather than $\theta_0$.

By combining the empirical risk $L$ in Eqn. (7) and the reconstruction loss $L_{rec}$ in Eqn. (10), we revise the objective function in Eqn. (7) and formulate the final objective function as:

$$\min_{\theta} \sum_{i=1}^{N_f} L(D^f_i; \theta_i) + \lambda L_{rec}.$$  \hspace{1cm} (13)

where the hyperparameter $\lambda$ is used to balance the value of two loss terms. We describe the whole meta-training process for RMLDP in Algorithm 1.

5 SYSTEM OVERVIEW

In this section, we introduce the system deployment pipeline of RMLDP in the real-world platform. In order to implement our algorithm in business scenarios and maintain the low coupling and high cohesion in system design, we independently abstract the strategy center from the E-commerce background management system.
In Figure 5, we show the online system. In E-commerce platform, the operators first define the market segments based on a series of field configurations (e.g., the consumer groups served, price ranges of goods, brand collections, etc.). Then, the predicted segments are selected and the strategy center provides the demand prediction results by using the proposed algorithm. Based on the prediction results, the reasonable market flow resource for each segment is determined and sent to the background system. The background system assigns the specific display time and the display channel for each segment. Finally, the segment-specific information are displayed in the consumers’ client App.

With the rapid development of cloud computing technology, behavioral logs (e.g., detailed product information, user’s click and purchase records on the client) are collected back to cloud storage in real-time. Based on this data, the MapReduce task deployed in the cloud extracts the required features and provides a steady feature stream. With the help of GPU, the neural network model is efficiently trained. The whole system forms a complete closed loop.

6 EXPERIMENTS
In this section, we conduct comprehensive experiments to evaluate our proposed RMLDP by answering the following major research questions: (1) How is the overall prediction performance of RMLDP compared with state-of-the-art baselines? (2) How do various components we proposed (e.g., market segment knowledge graph) impact the model’s performance? (3) How is the online performance of RMLDP based on our proposed method and system?

6.1 Experimental Setups
In the experiment settings, we describe two real-world datasets and the compared baselines. The mean absolute percentage error (MAPE) are used to evaluate the performance.

6.1.1 Dataset Description. To evaluate our proposed method, we collect the data from two large-scale marketing scenarios in Taobao, the largest E-commerce platform in China [2]. We detail the descriptions in the follows:

- **Juhuasuan** [1]: The first dataset is collected from Juhuasuan, one of the largest platform for group buying in China. There are more than 4000 segments over 800 days. We select 3 coarse-grained categories (electric appliances, clothing and daily supplies) with 9 fine-grained categories (large electric appliances, small electric appliances, digital electric appliances, women’s clothing, men’s clothing, sports, food, daily chemicals, daily sundry).

- **Tiantiantemai** [3]: Another dataset collected from Tiantiantemai, one of the largest platform for low-cost products in China. There are more than 6000 market segments over 189 days. The segments are selected from 3 coarse-grained categories (household, clothing, food) with 9 fine-grained categories (kitchenware, bedding, toiletries, women’s clothing, men’s clothing, children’s clothing, snacks, fresh, drink).

For both Juhuasuan and Tiantiantemai, we sort all segments by the number of purchasing records. We select top 70% segments with more records for meta-training and the rest for meta-validation and meta-testing. For each fine-grained category in Juhuasuan and Tiantiantemai, in this experiment, the averaged performance (MAPE) over all segments with this category are reported.

6.1.2 Hyperparameter Settings. In Table 3, we list all hyperparameters of Juhuasuan and Tiantiantemai.

6.1.3 Baselines. We compare our proposed method with the following four types of baselines: (1) **Basic regression methods**: Linear Regression, Support vector regression (SVR); (2) **Ensemble regression methods**: Random Forest, XGBoost [5]; (3) **Neural-network-based methods**: GRU, Dipole [23], LSTNet [14]. For GRU and Dipole, we use the MPFN as backbone models and denote these two methods as GRU-MPFN and Dipole-MPFN, respectively. (4) **Transfer methods**: Fine-tuning, MAML [7]. In Fine-tuning, we use the same strategy as GRU-MPFN to learn the model parameters. Then, we fine-tune the learned parameters for each target segment. For all baselines, we use the same features as RMLDP. In basic regression, ensemble regression and neural-network-based methods, the training dataset includes samples from all source segments and the training samples from target segments for fair comparison.

6.2 Results
6.2.1 Overall Performance. After implementing our proposed model and comparing with other baselines, we report the results for Juhuasuan and Tiantiantemai in Table 1 and Table 2, respectively. For each fine-grained category, the averaged MAPE over segments in this category are reported. According to these results, we draw the following conclusions:

- All other types of baselines significantly outperform the basic regression methods (i.e., Linear regression, SVR). The reason is that it is non-trivial to capture complex non-linear temporal patterns through the basic regression methods.
- All transfer learning methods (i.e., MAML, Finetune and RMLDP) achieves better performance than other non-transfer methods. The results suggest that finetuning the learned knowledge from other segments can capture the task-specific information in the target segment and further benefit the performance.
- In all cases, our RMLDP outperforms other baselines. Especially, RMLDP achieves better performance than MAML, which indicates the effectiveness of customizing model parameter initializations by leveraging the complex relations across market segments. Combining with the segment relations, the stability and diversity of transferred knowledge increases to the highest degrees.

6.2.2 Ablation Study. We further perform comprehensive ablation studies to demonstrate the effectiveness of proposed components. We describe the ablation models as follows:

- **RMLDP-d**: In RMLDP-d, we remove the data-driven segment representation and only use the segment knowledge graph representation to modulate the model parameter initializations.
- **RMLDP-g**: In RMLDP-g, the segment knowledge graph is removed and the data-driven segment representation is the only signal for customizing model parameter initializations.
- **RMLDP-szn**: We only consider the local temporal trend in RMLDP-szn, i.e., the GRU is removed in the base learner MPFN.
- **RMLDP-local**: Contrary to RMLDP-szn, in RMLDP-local, we remove GRU in the base learner MPFN.
Table 1: Overall Performance of Juhuasuan.

| Model          | Electric Appliances | Clothing | Daily Supplies |
|----------------|---------------------|----------|---------------|
|                | Large | Small | Digital | Women | Men | Sports | Food | Chemicals | Sundry |
| Linear Regression | 42.19% | 43.26% | 41.43% | 46.32% | 47.13% | 47.94% | 41.02% | 42.24% | 43.97% |
| SVR            | 30.16% | 29.26% | 30.06% | 34.45% | 35.56% | 36.94% | 30.11% | 30.31% | 33.12% |
| Random Forest  | 26.51% | 27.41% | 26.58% | 28.53% | 29.36% | 30.55% | 26.49% | 26.84% | 27.29% |
| XGBoost        | 25.38% | 26.81% | 25.19% | 27.49% | 28.37% | 28.76% | 26.93% | 26.53% | 26.07% |
| GRU+MPFN       | 25.62% | 26.34% | 25.46% | 27.51% | 27.34% | 29.06% | 27.09% | 26.67% | 26.33% |
| Dipole+MPFN    | 25.53% | 26.07% | 25.37% | 27.43% | 27.05% | 28.86% | 27.01% | 26.27% | 26.20% |
| LSTNet         | 25.98% | 26.66% | 26.31% | 27.48% | 27.56% | 29.13% | 28.93% | 26.94% | 26.89% |
| Fine-tuning+MPFN | 24.29% | 26.20% | 24.13% | 27.12% | 26.84% | 27.29% | 26.67% | 26.01% | 26.01% |
| MAML           | 24.21% | 26.08% | 23.53% | 28.53% | 29.36% | 30.55% | 27.09% | 25.04% | 25.99% |

*: comparing with MAML, the results of RMLDP are significant according to Student’s t-test at level 0.01.

Table 2: Overall Performance of Tiantiantemai.

| Model          | Household | Clothing | Food |
|----------------|-----------|----------|------|
|                | Kitchenware | Bedding | Toiletries | Women | Men | Children | Snacks | Fresh | Drink |
| Linear Regression | 56.89% | 47.29% | 55.68% | 47.51% | 49.98% | 46.45% | 49.64% | 53.53% | 54.11% |
| SVR            | 37.16% | 36.26% | 38.25% | 37.81% | 38.94% | 38.55% | 38.29% | 38.19% | 39.25% |
| Random Forest  | 31.21% | 30.41% | 31.45% | 29.40% | 31.35% | 29.15% | 30.81% | 30.84% | 31.92% |
| XGBoost        | 31.16% | 30.81% | 30.86% | 29.49% | 31.43% | 29.46% | 30.13% | 30.49% | 31.07% |
| GRU+MPFN       | 31.62% | 30.34% | 30.33% | 29.48% | 31.91% | 29.71% | 30.19% | 30.67% | 31.03% |
| Dipole+MPFN    | 30.97% | 29.89% | 30.01% | 29.41% | 30.69% | 30.12% | 30.14% | 30.33% | 30.66% |
| LSTNet         | 31.49% | 30.76% | 30.12% | 30.09% | 30.98% | 30.01% | 30.84% | 31.16% | 31.98% |
| Fine-tuning+MPFN | 30.49% | 29.56% | 29.54% | 29.27% | 30.86% | 29.49% | 30.10% | 30.01% | 30.52% |
| MAML           | 29.55% | 29.08% | 29.41% | 29.06% | 30.07% | 29.31% | 29.97% | 29.84% | 30.01% |

*: comparing with MAML, the results of RMLDP are significant according to Student’s t-test at level 0.01

Table 3: Hyperparameter Settings.

| Hyperparameter | Juhuasuan | Tiantiantemai |
|----------------|-----------|---------------|
| batch size     | 128       | 128           |
| feature dimension | 48        | 48            |
| sequence length | 30        | 30            |
| GRU embedding dimension | 128      | 128           |
| dimension of $q^d_i$ | 32        | 32            |
| dimension of $q^m_i$ | 16        | 16            |
| learning rate $\alpha$ | $10^{-4}$ | $10^{-4}$    |
| meta-learning rate $\beta$ | $10^{-3}$ | $10^{-3}$    |
| loss factor $\lambda$ | 0.5       | 0.5           |

The results for Juhuasuan and Tiantiantemai are reported in Table 4 and Table 5, respectively. The performance of RMLDP is also reported for comparison. From these tables, we observe that:

- Comparing with RMLDP, both RMLDP-d and RMLDP-g performs worse, indicating the effectiveness and complementarity of segment knowledge graph representation and data-driven segment representation.
- Comparing with RMLDP-d, RMLDP-g achieves better performance. The potential reason is that the data-driven market segment representations, which learned from training data of each segment, capture the segment-specific temporal patterns and provide more effective information.
- RMLDP significantly outperforms RMLDP-szn and RMLDP-local, indicating that both local and seasonal temporal patterns contribute to the model performance. The seasonal temporal patterns provide the basic estimation for the segment demand and the local temporal patterns further provide the calibration by using the most recent records.
### Table 4: Ablation studies of Juhuasuan.

| Model     | Electric Appliances | Clothing | Daily Supplies |
|-----------|---------------------|----------|---------------|
|           | Large | Small | Digital | Women | Men | Sports | Food | Chemicals | Sundry |
| RMLDP-d   | 24.18% | 26.03% | 23.41% | 26.45% | 26.47% | 27.84% | 25.63% | 25.01% | 25.97% |
| RMLDP-g   | 24.09% | 25.84% | 22.97% | 26.28% | 26.01% | 27.53% | 24.81% | 24.77% | 25.44% |
| RMLDP-szn | 28.01% | 27.97% | 26.67% | 33.29% | 34.72% | 35.88% | 29.31% | 29.34% | 31.51% |
| RMLDP-local | 30.42% | 31.38% | 31.23% | 34.98% | 35.87% | 37.32% | 33.09% | 31.09% | 33.49% |
| RMLDP     | 23.96% | 25.29% | 22.84% | 26.21% | 25.87% | 26.98% | 24.25% | 24.38% | 25.11% |

### Table 5: Ablation studies of Tiantiantemai.

| Model     | Household | Clothing | Food |
|-----------|-----------|----------|------|
|           | Kitchenware | Bedding | Toiletries | Women | Men | Children | Snacks | Fresh | Drink |
| RMLDP-d   | 29.35% | 29.07% | 29.36% | 29.01% | 29.96% | 28.53% | 29.89% | 29.78% | 29.86% |
| RMLDP-g   | 28.91% | 28.65% | 29.04% | 28.09% | 29.33% | 29.08% | 29.15% | 29.41% | 29.31% |
| RMLDP-szn | 33.83% | 33.54% | 34.99% | 37.36% | 36.78% | 37.98% | 37.54% | 36.54% | 38.27% |
| RMLDP-local | 41.96% | 37.29% | 42.53% | 40.15% | 42.39% | 40.49% | 41.04% | 42.93% | 41.45% |
| RMLDP     | 28.54% | 28.19% | 28.85% | 27.93% | 28.79% | 28.41% | 29.02% | 29.34% | 29.17% |

#### 6.2.3 Effect of Sequence Length.
In this section, we analyze the effect of sequence length (i.e., the value of $|T_c|$). We change the sequence length from 15 to 40 and the results for each large category of two datasets are shown in Figure 6. We can see that the MAPE decreases at the beginning and then keeps stable/slightly increases. The reason is that too short sequence may not provide enough information for accurate prediction. When the length of sequence increases, the information gradually become saturated and the results keep stable.

![Figure 6: (a), (b), (c): Prediction performance on each large category from Juhuasuan v.s. the sequence length $|T_c|$; (d), (e), (f): Prediction performance on each large category from Tiantiantemai v.s. the sequence length $|T_c|$.

#### 6.2.4 Analysis of Segment Representation.
In this section, we analyze the segment representation $q_i$ discussed Section 4.3.3. We use 694 testing segments from nine target segments in Juhuasuan. The results are shown in Figure 7. In this figure, we observe that the segment representations are capable of well-distinguishing different categories of segments and further provide qualitative evidence for effectiveness of RMLDP.

![Figure 7: Visualization of learned segment representation.

#### 6.2.5 Online Experiment.
To further evaluate the proposed model, we design the online experiments in Taobao mobile App. We conduct a bucket testing (i.e, A/B testing) in Tiantiantemai to test the consumers’ response to our RMLDP and baseline. For each segment, the higher demand prediction value it gets, the more opportunities of display it gains.

Without using prediction model, operators usually leverage the averaged demand value from the same period of the last year and the nearest month to predict the future demand. In offline evaluation, the MAPE for this statistical method is more than 0.8. We regard this statistical method as our baseline, and calculate five core indicators:
Relation-aware Meta-learning for Market Segment Demand Prediction with Limited Records

7 CONCLUSION

In this paper, we propose a novel relation-aware meta-learning framework, RMLDP, for market segment demand prediction with limited data by transferring knowledge from data-sufficient segments. Specifically, our proposed method incorporates the base demand prediction model (i.e., multi-pattern fusion network) into a meta-learning paradigm. The model parameter initializations are learned and transferred from source segments, which can be easily adapted to each target segment with limited data. Additionally, the relations across segments are learned and embedded into the representation of each segment. The segment representations are used to customize the model initializations. We conduct the experiments on two large-scale industry datasets and RMLDP consistently outperforms the state-of-the-art baselines. RMLDP is also deployed in the real-world platform with the positive bucket testing results.

REFERENCES

[1] 2020. Juhaasan. https://ju.taobao.com/.
[2] 2020. Taobao. https://www.taobao.com/.
[3] 2020. Tiantiantemai. https://tejia.taobao.com/.
[4] Marcin Andrychowicz, Misha Denil, Sergio Gomez, Matthew W Hoffman, David Plou, Tom Schaud, Brenden Shillingford, and Nando De Freitas. 2016. Learning to learn by gradient descent by gradient descent. In ICLR.
[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014).
[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks. In ICLR. 1126–1135.
[7] Chelsea Finn and Sergey Levine. 2018. Meta-learning and universality: Deep representations and gradient descent can approximate any learning algorithm. In ICLR.
[8] Chelsea Finn, Kelvin Xu, and Sergey Levine. 2018. Probabilistic Model-Agnostic Meta-Learning. In NeurIPS.
[9] Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. 2012. Geodesic flow kernel for unsupervised domain adaptation. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2066–2073.
[10] Jiatao Gu, Yong Wang, Li Wan, Kyunghyun Cho, and Victor OK Li. 2018. Meta-learning for low-resource neural machine translation. In EMNLP.
[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9, 8 (1997), 1735–1780.
[12] Chao Huang, Xian Wu, Xuchao Zhang, Chuzh Zhang, Jishu Zhao, Dawei Yin, and Nitesh V Chawla. 2019. Online Purchase Prediction via Multi-Scale Modeling of Behavior Dynamics. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2613–2622.

Table 6: The results of different prediction strategies

| Bucket     | PV    | UV    | #Seg | #Item | Ord  |
|------------|-------|-------|------|-------|------|
| Stat. Method | 1.93M | 0.64M | 1334 | 3708  | 91980|
| RMLDP      | 1.92M | 0.64M | 1458 | 3809  | 94780|

[13] Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. 2018. Modeling long-and short-term temporal patterns with deep neural networks. In SIGIR. 184–184.
[14] Nikolay Laptev, Jason Yosinski, Li Erzann Li, and Slawek Smyl. 2017. Time-series extreme event forecasting with neural networks at uber. In ICML \textit{Time Series Workshop}.
[15] Yoonho Lee and Seungjin Choi. 2019. Gradient-based meta-learning with learned layerwise metric and subspace. In ICML. 2933–2942.
[16] Zheng Li, Xin Li, Ying Wei, Lidong Bing, Yu Zhang, and Qiang Yang. 2019. Transferable end-to-end aspect-based sentiment analysis with selective adversarial learning. In EMNLP.
[17] Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In Thirty-Second AAAI Conference on Artificial Intelligence.
[18] Marco Lippi, Matteo Bertini, and Paolo Frasconi. 2013. Short-term traffic flow forecasting: An experimental comparison of time-series analysis and supervised learning. \textit{IEEE Transactions on Intelligent Transportation Systems} 14, 2 (2013), 871–882.
[19] Mingsheng Long, Yue Cao, Jiannin Wang, and Michael I Jordan. 2015. Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02971 (2015).
[20] Mingsheng Long, Jiannin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. 2013. Transfer learning with graph co-regularization. \textit{IEEE Transactions on Knowledge and Data Engineering} 26, 7 (2013), 1805–1818.
[21] Zhongjian Ly, Jiajie Xu, Kai Zheng, Hongzhi Yin, Pengpeng Zhao, and Xiaofang Zhou. 2018. LC-RNN: A Deep Learning Model for Traffic Speed Prediction. In IJCAI. 3470–3476.
[22] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In KDD, ACM, 1903–1912.
[23] Boris Oreshkin, Pau Rodríguez López, and Alexandre Lacoste. 2018. Task-dependent adaptive metric for improved few-shot learning. In NeurIPS. 721–731.
[24] Boris N Oreshkin, Dmitri Carpov, Nicolas Chapados, and Yoshua Bengio. 2020. N-BEATS Neural basis expansion analysis for interpretable time series forecasting. ICLR (2020).
[25] Bei Pan, Ugur Demiryurek, and Cyrus Shahabi. 2012. Utilizing real-world transportation data for accurate traffic prediction. In KDD. IEEE. 595–604.
[26] Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. IEEE TKDE 22, 10 (2009), 1345–1359.
[27] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. 2014. Deepwalk: Online learning of social representations. In KDD. 701–710.
[28] Yao Qin, Dongjin Song, Haisong Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. IJCAI (2017).
[29] Syama Sundar Rangapuram, Matthias W Seeger, Jan Gasthaus, Lorenzo Stella, Yuyang Wang, and Tim Januschowski. 2018. Deep state space models for time series forecasting. In NeurIPS. 7785–7794.
[30] Sachin Ravi and Hugo Larochelle. 2016. Optimization as a Model for Few-Shot Learning. ICLR (2016).
[31] David Salinas, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2019. DeepFactors for Forecasting. In Computer Vision and Pattern Recognition. 7167–7176.
[32] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Learning to learn by gradient descent by gradient descent. In ICLR.
[33] Chen Wu, Baosong Shi, Yuxiao Dong, Chao Huang, and Nitesh V Chawla. 2018. BEATS: Neural basis expansion analysis for interpretable time series forecasting. ICLR (2019).
[34] Yoshua Bengio. 2009. Learning Deep Architectures for AI. 1st Edition. The MIT Press.
[35] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Learning to learn by gradient descent by gradient descent. In ICLR.
[36] Zhirong Yang, Zhiyuan Liu, and Jianfeng Zhang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In Thirty-Second AAAI Conference on Artificial Intelligence.
[37] Tiantiantemai. https://tejia.taobao.com/.
[38] Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. 2020. Representation Learning on Variable Length and Incomplete Wearable-Sensory Time Series. IEEE Transactions on Knowledge and Data Engineering 32, 12 (2020), 2633–2647.
[39] Xian Wu, Chao Huang, Pablo Roblesgranda, and Nitesh Chawla. 2020. Representation Learning on Variable Length and Incomplete Wearable-Sensory Time Series. IEEE Transactions on Knowledge and Data Engineering 32, 12 (2020), 2633–2647.