Graph-Enhanced Multi-Task Learning of Multi-Level Transition Dynamics for Session-based Recommendation

Chao Huang[1], Jiahui Chen[2], Lianghao Xia[2], Yong Xu[2,3,4*], Peng Dai[1], Yanqing Chen[1], Liefeng Bo[1], Jiashu Zhao[5], Jimmy Xiangji Huang[6]
JD Finance America Corporation[1], USA
South China University of Technology[2], Peng Cheng Laboratory[3], China
Communication and Computer Network Laboratory of Guangdong[4], China
Wilfrid Laurier University[5], York University[6], Canada
chaohuang75@gmail.com, {201721041314, cslianghao.xia}@mail.scut.edu.cn, yxu@scut.edu.cn,
{peng.dai, yanqing.chen, liefeng.bo}@jd.com, jzhao@wlu.ca, jhuang@yorku.ca

Abstract

Session-based recommendation plays a central role in a wide spectrum of online applications, ranging from e-commerce to advertising services. However, the majority of existing session-based recommendation techniques (*e.g.*, attention-based recurrent network or graph neural network) are not well-designed for capturing the complex transition dynamics exhibited with temporally-ordered and multi-level inter-dependent relation structures. These methods largely overlook the relation hierarchy of item transitional patterns. In this paper, we propose a multi-task learning framework with Multi-level Transition Dynamics (MTD), which enables the jointly learning of intra- and inter-session item transition dynamics in automatic and hierarchical manner. Towards this end, we first develop a position-aware attention mechanism to learn item transitional regularities within individual session. Then, a graph-structured hierarchical relation encoder is proposed to explicitly capture the cross-session item transitions in the form of high-order connectivities by performing embedding propagation with the global graph context. The learning process of intra- and inter-session transition dynamics are associated by cross units, which seamlessly preserves the underlying low- and high-level item relationships in a common latent space. Extensive experiments on three real-world datasets demonstrate the superiority of MTD as compared to state-of-the-art baselines.

Introduction

Personalized recommendation has attracted a lot of attention in real-life applications, to alleviate information overload on the web (Xia et al. 2020). In various recommendation scenarios, session-based recommendation has become an important component in many online services (*e.g.*, retailing and advertising platforms) (Huang et al. 2004), to address the unavailability issue of user information in realistic scenarios (such as non-logged in customers or users without historical interactions) (Quadrana et al. 2017; Ren et al. 2019; Yuan et al. 2020). At its core is to predict the next interactive item based on a group of anonymous temporally-ordered behavior sequences of users (*e.g.*, clicked, browsed or purchased item sequences) (Liu, Zeng, and others 2018; Wang et al. 2020; Wang et al. 2019a). To facilitate the study of session-based recommendation, many efforts have been devoted to developing various deep neural network models, by exploring correlations between the future interested item and past interacted ones, which contributes to smarter recommendations.

Existing session-based recommendation methods for understanding the item transitional regularities can be grouped into several key paradigms. For example, one key research line aims to capture transitional patterns of interacted item sequence with recurrent neural network (Hidasi et al. 2015; Hidasi and Karatzoglou 2018). Along this line, to aggregate sequential embeddings into a more summarized session-level representation, researchers recently propose to augment recurrent session-based recommendation frameworks with attention mechanism (Li et al. 2017), or rely on the memory network (Liu, Zeng, and others 2018; Wang et al. 2019a). Furthermore, another recommendation paradigm utilizes graph neural network as the item transition encoder, to model long-term item dependencies within the session based on the structured relation graph (Wu et al. 2019).

Despite their effectiveness, we argue that these methods are not sufficient to yield satisfactory recommendation results, due to their failure in encoding complex item transition dynamics which are exhibited with multi-levels in nature (Song et al. 2019). Particularly, in the practical session-based recommendation scenarios, there exists session-specific short-term and long-term item transitions, as well as the long-range cross-session item dependencies in a global context (Al-Ghossein, Abdessalem, and Barré 2018). These different inter-correlations among items constitute the underlying multi-level item transition dynamics. As illustrated in Figure 1, while item t_7 and t_3 are not directly connected within the same session, there exist implicit inter-dependency among them, due to the item transitional relationship of $t_2 \rightarrow t_3$ and $t_7 \rightarrow t_2$ in session B and A, respectively. In such cases, items across different sessions are no longer independent. The dependent signals between interactive items may come from not only the intra-session transition regularities, but also inter-session item relations. However, to simplify the model design, most of cur-

*Corresponding author: Yong Xu

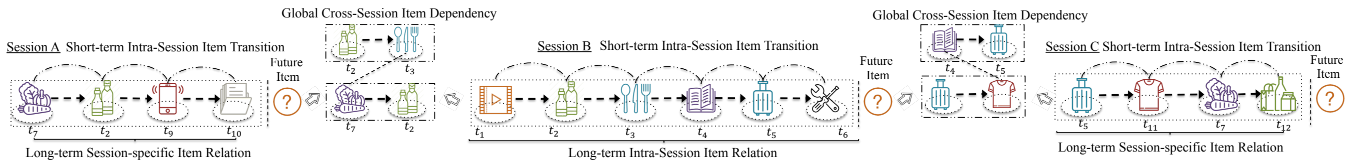


Figure 1: Illustrated example of session-based recommendation with multi-level transition dynamics.

rent session-based recommender systems only explore local contextual features, while the global item transitional patterns across exogenous sessions are neglected. This restricts the capabilities of current models in capturing the hierarchical transition signals for making recommendations.

While intuitively useful to perform the joint learning of item relation structures with multi-level transition dynamics, it is non-trivial to do it well. In particular, the item dependencies across different sessions can be complex. It is not necessary that a future interactive item is more relevant to items from a recent session that one that is further away (Kang and McAuley 2018). Hence, when tackling the cross-session item dependencies at various neighbor distances, the high-order relation structures exhibited with item transition patterns from a global perspective over all sessions, is necessary to be investigated in the relation embedding function. Additionally, intra-session item transition patterns vary by sessions. When modeling the time-evolving item correlation within a session, both the user's sequential behavior (short-term) and the overall session-specific purpose (long-term) should be taken into account (Liu, Zeng, and others 2018; Li, Wang, and McAuley 2020). Therefore, it is a significant challenge to jointly integrate the intra-session item correlations and inter-session item transition patterns into the recommendation framework in a fully adaptive manner.

Present Work. Motivated by the aforementioned challenges, we propose a new multi-task learning model with Multi-level Transition Dynamics (MTD) for session-based recommendation. In our MTD framework, we first devise a position-aware attention mechanism to jointly capture the intra-session sequential item transitions and session-specific main purchase with the incorporation of timestamp information. Specifically, we integrate a self-attention model with an attentive aggregation layer to capture the sequential transitional patterns of items within each individual session, without the rigid order assumption of user behavior (*i.e.*, latent states are propagated through temporally-ordered sequences in recurrent framework). To argument the representation learning ability over individual sessions, an attentive summarization layer is introduced to adaptively perform pattern aggregation. In the hierarchical attentive component, we also seek to explore the item positional information under a sequential encoding module to learn the influence of time factors. Additionally, inspired by the effectiveness of mutual information maximization in prioritizing global or local structural information in feature learning (Hjelm, Fedorov, and others 2019), we model the cross-session item dependencies in a hierarchical manner, *i.e.*, from item-level embedding learning to global graph-level representation. The developed hierarchically structured encoder via graphical mutual information maximization, endows the MTD with the capability to incorporate inter-session transitional signals from low-level to high-level across different sessions.

We highlight key contributions of this paper as follows:

- We exploit multi-level item transition dynamics in studying the session-based recommendation task. Towards this end, we propose a new recommendation framework which captures the item transition patterns, in the form of of short- and long-term intra-session item dependencies, as well as the cross-session item relation structures.
- We first develop a position-aware attentive mechanism to learn the evolving intra-session behavioral sequential signals and the summarized session-specific knowledge. Furthermore, a global context enhanced inter-session relation encoder is built upon the graphical neural network paradigm, to endow MTD for capturing the inter-session item-wise dependencies.
- Our extensive experiments on three real-world datasets demonstrate that MTD outperforms different types of baselines in yielding better recommendation results. Also, we show the efficiency of our developed model as compared to representative competitors and perform case studies with qualitative examples to investigate the interpretation capability of our MTD model.

Methodology

In this section, we present the technical details of our proposed recommendation framework MTD. We first formulate our studied session-based recommendation scenario as follows: Session-based recommendation aims to predict the next action of users based on their anonymous historical activity sequences (*e.g.*, clicks or purchases). Let $S = \{v_1, ..., v_m, ..., v_M\}$ denote the item candidate set, where M is the number of items. An anonymous session s is the temporally-ordered item sequence $s = [v_{s,1}, ..., v_{s,i}, ..., v_{s,I}]$, where $v_{s,i} \in S$ denotes the i-th item interested by the user in the session s, and I denotes the length of session s. The recommendation model outputs a list $Y = [y_1, y_2, ..., y_M]$ for each session s, where y_m denotes the probability that the next interacted item is v_m. We finally make recommendations based on the top-K ranked items in terms of their estimated probability values.

Intra-Session Item Relation Learning

To capture item transitional relationships within a session, we integrate two modules for learning the session-specific item transition patterns: (i) position-aware self-attention network for sequential transition modeling; (ii) attentive aggregation for session-specific knowledge representation.

Self-Attentive Item Embedding Layer. In MTD framework, we leverage the self-attention mechanism to learn the relevance scores over historical interested items within the session and draw the sequential contextual signals. Motivated by the attentive neural network in relation learning, self-attention mechanism has been proposed to tackle various sequence modeling tasks such as machine translation (Yang et al. 2019) and spatial-temporal data prediction (Wu et al. 2020)). Different from the standard attention module, self-attention could bring the benefits of capturing the relevance of past instances (*e.g.*, words or behaviors), and refine the representation process on the single sequence at various distance (Vaswani et al. 2017). Following the transformer network, we build the intra-session transition modeling layer upon the dot-product attention which consists of query, key and value dimensions. The weight matrices $\mathbf{W}_Q$, $\mathbf{W}_K$, $\mathbf{W}_V \in \mathbb{R}^{d\times d}$ respectively corresponds to the query, key, value vectors, to map initial item embeddings $\mathbf{E}_s \in \mathbb{R}^{I\times d}$ of session s into latent representations. The operations of self-attention network are defined as follows:

$$\begin{bmatrix}\mathbf{Q}\\\mathbf{K}\\\mathbf{V}\end{bmatrix} = \mathbf{E}_s \begin{bmatrix}\mathbf{W}_Q\\\mathbf{W}_K\\\mathbf{W}_V\end{bmatrix};\ \ \mathrm{Att}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \delta(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V} \tag{1}$$

where we define $\mathbf{X}_s \in \mathbb{R}^{I\times d} = \mathrm{Att}(\mathbf{Q},\mathbf{K},\mathbf{V})$ to represent the learned item embeddings with the modeling of pairwise relations between items $[v_{s,1}, ..., v_{s,i}, ..., v_{s,I}]$ in session s. $\delta(\cdot)$ denotes the softmax function and $\sqrt{d}$ is the scaling factor during the inner product operation.

We further enhance the self-attentive transition learning module with the modeling of non-linearities with the feed-forward network as shown below:

$$\widetilde{\mathbf{X}}_s = \mathrm{FFN}(\mathbf{X}_s) = \varphi(\mathbf{X}_s \cdot \mathbf{W}_1 + \mathbf{b}_1) \cdot \mathbf{W}_2 + \mathbf{b}_2 \tag{2}$$

we utilize $\varphi(\cdot)$=ReLU as the activation function and $\mathbf{W}_1$, $\mathbf{W}_2 \in \mathbb{R}^{d\times d}$ and $\mathbf{b}_1$, $\mathbf{b}_2 \in \mathbb{R}^d$ are trainable weight matrices and bias terms. After integrating the self-attention layer with the feed-forward network, we generate the embeddings $\widetilde{\mathbf{X}}_s \in \mathbb{R}^{I\times d}$ for all items $[v_{s,1}, ..., v_{s,I}]$ in each session.

Position-aware Item-wise Aggregation Module. We further design a position-aware attentive aggregation component to fuse the encoded item-wise relations for capturing the user main purpose within individual session s. We assign larger importance to the item states in which they have more contextual relations with future interested item. In particular, for the set of items in session s, we learn a set of weights $\{\alpha_1,...,\alpha_t,...,\alpha_I\}$ corresponding to the set of learned item embeddings $\widetilde{\mathbf{X}}_s = \{\mathbf{x}_{s,1}, ..., \mathbf{x}_{s,i}, ..., \mathbf{x}_{s,I}\}$. Formally, α_i is calculated as follows:

$$\alpha_i = \delta(\mathbf{g}^T \cdot \sigma(\mathbf{W}_3 \cdot \mathbf{x}_{s,I} + \mathbf{W}_4 \cdot \mathbf{x}_{s,i})) \tag{3}$$

where $\mathbf{g} \in \mathbb{R}^d$ is a linear projection vector for generating the weight scalar α_i. $\mathbf{W}_3, \mathbf{W}_4 \in \mathbb{R}^{d\times d}$. $\sigma(\cdot)$ and $\delta(\cdot)$ denotes the sigmoid and softmax function, respectively. The aggregated session representation as $\mathbf{x}_s^*$, *i.e.*, $\mathbf{x}_s^* = \sum_{i=1}^{I} \alpha_i \cdot \mathbf{x}_{s,i}$.

We further augment the intra-session item-wise fusion module with the injection of positional information, to capture the session-specific temporally-order signals of items. The dimensionality of positional representation is also set as d. This endows the modeling of relative positions with the incorporation of decay factor into linear transformations:

$$\mathbf{p}_s = \sum_{i=1}^{I} \omega_i \cdot \mathbf{x}_{s,i};\quad \omega_i =\propto \exp(|i - I| + 1) \tag{4}$$

where $\mathbf{p}_s$ denotes the fused representation with the preservation of relative positional information across different items. We construct a concatenated embedding for individual session of s as $\mathbf{q}_s = \mathbf{W}_c[\mathbf{x}_{s,I}, \mathbf{x}_s^*, \mathbf{p}_s]$, where $\mathbf{W}_c \in \mathbb{R}^{d\times 3d}$ performs the transformation operation. After that, following the implicit feedback-based recommendation paradigm in (He et al. 2020; Wang et al. 2019c), we utilize the inner product between $\mathbf{q}_s$ and embedding of item candidate $\mathbf{v}_m$ as $\mathbf{z}_m = \mathbf{q}_s^T\mathbf{v}_m$ and define our loss function of intra-session item relation learning with the cross-entropy as follows:

$$\mathcal{L}_{in} = -\sum_{n}^{N} \mathbf{y}_n \mathrm{log}(\tilde{\mathbf{y}}_n) + (1 - \mathbf{y}_n)\mathrm{log}(1 - \tilde{\mathbf{y}}_n) \tag{5}$$

where $\mathbf{y}_n$ denotes the ground truth label of n-th instance and $\tilde{\mathbf{y}}_n$ is the corresponding estimated result (*i.e.*, $\tilde{\mathbf{y}}_n = \delta(\mathbf{z}_n)$).

Global Transition Dynamics Modeling

To comprehensively capture the global cross-session transition dynamics among items, we develop a graph neural network architecture (as illustrated in Figure 2) to inject high-order dependent signals across different sessions into session representations. In particular, we first formulate a cross-session item graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ in which nodes $\mathcal{V}$ and $\mathcal{E}$ are generated from all historical sessions. Each session s can be regarded as a path which starts from $v_{s,1}$ and ends at $v_{s,I}$ in graph $\mathcal{G}$. The adjacent matrix $\mathcal{A}$ is constructed where each entry $a_{m,m'} = 1$ if there exists a transition relation from item v_m to $v_{m'}$ and $a_{m,m'} = 0$ otherwise.

We first propose a graph-structured message passing architecture to model the local context of transitional signals between different items. We formally define the corresponding encoding function as follows:

$$\mathbf{H}^{(l+1)} = \varphi(\mathbf{A}, \mathbf{H}^l\mathbf{W}^l) = \varphi(\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}\mathbf{H}^l\mathbf{W}^l) \tag{6}$$

where $\mathbf{H}^{(l+1)} \in \mathbb{R}^{M\times d}$ denotes the learned representations over items under the l-th propagation layer. With the aim of incorporating the self-propagated signals, we update the adjacent matrix with the summation of identify matrix $\mathbf{I}$ and the original adjacent matrix $\mathbf{A}$ as $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$. Then, we further apply the symmetric normalization strategy to conduct the information aggregation as: $\hat{\mathbf{D}}^{-\frac{1}{2}}\hat{\mathbf{A}}\hat{\mathbf{D}}^{-\frac{1}{2}}$, where $\hat{\mathbf{D}}$ is the diagonal node degree matrix of matrix $\mathbf{A}$.

Global Dependency Representation. After obtaining $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_m, ...\mathbf{h}_M\}$, we propose to capture the high-order global dependencies across correlated items from different sessions. Different from the session-based recommender system (Xu et al. 2020b) which replies on the random walk to generate item path, we leverage graph neural networks to consider the global item dependency across different sessions. Specifically, we first generate a fused graph-level emebdding with the aggregation function as: $\mathbf{z} = \tau(\mathbf{H})$

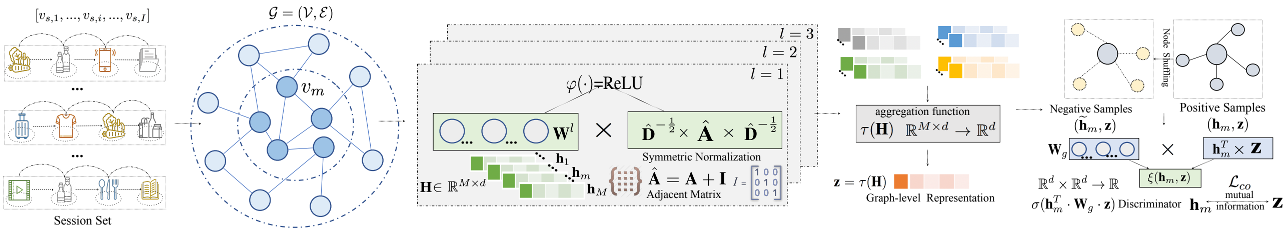


Figure 2: Global transition dynamics modeling

($\mathbb{R}^{M\times d} \rightarrow \mathbb{R}^d$), where $\tau(\cdot)$ denotes the mean pooling operation. Motivated by the paradigm of global feature representation with mutual information (Veličković et al. 2019a), we enhance our cross-session item relation encoder with the global context of the mutual information between patch-level embedding ($\mathbf{H}$) and graph-level representation $\mathbf{z}$.

We develop a classifier to perform the global dependency representation under the mutual information learning paradigm. It aims to differentiate positive $(\mathbf{h}_m, \mathbf{z})$ and negative instances $(\widetilde{\mathbf{h}}_m, \mathbf{z})$ in graph $\mathcal{G}$ by preserving the underlying cross-session item transition dynamics, the negative sample pair $(\widetilde{\mathbf{h}}_m, \mathbf{z})$ are generated by associating sampled item nodes with the fake embeddings based on the node shuffling strategy (Velickovic et al. 2019b). Then, both the positive and negative instances are fed into the classifier for classification task with the encoding function $\xi(\cdot)$:

$$\xi(\mathbf{h}_m, \mathbf{z}) = \sigma(\mathbf{h}_m^T \cdot \mathbf{W}_g \cdot \mathbf{z}); \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R} \tag{7}$$

where $\mathbf{W}_g \in \mathbb{R}^{d\times d}$ is the projection matrix. The classifier function outputs a probability score of the target node belongs to $\mathcal{G}$ given the corresponding embedding pair $(\mathbf{h}_m, \mathbf{z})$. The loss function of our graph-level global dependency representation component is defined as follows:

$$\mathcal{L}_{co} = -\frac{1}{N_{pos} + N_{neg}} \Big(\sum_{i=1}^{N_{pos}} \rho(\mathbf{h}_m, \mathbf{z}) \cdot log\xi(\mathbf{h}_m, \mathbf{z}) + \sum_{i=1}^{N_{neg}} \rho(\widetilde{\mathbf{h}}_m, \mathbf{z}) \cdot log[1 - \xi(\widetilde{\mathbf{h}}_m, \mathbf{z})] \Big) \tag{8}$$

where $\rho(\cdot)$ is an indicator function where $\lambda(\mathbf{h}_m, \mathbf{z}) = 1$ and $\rho(\widetilde{\mathbf{h}}_m, \mathbf{z}) = 1$ corresponds to positive and negative instance pairs, respectively. We define the number of positive and negative samples as N_{pos} and N_{neg}. By minimizing $\mathcal{L}_{co}$ (maximizing the mutual information between patch-level and graph-level representations), we could generate the enhanced user representations $\mathbf{H}^* \in \mathbb{R}^{M\times d}$ by encoding cross-session item transitional patterns from low-level (locally) to high-level (globally).

Learning Process of MTD

Based on the multi-task learning framework of MTD, we define our loss function with the integration of both intra- and inter-session transition dynamics as follows:

$$\mathcal{L} = \mathcal{L}_{cr} + \lambda_1 \mathcal{L}_{in} + \lambda_2 \|\Theta\|_2^2 \tag{9}$$

where Θ are learnable parameters. λ_1 and λ_2 balance the losses from two module and prevent over-fitting, respectively. Since the input of cross-session relation encoder and dual-stage attention network are different, we employ mini-batch Adam to optimize $\mathcal{L}_{in}$ and $\mathcal{L}_{cr}$ alternatively. We further define additional parameter f to denote the training frequency of $\mathcal{L}_{in}$ optimization for loss balance. In each epoch, we first optimize the graph-structured relation encoder and initialize the item representations with the current embeddings. Note that the patch representation $\mathbf{H}$, which are generated by the graph neural network, implies the global transition of items. To capture the global signal in recommendation module, we update the embedding table of items with $\mathbf{H}$ after the optimization step of $\mathcal{L}_{cr}$.

Complexity Analysis of MTD Framework. the intra-session item relation learning requires $O(I \times d^2 + I^2 \times d)$ calculations to compute the $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ and attentive embeddings $\mathbf{X}_s$ in the self-attention layer. After that, the rest of the intra-session learning spends most complexity on transformations in the d-dimensional hidden space (*e.g.* the two-layer feed-forward network), which costs $O(I \times d^2)$ complexity, and results in $O(L_1 \times I \times d^2 + I^2 \times d)$ overall complexity, where L_1 denotes the number of $d \times d$ transformations. Furthermore, the graph-based inter-session item transition modeling component requires $O(|\mathbf{A}| \times d + M \times d^2)$ complexity for message passing and embedding transformation, where $|\mathbf{A}|$ denotes the number of neighboring item pairs. Overall, the time complexity of MTD is comparable to most of current session-based recommender systems (*e.g.*, recurrent or graph neural networks), which is validated in our experiments through model efficiency investigation.

Evaluation

In this section, we perform extensive experiments on three publicly available real-life recommendation datasets and compare *MTD* with various state-of-the-art techniques. Particularly, we aim to answer the following research questions:

- **RQ1**: Does *MTD* consistently outperform other baselines by yeilding better recommendation performance?
- **RQ2**: How do different sub-modules in our *MTD* framework affect the recommendation performance?
- **RQ3**: What is the influence of hyperparameter settings in *MTD* for the model performance?
- **RQ4**: How is the model interpretation capability of *MTD*?
- **RQ5**: How is the scalability of the *MTD* method ?

Experimental Settings

Data Description. The data statistics with training/test detailed split settings are shown in Table 1.

Table 1: Statistics of the experimented datasets.

Dataset	Yoochoose	Diginetica	RetailRocket
# Train Sessions	369,859	719,470	433,648
# Test Sessions	55,400	60,858	15,132
# All Items	17,376	43,097	36,968
Average Length	6.15	5.13	9.93

Yoochoose Data[1]. This data comes from an online retailing size to logs half year of user clicks (released by Recsys'15 Challenge). Following the pre-processing strategies in (Li et al. 2017; Liu, Zeng, and others 2018), the sessions with the length of ≥ 2 and items with the appearing frequency of ≥ 5 are kept in the training and test (from the last day) set.

Diginetica Data[2]. This data is collected from the CIKM Cup 2016 which records the user clicks from the time period of six months. To be consistent with the settings in (Wu et al. 2019; Liu, Zeng, and others 2018), we do not include the sessions that contains single clicked item. Sessions in the test set are generated from the last week.

Retailrocket Data[3]. It contains the user browse data within six months from another e-commerce company. Following the same settings in (Xu et al. 2019), we filter out the items with the browsed frequency less than 5 and sessions with the length of less than 2. We set the data from the last week for test and the remaining part for training.

Evaluation Metrics. We leverage two metrics which are widely adopted in the session-based recommendation applications: **Precision@**K (Pre@K) and **Mean Reciprocal Rank@**K (MRR@K). Following the same rubric in (Wu et al. 2019; Li et al. 2017), MRR@K=0 if the first correctly recommended items is not among the top-K ranked items. Note that larger Pre@K and MRR@K scores indicate better performance.

Methods for Comparison. In our experiments, we consider the following baselines for performance comparison.
Frequency-based Recommendation Strategy.

- **POP**: it explores users' past interested items and makes recommendations with the identified most frequent items.
- **S-POP**: it recommends the most popular items to users by considering the their activities from the current session.

Neighboring Relation Modeling Algorithm.

- **ItemKNN** (Davidson et al. 2010): it considers the item correlations using k-nearest neighbors algorithm based on items' cosine similarity.

Recurrent Session-based Recommender system.

- **GRURec** (Hidasi et al. 2015): it is a representative session-based recommendation approach using the gated recurrent unit to encode the transitional regularities.

Attention-based Recommendation Frameworks.

- **NARM** (Li et al. 2017): it is a neural attention model to argument recurrent network for session representations, by attending deferentially to sequential items.
- **STAMP** (Liu, Zeng, and others 2018): it is attention model to capture user's temporal interests from historical clicks in a session.
- **SASRec** (Kang and McAuley 2018): this method is built upon the self-attention architecture to model the long-term item transition dynamics.

Session-based Recommendation with GNN.

- **SR-GNN** (Wu et al. 2019): it proposes a graph neural network model to encode item transitions within a session to generate its embedding.

Hybrid Session-based Recommendation Model.

- **CSRM** (Wang et al. 2019a): it integrates the inner memory encoder through an outer memory network by considering correlations between neighborhood sessions.
- **CoSAN** (Luo et al. 2020): it designs self-attention networks to model the collaborative feature information of items from neighborhood sessions.

Parameter Settings

Our implement is based on Tensorflow and conduct experiments on a GeForce GTX TITAN V GPU in Linux. In our experiments, the embedding dimensionality d is set as 100. We assign the regularization penalty $\lambda_2 = 10^{-6}$. All models are optimized using the Adam optimizer with the batch size and learning rate as 512 and $1e^{-3}$, respectively. The training frequency f in each epoch is set as 1, 4, 6 corresponding to the Yoochoose, Diginetica, Retailrocket, respectively. Furthermore, the dropout technique is applied in the training phase to alleviate the overfitting issue, with the ratio of 0.2. Furthermore, experiments of most baselines are conducted with their release source code.

Performance Validation (RQ1)

Superiority of *MTD* over Baselines. We present evaluation results of all methods in Table 2, and show performance of several representative baselines when varying the value of K in Table 3 (due to space limit). we can observe that *MTD* consistently outperforms other baselines in most cases, which justifies the effectiveness of our model in comprehensively capturing multi-level transition dynamics from intra-session and inter-session relations in a hierarchical manner.

Performance Gap among Baselines. The naive frequency (POP and S-POP) and similarity (ItemKNN) based recommendation approaches perform much worse than other baselines due to their limitations in capturing the dynamic sequential patterns of item transitions. Additionally, the attention-based recommendation techniques (NARM and STAMP) outperform the mere RNN approach (GRU4REC)–considering singular level of item relations. However, the significant improvement between *MTD* and attentive recommendation model suggests that only considering the intra-session item transitions is insufficient to fully capture the complex item transition dynamics from both local and global perspectives. While SR-GNN tries to encode the long-term item dependencies using the graph neural network, it still yields suboptimal performance because its failure in learning cross-session temporal signals.

[1]http://cikm2016.cs.iupui.edu/cikm-cup
[2]http://2015.recsyschallenge.com/challenge.html
[3]https://www.kaggle.com/retailrocket/ecommerce-dataset

Table 2: Recommendation performance comparison of all methods in terms of Pre@10 and MRR@10

Dataset	Metric	POP	S-POP	It-KNN	GRURec	NARM	STAMP	SASRec	SR-GNN	CSRM	CoSAN	*MTD*
Diginetica	Pre	0.58	20.66	26.46	20.31	36.72	37.05	38.42	38.40	38.56	37.58	**40.22**
	MRR	0.19	13.59	10.91	7.78	15.00	16.05	16.27	17.04	16.23	15.57	**17.58**
Yoochoose	Pre	4.59	28.61	43.40	55.13	60.19	58.79	60.42	60.84	60.46	61.01	**61.83**
	MRR	1.51	18.45	21.39	25.76	29.03	29.44	30.47	30.57	30.37	30.21	**30.83**
Retailrocket	Pre	1.59	29.67	21.41	31.01	44.74	43.14	46.39	44.88	47.21	45.83	**47.93**
	MRR	0.44	21.51	9.78	15.37	25.54	26.65	26.74	26.95	27.14	26.01	**28.51**

Table 3: Ranking performance with different K values.

Data	Metric	NARM	STAMP	SR-GNN	CSRM	CoSAN	*MTD*
Digi	Pre@5	24.80	25.72	27.15	26.38	25.72	**28.29**
	Pre@10	36.72	37.05	38.40	38.56	37.58	**40.22**
	Pre@20	50.32	49.86	51.57	52.56	50.94	**53.92**
Reta	Pre@5	36.25	36.45	37.38	38.65	37.07	**39.64**
	Pre@10	44.74	47.54	44.88	47.21	45.83	**47.93**
	Pre@20	52.58	55.56	52.27	55.04	54.87	**55.95**

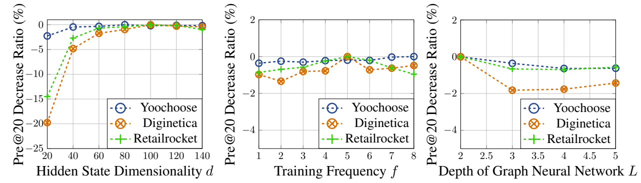


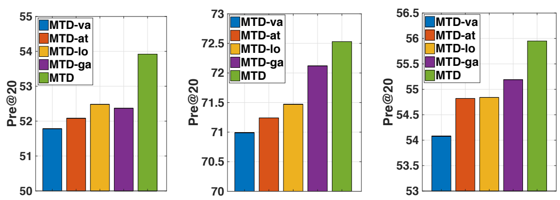
Figure 4: Hyper-parameter study of *MTD*.

Model Ablation and Effect Analyses (RQ2)

We consider several model variants to investigate the efficacy of key modules in the joint learning framework of *MTD*

Effect of Hierarchical Attention Network. We design two contrast models: i) *MTD*-va generates the session-level embeddings with the vanilla attention layer; ii) *MTD*-at further incorporates the temporal factor into the *MTD*-va method.

Effect of Cross-Session Dependency Encoder. i) *MTD*-lo only encodes the local-level item transition patterns without the cross-session dependency encoder; ii) *MTD*-ga replaces our graph-structured hierarchical relation encoder with the graph attention network operated on all relevant sessions.



(a) Diginetica (b) Yoochoose (c) Retailrocket

Figure 3: Model ablation study of *MTD*.

We report the results in Figure 3 and observe that *MTD* outperforms all other variants on all datasets in terms of $Pre@K$ and $MRR@K$ under $K = 20$, which justifies the effectiveness of the design of individual component in our *MTD* framework. In particular: (1) The performance gap among *MTD*-va, *MTD*-at, and *MTD*-lo illustrates the effectiveness of our time-aware hierarchical attention network in modeling the local item transitions. (2) Without the consideration of cross-session item dependencies, *MTD*-lo performs worse than *MTD*. It suggests the necessity of modeling the inter-session item correlations based on our developed graph-structured framework; (3) While the graph attention network (*MTD*-ga) could learn global-level item relations, it still falls behind *MTD* since it does not capture the hierarchical informativeness across relevant sessions.

Hyperparameter Study of MTD (RQ3)

We further investigate the hyperparameter sensitivity of our *MTD* (shown in Figure 4) and summarize the following observations. To save space and integrate results on different datasets with different performance scales into the one figure, we set y-axis as the performance degradation ratio compared to the best performance.

(1) **Effect of Hidden Dimensionality** d. The performance saturates as the hidden dimensionality d reaches around 100. This is because a larger dimensionality d brings a stronger representation ability at the early stage, but might lead to overfitting as the continuously increasing of d.

(2) **Impact of Training Frequency** f. We perform the training frequency study by varying f from 1 to 8, and could notice that a large value of f (≥ 5) will degrade the performance by misleading the objective function optimization.

(3) **Influence of Depth in Graph Neural Architecture**. Stacking more graph convolution layers with the adjacent matrix-based aggregation will more involve more redundant information of high-order connectivity, which hinders the learning process of global item relational structures in *MTD*. This observation also suggests the rationality of our designed graph neural component in simplifying and powering the cross-session item dependency learning, via the exploration of mutual relations between patch-level item embeddings and high-level graph representation.

Case Studies: Model Interpretation (RQ4)

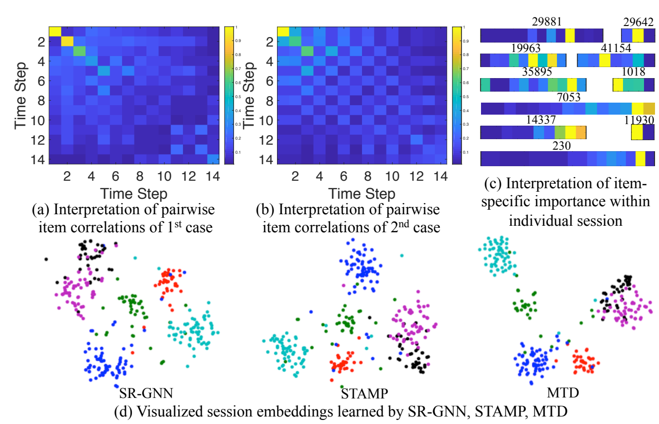


Figure 5: Case studies of *MTD* framework

Hierarchical Relation Interpretation across Items. We visualize the hierarchical item relations with quantitative weights learned from our time-aware hierarchical attention network on Diginetica. Figure 5 (a) and Figure 5 (b) show the encoded pairwise item correlations in modeling the intra-session sequential patterns of two sampled sessions across different time steps. From Figure 5 (c), we can observe that different items contribute differently to summarize the session-specific main purchase with hidden representations.

Visualizations of Learned Session Embeddings. We further visualize the projected session representations by our *MTD* and two state-of-the-arts: SR-GNN and STAMP (as shown in Figure 5 (d)). We randomly sample 180 session instances and label each one with its corresponding next clicked item (ground truth). It is easy to see that embeddings of sessions with the same label (6 classes and each one is represented with the same color) cluster closely and can be better distinguished by *MTD* as compared to other two methods, which demontrate the effectiveness of our learned item transitional patterns with session embeddings.

Model Scalability Study (RQ5)

Since efficiency is a key factor in many real-life recommendation applications, we finally investigate the computational cost (measured by running time of individual epoch) of our *MTD* and other state-of-the-art recommendation models. Our experiments are conducted on a GeForce GTX TITAN V GPU and results on three different datasets are summarized in Table 4. We can observe that *MTD* outperforms most competitive baselines with different deep neural network architectures (*e.g.*, attention mechanisms and graph-based message passing frameworks). Particularly, SR-GNN involves much computation cost in the gating mechanisms from neural network over each constructed session graph. Additionally, it is time-consuming to discover collaborative neighborhood sessions for each batch during the training phase of CSRM method. In the occasional cases that *MTD* miss the best performance (as compared to a streaming algorithm STAMP–only using attention mechanism for transition aggregation), *MTD* still achieves competitive model efficiency. Overall, the proposed *MTD* is efficient and scalable for large-scale session-based recommendation applications.

Table 4: Computational time cost (seconds) investigation.

Models	Yoochoose	Diginetica	RetailRocket
NARM	35	66	81
STAMP	9	24	14
SASRec	18	28	42
TiSA	82	160	100
SR-GNN	1401	2586	2502
CSRM	530	556	228
MTD	24	40	53

Related Work

Session-based Recommender Systems. To model sequential patterns of user behaviors, many recommender systems have been proposed to predict future interactions based on users' historical observations (Huang et al. 2019). In recent years, many session-based recommendation techniques have been developed based on various neural network architectures (Qiu et al. 2020a). Particularly, one intuitive approach is to apply the recurrent neural network (*e.g.*, GRU) for modeling the item sequential correlations (Hidasi et al. 2015). Furthermore, attention mechanisms have been adopted for pattern aggregation through relation weight learning, such as NARM (Li et al. 2017) and STAMP (Liu, Zeng, and others 2018). Another paradigm of session-based recommendation models lie in utilizing graph neural networks to capture the graph-structured item dependencies, such as attributed graph neural network for streaming recommendation (Qiu et al. 2020b) and graph-based message passing architectures (Wu et al. 2019). Different from the above work, our MTD framework aims to jointly captures the local and global item transitional signals in a hierarchical manner.

Graph Neural Networks for Recommendation. Recently emerged graph neural networks shine a light on performing information propagation over user-item graph for recommendation. Inspired by the graph convolution, several efforts have been devoted to capturing collaborative signals from the graph-based interacted neighbors, such as and LightGCN (He et al. 2020) and PinSage (Ying et al. 2018). Additionally, graph neural networks have also been integrated for recommendation to aggregate external knowledge from user side (Xu et al. 2020a) or item side (Wang et al. 2019b). Motivated by the success of graph neural networks, we propose to capture cross-session item dependencies in a hierarchical manner upon a global context enhanced graph network.

Conclusion

This work develops a new multi-task learning framework–MTD, which aims to inject multi-level transition dynamics into the session-based recommendation. By integrating a time-aware dual-stage attention network and graph hierarchical relation encoder, MTD not only models the intra-session sequential transitions, but also derives the high-order item relationships across long-range sessions. Experimental results on different real-world datasets show that MTD is superior to state-of-the-art baselines. In the future, we will incorporate item content information (*e.g.*, item text description or reviews) into MTD to deal with external attributes in learning semantic-aware item transitions.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and comments. This work is supported by National Nature Science Foundation of China (62072188, 61672241), Natural Science Foundation of Guangdong Province (2016A030308013), Science and Technology Program of Guangdong Province (2019A050510010). This work is also partially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the York Research Chairs (YRC) program.

References

[Al-Ghossein, Abdessalem, and Barré 2018] Al-Ghossein, M.; Abdessalem, T.; and Barré, A. 2018. Dynamic local models for online recommendation. In *WWW*, 1419–1423.

[Davidson et al. 2010] Davidson, J.; Liebald, B.; Liu, J.; Nandy, P.; Van Vleet, T.; et al. 2010. The youtube video recommendation system. In *Recsys*, 293–296. ACM.

[He et al. 2020] He, X.; Deng, K.; Wang, X.; Li, Y.; Zhang, Y.; and Wang, M. 2020. Lightgcn: Simplifying and powering graph convolution network for recommendation. *SIGIR*.

[Hidasi and Karatzoglou 2018] Hidasi, B., and Karatzoglou, A. 2018. Recurrent neural networks with top-k gains for session-based recommendations. In *CIKM*, 843–852.

[Hidasi et al. 2015] Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; and Tikk, D. 2015. Session-based recommendations with recurrent neural networks. In *ICLR*.

[Hjelm, Fedorov, and others 2019] Hjelm, R. D.; Fedorov, A.; et al. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

[Huang et al. 2004] Huang, X.; Peng, F.; An, A.; and Schuurmans, D. 2004. Dynamic web log session identification with statistical language models. *Journal of the American Society for Information Science and Technology* 55(14):1290–1303.

[Huang et al. 2019] Huang, C.; Wu, X.; Zhang, X.; Zhang, C.; et al. 2019. Online purchase prediction via multi-scale modeling of behavior dynamics. In *KDD*, 2613–2622.

[Kang and McAuley 2018] Kang, W.-C., and McAuley, J. 2018. Self-attentive sequential recommendation. In *ICDM*, 197–206. IEEE.

[Li et al. 2017] Li, J.; Ren, P.; Chen, Z.; Ren, Z.; Lian, T.; and Ma, J. 2017. Neural attentive session-based recommendation. In *CIKM*, 1419–1428. ACM.

[Li, Wang, and McAuley 2020] Li, J.; Wang, Y.; and McAuley, J. 2020. Time interval aware self-attention for sequential recommendation. In *WSDM*, 322–330.

[Liu, Zeng, and others 2018] Liu, Q.; Zeng, Y.; et al. 2018. Stamp: short-term attention/memory priority model for session-based recommendation. In *KDD*, 1831–1839. ACM.

[Luo et al. 2020] Luo, A.; Zhao, P.; Liu, Y.; Zhuang, F.; Wang, D.; Xu, J.; et al. 2020. Collaborative self-attention network for session-based recommendation. In *IJCAI*.

[Qiu et al. 2020a] Qiu, R.; Huang, Z.; Li, J.; and Yin, H. 2020a. Exploiting cross-session information for session-based recommendation with graph neural networks. *TOIS* 38(3):1–23.

[Qiu et al. 2020b] Qiu, R.; Yin, H.; Huang, Z.; et al. 2020b. Gag: Global attributed graph neural network for streaming session-based recommendation. In *SIGIR*, 669–678.

[Quadrana et al. 2017] Quadrana, M.; Karatzoglou, A.; Hidasi, B.; and Cremonesi, P. 2017. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Recsys*, 130–137.

[Ren et al. 2019] Ren, P.; Chen, Z.; Li, J.; Ren, Z.; Ma, J.; and de Rijke, M. 2019. Repeatnet: A repeat aware neural recommendation machine for session-based recommendation. In *AAAI*, volume 33, 4806–4813.

[Song et al. 2019] Song, K.; Ji, M.; Park, S.; and Moon, I.-C. 2019. Hierarchical context enabled recurrent neural network for recommendation. In *AAAI*, volume 33, 4983–4991.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; et al. 2017. Attention is all you need. In *NIPS*, 5998–6008.

[Veličković et al. 2019a] Veličković, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2019a. Deep graph infomax. In *ICLR*.

[Velickovic et al. 2019b] Velickovic, P.; Fedus, W.; Hamilton, W. L.; Liò, P.; Bengio, Y.; and Hjelm, R. D. 2019b. Deep graph infomax. In *ICLR*.

[Wang et al. 2019a] Wang, M.; Ren, P.; Mei, L.; Chen, Z.; Ma, J.; and de Rijke, M. 2019a. A collaborative session-based recommendation approach with parallel memory modules. In *SIGIR*, 345–354.

[Wang et al. 2019b] Wang, X.; He, X.; Cao, Y.; Liu, M.; and Chua, T.-S. 2019b. Kgat: Knowledge graph attention network for recommendation. In *KDD*, 950–958.

[Wang et al. 2019c] Wang, X.; He, X.; Wang, M.; Feng, F.; and Chua, T.-S. 2019c. Neural graph collaborative filtering. In *SIGIR*, 165–174.

[Wang et al. 2020] Wang, Z.; Wei, W.; Cong, G.; Li, X.-L.; Mao, X.-L.; and Qiu, M. 2020. Global context enhanced graph neural networks for session-based recommendation. In *SIGIR*, 169–178.

[Wu et al. 2019] Wu, S.; Tang, Y.; Zhu, Y.; Wang, L.; Xie, X.; and Tan, T. 2019. Session-based recommendation with graph neural networks. In *AAAI*, 346–353.

[Wu et al. 2020] Wu, X.; Huang, C.; Zhang, C.; et al. 2020. Hierarchically structured transformer networks for fine-grained spatial event forecasting. In *WWW*, 2320–2330.

[Xia et al. 2020] Xia, L.; Huang, C.; Xu, Y.; Dai, P.; Zhang, B.; and Bo, L. 2020. Multiplex behavioral relation learning for recommendation via memory augmented transformer network. In *SIGIR*, 2397–2406.

[Xu et al. 2019] Xu, C.; Zhao, P.; Liu, Y.; Sheng, V. S.; Xu, J.; Zhuang, F.; Fang, J.; and Zhou, X. 2019. Graph contextualized self-attention network for session-based recommendation. In *IJCAI*, 3940–3946.

[Xu et al. 2020a] Xu, H.; Huang, C.; Xu, Y.; Xia, L.; Xing, H.; et al. 2020a. Global context enhanced social recommendation with hierarchical graph neural networks. In *ICDM*.

[Xu et al. 2020b] Xu, Y.; Chen, J.; Huang, C.; Zhang, B.; et al. 2020b. Joint modeling of local and global behavior dynamics for session-based recommendation. In *ECAI*.

[Yang et al. 2019] Yang, B.; Wang, L.; Wong, D.; Chao, L. S.; and Tu, Z. 2019. Convolutional self-attention networks. In *AAAI*.

[Ying et al. 2018] Ying, R.; He, R.; Chen, K.; Eksombatchai, P.; Hamilton, W. L.; and Leskovec, J. 2018. Graph convolutional neural networks for web-scale recommender systems. In *KDD*, 974–983.

[Yuan et al. 2020] Yuan, F.; He, X.; Jiang, H.; Guo, G.; Xiong, J.; Xu, Z.; and Xiong, Y. 2020. Future data helps training: Modeling future contexts for session-based recommendation. In *WWW*, 303–313.