JCST Journal of Computer Science & Technology

Indexed in: SCIE, EI, INSPEC, CBST, DBLP, etc.

Sponsored by: ICT, CAS & CCF

Multimodal Interactive Network for Sequential Recommendation

Han Teng-Yue, Wang Peng-Fei, Niu Shao-Zhang

View online: http://doi.org/10.1007/s11390-022-1152-7

Articles you may be interested in

Sequential Recommendation via Cross-Domain Novelty Seeking Trait Mining

Fu-Zhen Zhuang, Ying-Min Zhou, Hao-Chao Ying, Fu-Zheng Zhang, Xiang Ao, Xing Xie, Qing He, Hui Xiong

Journal of Computer Science and Technology. 2020, 35(2): 305-319 http://doi.org/10.1007/s11390-020-9945-z

ATLRec: An Attentional Adversarial Transfer Learning Network for Cross-Domain Recommendation

Ying Li, Jia-Jie Xu, Peng-Peng Zhao, Jun-Hua Fang, Wei Chen, Lei Zhao

Journal of Computer Science and Technology. 2020, 35(4): 794-808 http://doi.org/10.1007/s11390-020-0314-8

Exploiting Pre-Trained Network Embeddings for Recommendations in Social Networks

Lei Guo, Yu-Fei Wen, Xin-Hua Wang

Journal of Computer Science and Technology. 2018, 33(4): 682-696 http://doi.org/10.1007/s11390-018-1849-9

Exploiting Structural and Temporal Influence for Dynamic Social-Aware Recommendation

Yang Liu, Zhi Li, Wei Huang, Tong Xu, En-Hong Chen

Journal of Computer Science and Technology. 2020, 35(2): 281-294 http://doi.org/10.1007/s11390-020-9956-9

Discovering Functional Organized Point of Interest Groups for Spatial Keyword Recommendation

Yan-Xia Xu, Wei Chen, Jia-Jie Xu, Zhi-Xu Li, Guan-Feng Liu, Lei Zhao

Journal of Computer Science and Technology. 2018, 33(4): 697-710 http://doi.org/10.1007/s11390-018-1850-3

Hashtag Recommendation Based on Multi-Features of Microblogs

Fei-Fei Kou, Jun-Ping Du, Cong-Xian Yang, Yan-Song Shi, Wan-Qiu Cui Mei-Yu Liang, Yue Geng

Journal of Computer Science and Technology. 2018, 33(4): 711-726 http://doi.org/10.1007/s11390-018-1851-2

Han TY, Wang PF, Niu SZ. Multimodal interactive network for sequential recommendation. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(4): 911-926 July 2023. DOI: 10.1007/s11390-022-1152-7

Multimodal Interactive Network for Sequential Recommendation

Teng-Yue Han[1] (韩滕跃), Peng-Fei Wang[2,*] (王鹏飞), and Shao-Zhang Niu[1] (牛少彰)

[1] *Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China*

[2] *School of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China*

E-mail: hantengyue@bupt.edu.cn; wangpengfei@bupt.edu.cn; szniu@bupt.edu.cn

Received November 12, 2020; accepted February 22, 2022.

Abstract Building an effective sequential recommendation system is still a challenging task due to limited interactions among users and items. Recent work has shown the effectiveness of incorporating textual or visual information into sequential recommendation to alleviate the data sparse problem. The data sparse problem now is attracting a lot of attention in both industry and academic community. However, considering interactions among modalities on a sequential scenario is an interesting yet challenging task because of multimodal heterogeneity. In this paper, we introduce a novel recommendation approach of considering both textual and visual information, namely Multimodal Interactive Network (MIN). The advantage of MIN lies in designing a learning framework to leverage the interactions among modalities from both the item level and the sequence level for building an efficient system. Firstly, an item-wise interactive layer based on the encoder-decoder mechanism is utilized to model the item-level interactions among modalities to select the informative information. Secondly, a sequence interactive layer based on the attention strategy is designed to capture the sequence-level preference of each modality. MIN seamlessly incorporates interactions among modalities from both the item level and the sequence level for sequential recommendation. It is the first time that interactions in each modality have been explicitly discussed and utilized in sequential recommenders. Experimental results on four real-world datasets show that our approach can significantly outperform all the baselines in sequential recommendation task.

Keywords multi-modality, interactive network, sequential recommendation

1 Introduction

The goal of sequential recommendation is to recommend the next item or the next few items based on users' sequential behaviors. It plays a central role in online shopping scenarios by sifting items from the huge corpora. With the ever prospering of neural networks, recent years have witnessed strong progress[1-3] on exploring interactions among users and items for a better recommendation. Inferring the auxiliary properties from multiple modalities is an important factor to improve the recommendation performance.

Naturally, the textual or visual information of an item plays an important role for the user to make a purchase decision. It is widely recognized that utilizing the auxiliary information would largely improve the recommendation accuracy. Following this line, the visual-content[4-6] and the language-content[7-9] models are two main modeling paradigms to explore interactions among multimedia contents to improve the recommendation accuracy. Concerning the unique value of each modality, some work[10-11] tried to integrate the multimodal data to improve the performance. Little work attempted to explore the relationships among modalities for sequential data.

To explain the relationships among modalities, we present an illustrative example in Fig.1. There is a sequence of items purchased by a user. From the item level we can infer the user's real intention at each step. For example, the user bought a "blouse" in the

Regular Paper

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61802029, U1536121, and 61370195.

*Corresponding Author

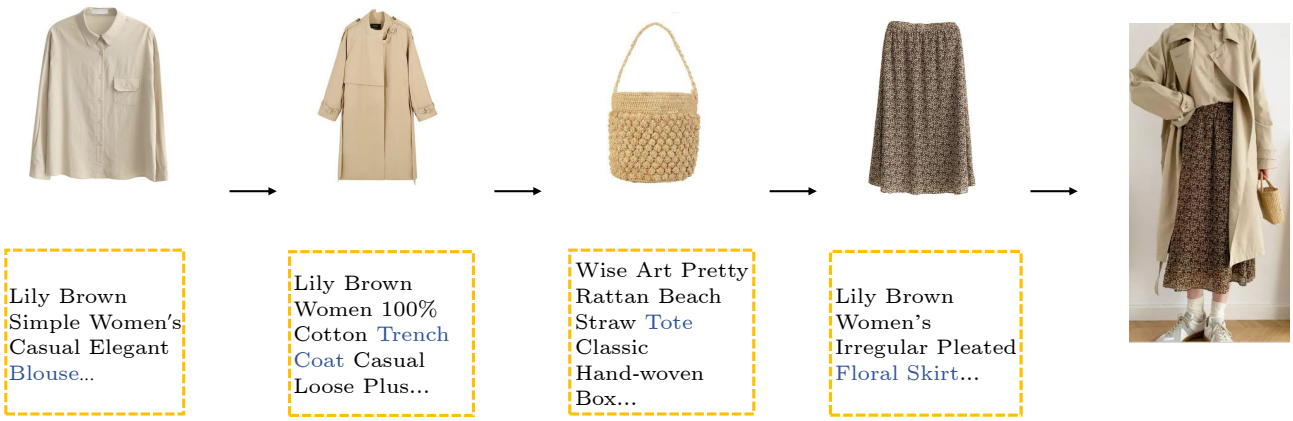

Fig.1. Example of a user's purchase sequence, where a suit of clothing in light color is purchased sequentially. Each item owns an image, the brand information, the title and a sentence (describing the products).

first time, and purchased a "trench coat" in the second time (in blue), while from the sequence level we know that the user aimed to buy a suit and preferred clothes with Lily Brown brand (from the textual sequence) and light color (from the visual sequence).

Modalities can provide complementary and mutually informative properties to describe users' preferences from both the item level and the sequence level. It is necessary to capture the item-level preferences and the sequence-level preferences from multiple modalities.

Although this seems to be a promising direction, it is non-trivial to model both the item-level interactions and the sequence-level interactions to facilitate recommendation due to the following challenges.

1) *Difficulties in Extracting a Union Preference from Each Modality.* How to extract these properties from a series of images and texts is a challenging task. In addition, for heterogeneous information, it is hard to define a unified extraction strategy to mine sequential properties.

2) *Difficulties in Integrating These Complicated Interactions from Different Levels for Recommendation.* Though features mined from both the item level and the sequence level are useful in revealing users' preferences, it is challenging to leverage these features and infer valuable information for a better recommendation.

3) *Difficulties in Utilizing the Intrinsic Data Correlations to Enhance the Data Representations.* Learning the correlations among the visual representation, the textual representation and the identity document (ID) representation for each item is a key factor to improve the performance. Thus, how to make up for sequential recommendation by utilizing multimodal information is still a challenging and unresolved task.

To address these issues, based on the item ID sequence, the textual sequence and the visual sequence, we propose Multimodal Interactive Network (MIN) to learn and leverage interactions among modalities from the item level and the sequence level for sequential recommendation. Specifically, MIN has two interactive layers to leverage interactions from different levels for a better recommendation. The item-wise interactive layer is designed to transform both visuals and texts of items to the same latent factor space to make them directly comparable. For each item an encoder-decoder mechanism is applied to match its image and text to select valuable information. The sequence-wise interactive layer applies a self-attention mechanism over each single modality to generate its sequential properties, and applies a cross-attention mechanism over multiple modalities (ID-visual and ID-text) to infer useful information. The item-wise interactions and the sequence-wise interactions are used for capturing the item-level preferences and the sequence-level preferences respectively. MIN finally concatenates the information to predict the successive items. Based on the integrated representations MIN can seamlessly incorporate with the classic sequential models. Effective data representation has been a key factor to improve the performances of existing models. Besides, we integrate two kinds of auxiliary tasks into the training of our network to enhance our model and improve the recommendation performance. One is a self-supervised task by maximizing the mutual information between the item representation and its visual (textual) representation. The item representation and its visual (textual) representation know little about each other. The other is the modality transformation task by optimizing the modal transformation loss. To

evaluate the proposed model, we construct some extensive experiments on four datasets by comparing it with several competitive baselines. Experimental results show that our model can significantly outperform all the baselines ranging from sequential approaches to multimodal approaches.

The contributions of this paper are summarized as follows.

• We systemically investigate the complex interactions among modalities for sequential recommendation.

• We propose a novel Multimodal Interactive Network (MIN) to employ multimodal information for sequential recommendation by integrating both the item-level interactions and the sequence-level interactions among modalities. Besides, we design two self-supervised tasks by maximizing the mutual information between the item representation and the visual (textual) representation to enhance the data representations.

• Experimental results show that our model can consistently outperform state-of-the-art baselines under different metrics for sequential recommendation.

The rest of the paper is organized as follows. Section 2 reviews the related work of sequential recommendation, multimedia recommendation and self-supervised learning. Section 3 reviews the preliminary knowledge. MIN is introduced in Section 4 and extensive experiments are conducted in Section 5. At last, we conclude the paper in Section 6.

2 Related Work

We briefly review three research areas related to our work: sequential recommendation, multimedia recommendation and self-supervised learning.

2.1 Sequential Recommendation

With the prosperity of neural networks, many studies try to design powerful sequential neural models for sequential recommendation. Among these models, recurrent neural network (RNN) based models[1–3, 12–14], convolutional neural network (CNN) based models[15, 16] and graph-based models[17–21] are three main approaches.

RNN-based models focus on exploiting sequential dependencies from interactions for recommendation. For this line, Hidasi *et al.*[1] and Yu *et al.*[13] proposed models that first apply RNN on sequential recommendation and demonstrated their effectiveness. Afterwards, a series of RNN-based models have been developed. Huang *et al.*[22] adopted a gated recurrent unit (GRU) component for capturing the sequential dependency and further incorporated knowledge graph for enhancing the modeling ability of attribute-level user preference. Kang and McAuley[23] introduced a novel self-attention based sequential approach to model the entire user sequence, and adaptively considered the consumed items for prediction. Sun *et al.*[24] used a bidirectional self-attention network to model users' sequential behaviors. Zhang *et al.*[25] integrated various heterogeneous features of items into feature sequences through a vanilla attention mechanism and then applied separated self-attention blocks on the item-level sequences and the feature-level sequences. Zhou *et al.*[26] utilized the intrinsic data correlations to derive the self-supervision signals and enhanced the data representations via pre-training methods for sequential recommendation. The CNN-based models have been recently introduced in the domain of sequential recommendation. For example, Tang and Wang[15] transformed a sequence of items into a 2-dimensional latent matrix and applied a convolution on it to model the stream-level features. Based on [27], He and Chua[28] integrated the future data into the model training to fill the gap between the historical and the future data. With the prosperity of graph neural networks, some new models were proposed. Wu *et al.*[17] modeled session sequences as graph structured data and captured the complex transitions of items based on the constructed session graphs. Wang *et al.*[29] adopted a hypergraph to represent the short-term item correlations and applied multiple convolutional layers to capture the semantic information behind items. Xia *et al.*[30] utilized dual channel hypergraph convolutional networks to model session-based data and integrated self-supervised learning into the framework to train the network.

2.2 Multimedia Recommendation

To alleviate the sparse problem, recent efforts have been made to leverage interactions among multiple modalities for recommendation. They can be roughly categorized into visual-content learning approaches[4, 5, 10, 31, 32] and text-context learning approaches[7, 9, 33-35].

Some work considered visual features for the task recommendation[10, 31, 32, 36]. Chen *et al.*[8] gave out

some visual explanations by applying an attention layer on the fine-grained visual preferences. Cui *et al.*[37] made sequential recommendation utilizing an RNN-based network by incorporating a multimodal representation with each item.

Another line is to explore the semantic information from textual reviews. For example, Li *et al.*[9] proposed a novel review-driven neural model to enhance the rich semantics from reviews for sequential recommendation. Cheng *et al.*[33] proposed an aspect-aware latent factor model for rating prediction by combining reviews and ratings effectively.

2.3 Self-Supervised Learning

Self-supervised learning is a new pattern of training networks on an auxiliary objective. The training signals are constructed by the correlations within the raw data[38]. Self-supervised learning has been introduced into several domains such as computer vision[39, 40] and natural language processing[41, 42].

For computer vision, Hjelm *et al.*[39] split the input data into multiple (possibly overlapping) views and maximized the mutual information between representations of these views. The views derived from other inputs were used as negative samples. For language modeling, Devlin *et al.*[41] learned to predict the next word or the next sentence given the previous sequences. The learned representations of words or sequences can improve the performance of downstream tasks.

3 Preliminary

In this part, we first introduce the symbols used in this paper, and then formalize the task for sequential recommendation. Furthermore, since our model is based on the classic transformer blocks[23], we briefly describe the self-attention mechanism used in the basic transformer blocks.

3.1 Notations

Let $\mathcal{U}$ and $\mathcal{I}$ denote the set of users and the set of items respectively. For modality information, we use $\mathcal{D}$ and $\mathcal{X}$ to denote the set of visuals and the set of texts respectively. For each user $u \in \mathcal{U}$, we use $i^u_{1:n} = \{i^u_1, i^u_2, \ldots, i^u_n\}$ to denote the ID sequence, and $v^u_{1:n} = \{v^u_1, v^u_2, \ldots, v^u_n\}$ and $t^u_{1:n} = \{t^u_1, t^u_2, \ldots, t^u_n\}$ to denote the visual sequence and the textual sequence respectively, where n represents the sequence length. We use i^u_k to represent the item that u has interacted with at the k-th time step, and v^u_k and t^u_k are the visual information and the textual information of item i^u_k respectively.

3.2 Task Definition

Based on above notations, our task of sequential recommendation is to learn the objection function to recommend the next $(n+1)$-th item that a user u will interact with at time t_{n+1}, given modalities of $i^u_{1:n}$, $v^u_{1:n}$, and $t^u_{1:n}$:

$$f(i^u_{1:n}, v^u_{1:n}, t^u_{1:n}) \rightarrow i^u_{n+1}.$$

3.3 Self-Attention Mechanism

The self-attention is a special attention mechanism that has been shown to be effective in various tasks[23]. Essentially the idea refers to using a content-based information extractor from a set of queries $\boldsymbol{Q}$, keys $\boldsymbol{K}$, and values $\boldsymbol{V}$. Given the input queries, keys, and values, the output is a weighted sum of the values according to a scaled dot-product attention:

$$Att(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V},$$

where d is the scaling factor, which is used to avoid overly large values of the inner product.

The advantage of self-attention is that it can keep the contextual sequential information and capture the relationships between elements in the sequence, regardless of their distance. In this paper, we use this mechanism to model the sequence-wise interactions among modalities.

4 Our Approach

We introduce the proposed model MIN in detail. The structure of MIN is shown in Fig.2.

The model consists of two modules for integrating interactions among modalities. 1) An item-wise interactive layer transforms each modality into the same latent factor space to make two modalities directly comparable. An encoder-decoder component is used to learn the interactions to match the relevant features between two distinct modalities. 2) A sequence-wise interactive layer extracts sequential properties from each modality according to the self-atten-

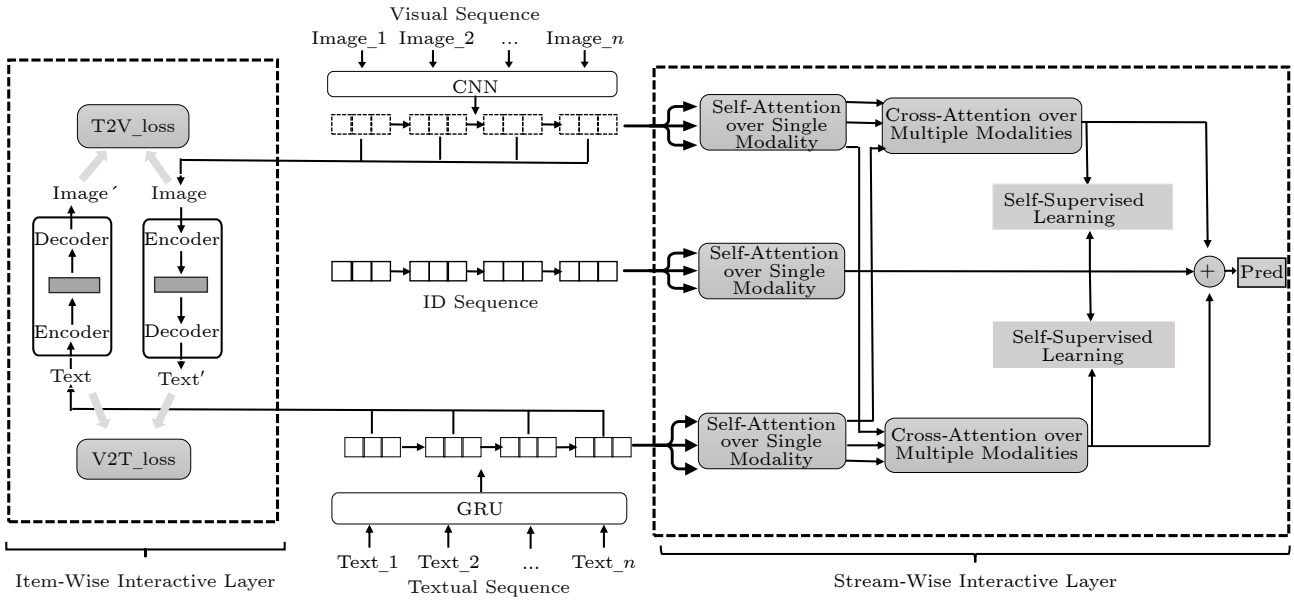

Fig.2. Over architecture of MIN. MIN contains two layers to fully integrate interactions among modalities. The item-wise interactive layer transforms modalities into a same embedding space. V2T and T2V are two encoder-decoder components, and they are applied on the visual and text of each item to model their interactions from the item level. The sequence-wise interactive layer first utilizes self-attention over single modality to obtain sequence-level properties from each sequence, and then cross-attention over multiple modalities is applied on the textual-visual pair and visual-textual pair respectively to infer relevant information from the sequence level. The relevant information is concatenated to predict the next item. Pred: Prediction.

tion mechanism and the cross-attention mechanism. MIN concatenates them to enrich the semantics of ID representations, and the enriched ID representations are used to predict the item that the user will buy. In the following part, we will present the design of each module.

For simplicity, we describe the approach for a single user u, and it is straightforward to extend the following formulas to a set of users. We drop the superscript of u in the notations for ease of reading.

4.1 Item-Wise Interactive Layer

As mentioned before, MIN totally considers three heterogeneous modalities for sequential recommendation. For the ID sequence, it is easy to handle it by mapping each item ID to a vector in a continuous space. Each item k is represented as a vector $\boldsymbol{e}_k$. For the visual sequence, we follow the previous work[4] to extract the visual feature of item k according to CNN models. The visual feature vector is denoted by $\boldsymbol{e}_{v_k}$. For the description text t_k, we adopt GRU to encode its word sequence, and use the final hidden state vector to represent the description, denoted by $\boldsymbol{e}_{t_k}$.

Based on the embedded representations over visuals and texts, we then apply two encoder-decoder components to align semantically similar concepts across the two modalities from the item level, denoted by V2T and T2V respectively. V2T encodes the visual modality to a latent space and then decodes this latent space into the textual modality. T2V encodes the textual modality to a latent space and then decodes this latent space into the visual space. As shown in Fig.2, the item-wise interactive layer is used to minor the gap between the two modalities. They are formalized as follows:

$$\hat{\boldsymbol{e}}_{t_k} = f_{\text{de}}^{\text{T2V}}(f_{\text{en}}^{\text{T2V}}(\boldsymbol{e}_{t_k})),$$
$$\hat{\boldsymbol{e}}_{v_k} = f_{\text{de}}^{\text{V2T}}(f_{\text{en}}^{\text{V2T}}(\boldsymbol{e}_{v_k})),$$

where functions $f_{\text{en}}^{\text{T2V}}(\cdot)$ and $f_{\text{en}}^{\text{V2T}}(\cdot)$ represent two encoders respectively, $f_{\text{de}}^{\text{T2V}}(\cdot)$ and $f_{\text{de}}^{\text{V2T}}(\cdot)$ represent two decoders respectively, and $\hat{\boldsymbol{e}}_{t_k}$ and $\hat{\boldsymbol{e}}_{v_k}$ are transformed representations from $\boldsymbol{e}_{t_k}$ and $\boldsymbol{e}_{v_k}$ respectively. According to the encoder-decoder component, we aim to match the relevant semantic features in different modalities. For simplicity, we use multilayer perceptron (MLP) components to model the encoders and the decoders.

Finally, we use the following functions to enhance the correlations between visuals and texts:

$$V2T\_loss_{\boldsymbol{e}_{t_k}} = \|\boldsymbol{e}_{t_k} - \hat{\boldsymbol{e}}_{t_k}\|^2,$$
$$T2V\_loss_{\boldsymbol{e}_{v_k}} = \|\boldsymbol{e}_{v_k} - \hat{\boldsymbol{e}}_{v_k}\|^2.$$

Then we use the following function to represent the item-level loss:

$$L_i = \sum_u \left(\sum_{t_k} V2T\_loss_{\boldsymbol{e}_{t_k}} + \sum_{v_k} T2V\_loss_{\boldsymbol{e}_{v_k}} \right).$$

As we can see, L_i can be regarded as an auxiliary task to compulsively model the interactions among modalities from the item level.

4.2 Sequence-Wise Interactive Layer

Based on the encoded representations of modalities, the sequence-wise interactive layer then models both intra-modality and cross-modality relationships from the sequence level. Specifically, it first extracts the sequence-level preference through the self-attention mechanism over each single modality, and further infers informative features according to the cross-attention mechanism over multiple modalities. In this way, sequential information is captured by the self-attention mechanism and the cross-attention mechanism. To explain this in detail, we plot Fig.3 to explain the sequential interactions captured by the self-attention mechanism and the cross-attention mechanism. Fig.3(a) and Fig.3(b) demonstrate interactions of sequential information on visual modality and textual modality respectively. Each item will interact with all the items in front of it. The specific process will be formalized in Subsection 4.2.1. Fig.3(c) and Fig.3(d) demonstrate the interactions of sequential information on visual-to-textual modality and textual-to-visual modality respectively, which will be formalized in Subsection 4.2.2. The visual information of each item will interact with the textual information of all items in front of it, and the visual information is query $\boldsymbol{Q}$. Symmetrically the textual information of each item will interact with the visual information of all items in front of it, and the textual information is query $\boldsymbol{Q}$. It is worth noting that the interactions are different because of different queries.

4.2.1 *Self-Attention over Single Modality*

The key point of the block is to assign a weight automatically on each element of the modality to derive the sequence-level representations. The interactions are shown in Fig.3(a) and Fig.3(b). Since the architecture of the self-attention mechanism over each modality is similar, here we only introduce the block from the ID perspective. Specifically, given an ID embedding matrix $\boldsymbol{E}_i = (\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \ldots, \boldsymbol{e}_{i_n})$, the computation rule of the self-attention mechanism is presented as follows:

$$\boldsymbol{E}_i^s = \left(\boldsymbol{e}_{i_1}^s, \ldots, \boldsymbol{e}_{i_n}^s\right) = Att(\boldsymbol{E}_i \boldsymbol{W}_s^q, \boldsymbol{E}_i \boldsymbol{W}_s^k, \boldsymbol{E}_i \boldsymbol{W}_s^v),$$

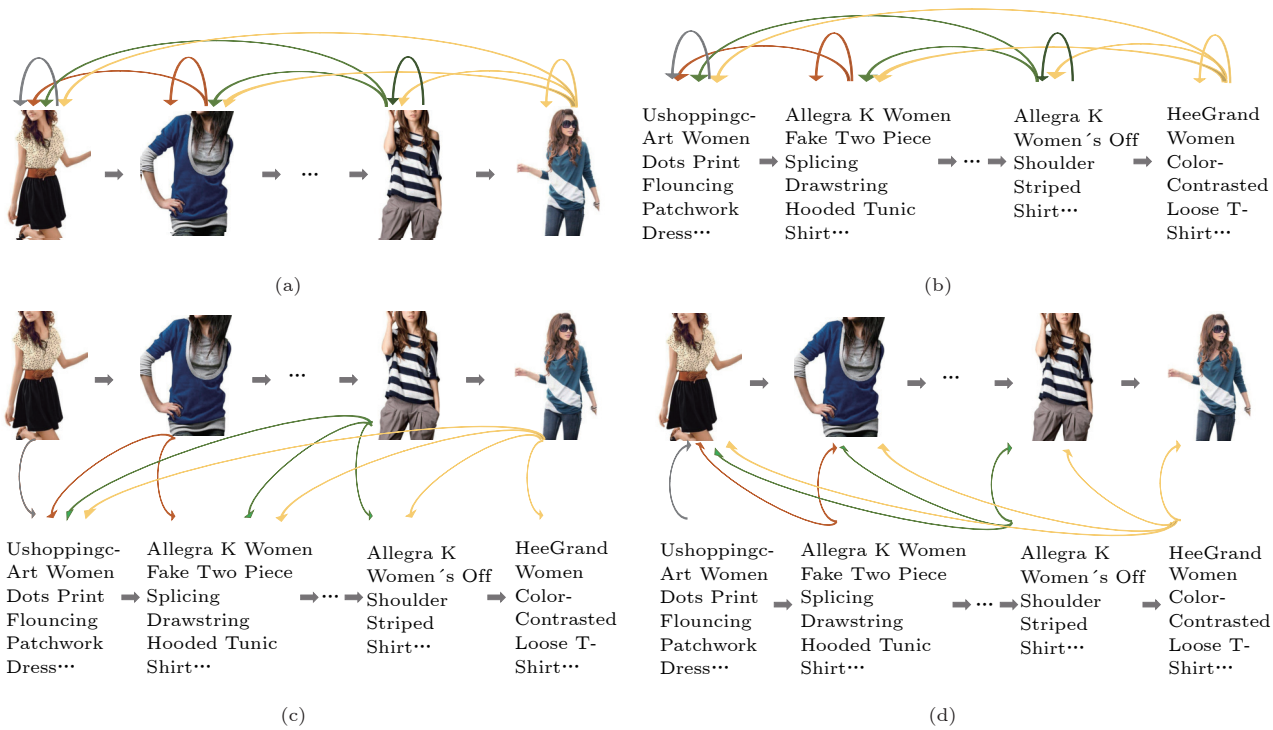

Fig.3. Explanations for the sequential interactions captured by the self-attention mechanism and the cross-attention mechanism. (a) Interactions of sequential information on visual modality. (b) Interactions of sequential information on textual modality. (c) Interactions of sequential information on visual-to-textual modality. (d) Interactions of sequential information on textual-to-visual modality.

where $\boldsymbol{W}_s^q$, $\boldsymbol{W}_s^k$ and $\boldsymbol{W}_s^v$ are three projection matrices respectively. $\boldsymbol{e}_{i_k}^s$ is the k-th column of $\boldsymbol{E}_i^s$. Similarly, we apply the self-attention mechanism on visual and textual sequences to obtain $\boldsymbol{E}_t^s$ and $\boldsymbol{E}_v^s$ respectively. By this way we can obtain the unique sequential patterns from each sequence.

4.2.2 Cross-Attention over Multiple Modalities

The purpose of this block is to infer the attentive semantics and the structure information from visual and textual sequences to enrich item representations for sequential recommendation. The interactions are shown in Fig.3(c) and Fig.3(d). Given textual and visual pairs, the functions are written as follows:

$$\begin{aligned}\boldsymbol{E}_{v\to t}^c &= \left[\boldsymbol{e}_{v_1\to t_1}^c, \ldots, \boldsymbol{e}_{v_n\to t_n}^c\right] \\ &= Att(\boldsymbol{E}_v^s\boldsymbol{W}_c^q, \boldsymbol{E}_t^s\boldsymbol{W}_c^k, \boldsymbol{E}_t^s\boldsymbol{W}_c^v), \\ \boldsymbol{E}_{t\to v}^c &= \left[\boldsymbol{e}_{t_1\to v_1}^c, \ldots, \boldsymbol{e}_{t_n\to v_n}^c\right] \\ &= Att(\boldsymbol{E}_t^s\boldsymbol{W}_c^q, \boldsymbol{E}_v^s\boldsymbol{W}_c^k, \boldsymbol{E}_v^s\boldsymbol{W}_c^v),\end{aligned}$$

where $\boldsymbol{e}_{v_k\to t_k}^c$ and $\boldsymbol{e}_{t_k\to v_k}^c$ represent the valuable representation queried from the visual and textual sequences respectively.

As we can see, in the self-attention mechanism over each single modality, we use a "homogeneous" query to model the intra-interactions among modalities. By assigning weights on the elements of sequences, we highlight the informative semantics from the sequence level, while in cross-attention over multiple modalities we utilize a "heterogeneous" query $\boldsymbol{E}_v^s\boldsymbol{W}_c^q$ to learn the inter-interactions to integrate diverse features from different modalities. Based on the queries that represent the real need of the ID sequences, the visual sequences and the textual sequences output the attentioned heterogeneous informative patterns. By this way our model is capable of inferring the useful cross-modal alignments from the visual sequences and the textual sequences respectively.

After obtaining the inferred features from the textual-visual sequence pair and the visual-textual sequence pair, the two kinds of features describe different perspectives for items, which can be complementary with each other. It is necessary to integrate these features together for a better recommendation. Hence, we fuse three representations for each item to derive the final comprehensive representations, denoted as:

$$\boldsymbol{e}_{i_k}^{\text{hybrid}} = \boldsymbol{e}_{i_k}^s \oplus \boldsymbol{e}_{v_k\to t_k}^c \oplus \boldsymbol{e}_{t_k\to v_k}^c, \tag{1}$$

where $\oplus$ is the concatenation operator. As we can see, our MIN is a very flexible model. We can apply any sequential recommenders based on the enhanced item representations, including the traditional collaborative filtering models or deep models. For simplicity, we output the probability of buying the next item:

$$P(i_{n+1}|i_{1:n}, t_{1:n}, v_{1:n}) = \frac{\exp(\boldsymbol{e}_{i_n}^{\text{hybrid}} \cdot \boldsymbol{e}_{i_{n+1}}^{\text{hybrid}})}{\sum_{k=1}^{|I|} \exp(\boldsymbol{e}_{i_n}^{\text{hybrid}} \cdot \boldsymbol{e}_{i_k}^{\text{hybrid}})}.$$

By considering all instances we obtain our learning approach as follows:

$$L_s = \sum_u \sum_{k=1}^{n} \log P(i_{k+1}|i_{1:k}, t_{1:k}, v_{1:k}).$$

However, the direct optimization of task L_s according to (2) has high computational complexity. Therefore, we adopt the negative sampling technique for efficient optimization, which approximates the original objective L_s with the following objective function:

$$\begin{aligned}L_s^{neg} = \sum_u \sum_{k=1}^{n} \Big(&\log \sigma(\boldsymbol{e}_{i_k}^{\text{hybrid}} \cdot \boldsymbol{e}_{i_{k+1}}^{\text{hybrid}}) + \\ &neg \times E_{i_{neg}\sim P_{\mathcal{I}}} \left(\log \sigma(-\boldsymbol{e}_{i_k}^{\text{hybrid}} \cdot \boldsymbol{e}_{i_{neg}}^{\text{hybrid}})\right) \Big),\end{aligned} \tag{2}$$

where $\sigma(x)$ is the sigmoid function $\sigma(x) = 1/(1+\mathrm{e}^{-x})$, neg is the number of "negative" samples, i_{neg} is the negative item sampled from $\mathcal{I}$, and $P_{\mathcal{I}}$ is the empirical distribution over all items.

4.3 Learning and Prediction

In this paper, we integrate two kinds of auxiliary tasks into the training of our network to enhance our model and improve the recommendation performance. One is the self-supervised task, which has two parts. One part is the mutual information maximization between the item representation and the visual representation. The other part is the mutual information maximization between the item representation and the textual representation. The other task is the modality transformation task by optimizing the modal transformation loss. The loss of modality transformation task L_i is introduced in Subsection 4.1. We will introduce the loss of the self-supervised learning.

In model MIN, we learn three kinds of representa-

tions of each item, namely item ID representations, item visual representations and item textual representations. Unfortunately, the item representation and the visual (textual) representation know little about each other. For the mini-batch item representation, the item visual (textual) representation can be the ground truth of the self-supervised learning. We adopt InfoNCE[38] with a standard binary cross-entropy loss between the samples from the ground truth and the corrupted samples as our learning objective:

$$\begin{aligned} & L_{\mathrm{ss}} \\ = & -\log \sigma(f(\boldsymbol{e}_{i_k}^s, \boldsymbol{e}_{v_k \to t_k}^c)) - \log \sigma(1 - f(\boldsymbol{e}_{i_{k'}}^s, \boldsymbol{e}_{v_{k'} \to t_{k'}}^c)) - \\ & \log \sigma(f(\boldsymbol{e}_{i_k}^s, \boldsymbol{e}_{t_k \to v_k}^c)) - \log \sigma(1 - f(\boldsymbol{e}_{i_{k'}}^s, \boldsymbol{e}_{t_{k'} \to v_{k'}}^c)). \end{aligned}$$

As we need to optimize both the item-level loss and the sequence-level loss, we obtain our final objective function as follows:

$$L_{\mathrm{MIN}} = L_s^{neg} + \alpha L_i + \beta L_{\mathrm{ss}} + \lambda \left\| \boldsymbol{\Theta} \right\|^2, \tag{3}$$

where λ is the regularization constant, and α and β are the hyperparameters to control the weight of L_i and L_{ss} respectively. $\boldsymbol{\Theta}$ represents all parameters that need to learn. We use a multi-task learning style to train the framework.

With the learned MIN, given a user u and his/her interaction sequence $i_{1:n}^u$, for each item we calculate the purchasing probability according to Subsection 4.2.2. We then rank the items according. their purchasing probabilities, and select the top N results as the final recommendations.

5 Experiments

We evaluate our proposed model focusing on the effects of interactions among modalities.

5.1 Experimental Setup

5.1.1 Datasets

We conduct experiments on the publicly available Amazon dataset[5] as it offers both the item visual information and the item textual information. The dataset fits our task. Considering the diversity of scales and species, we choose three categories from the Amazon dataset. They are Cell Phones & Accessories (Cell Phones), Sports and Outdoors (Sports), and Clothing, Shoes & Jewelry (Clothing). Besides, we use the MovieLens dataset[24]. We transform the users' ratings into implicit feedback data, in which each entry is marked as 1 indicating that a user has rated the item, and 0 otherwise. We extract the key frame of each trailer for each movie to extract the visual feature. We crawl the corresponding movie description.

Considering that in our model we utilize multiple modalities to improve the recommendation performance, we remove those items missing visuals or texts. Based on this, we follow the work[23] to filter unpopular items and inactive users with fewer than five records. The statistics of the four datasets are shown in Table 1.

Table 1. Statistics of Datasets for Experiments

Dataset	Number of Users	Number of Items	Number of Feedbacks
Cell Phones[5]	27 804	10 192	191 396
Sports[5]	35 597	18 267	295 091
Clothing[5]	39 386	23 010	278 406
MovieLens[24]	55 485	5 986	1 239 508

For each user, we sort her/his records according to the timestamp to form the interaction sequence. Based on the sorted sequences, we treat the last item of each sequence as the test data, and the second last item as the validation data. Specifically, for each user u we randomly sample 1 000 negative items to avoid heavy computation, and rank these items with the ground-truth item.

5.1.2 Evaluation Metrics

We provide a top-N recommendation list for each item in the testing set, where $N = 20$. For evaluation, we employ the commonly-used hit-ratio and NDCG[23] as our evaluation metrics. We perform significant tests using the paired t-test. Differences are considered statistically significant when the p-value is lower than 0.05. We repeat each experiment for five times, and the final results are the average of the five times.

5.1.3 Baselines

We use the following recommendation models to justify the effectiveness of our model, including the sequential recommenders and multimedia recommenders.

For sequential models, we concern both the shallow models and the deep models.

- *GRU4Rec*[1]. It is an RNN-based model which

uses GRU units and utilizes session-parallel mini-batches to make session-based recommendation.

• *Caser*[15]. Caser captures the stream-level sequential patterns by utilizing CNN on the adjacent items. Caser aims to model the relations among the history interactions.

• *SASRec*[23]. It is a self-attention based sequential model that captures the long-term semantics for recommendation.

• *BERT4Rec*[24]. It is a bidirectional self-attention network which learns useful sequential patterns to make recommendations.

• *FDSA*[25]. It is a self-attention network which integrates various heterogeneous features of items into the feature sequences.

• *S^3-Rec*[26]. It is a pre-trained model which utilizes the intrinsic data correlations to derive the self-supervision signals for sequential recommendation.

• *HyperRec*[29]. It is a hypergraph convolution network which adopts a hypergraph to represent the short-term item correlations and applies multiple convolutional layers to capture the semantic information behind items.

• *DHCN*[30]. It is a dual channel hypergraph convolutional network which models session-based data and integrates self-supervised learning into the framework.

For multimedia recommenders, we use the following methods.

• *VBPR*[4]. It is a visual bayesian personalized ranking model which is a well-known method based on visual features in the field of fashion recommendation.

• *JRL*[11]. JRL is a joint representation learning framework which incorporates heterogeneous information sources for recommendation.

• *MV-RNN*[37]. It is an RNN-based model which utilizes the visual information and the textual information to enhance the representation of items.

5.1.4 Parameter Settings

For baselines we utilize the recommended setting by their original work. For Caser[15] and SASRec[23], we use the codes released by their authors. The rest models are implemented in PyTorch. For MIN, we use Adam to train our model. When MIN achieves the best performance, the parameters are set as follows: the learning rate is set to 10^{-4}, the regularization coefficient is 10^{-4}, and the embedding size is set to 100. For the positional embedding used in the attention mechanism, we set the size to 50 on all four datasets. If the sequence length is greater than 50, we consider its recent 50 elements. If the sequence length is shorter than 50, we fill zero vectors for each sequence. For the self-attention block, the number of head is 1, and the hidden size of the point-wise feed forward layer is 512. The dropout ratio is set to 0.1.

5.2 Comparison Against Baselines

We compare our MIN model against several competitive baselines on the next-item recommendation task. We present the comparison results in Table 2. From this table, we have the following observations.

Table 2. Performance Comparison for Baselines and MIN for Next-Item Recommendation

Model	Cell Phones		Sports		MovieLens		Clothing	
	Hit-Ratio@20	NDCG@20	Hit-Ratio@20	NDCG@20	Hit-Ratio@20	NDCG@20	Hit-Ratio@20	NDCG@20
VBPR[4]	0.147 4	0.069 0	0.114 1	0.052 2	0.204 9	0.097 0	0.080 2	0.044 7
JRL[11]	0.181 0	0.091 4	0.137 1	0.061 6	0.249 0	0.116 3	0.106 4	0.046 9
GRU4Rec[1]	0.225 0	0.103 2	0.173 1	0.072 3	0.278 3	0.129 7	0.119 5	0.052 2
Caser[15]	0.247 9	0.108 4	0.198 4	0.085 7	0.297 2	0.136 9	0.124 2	0.056 9
MV-RNN[37]	0.268 1	0.113 7	0.216 4	0.088 3	0.319 4	0.145 2	0.136 8	0.063 7
SASRec[23]	0.261 9	0.116 0	0.215 0	0.093 8	0.325 9	0.152 8	0.151 5	0.061 0
BERT4Rec[24]	0.270 8	0.118 5	0.224 5	0.094 8	0.332 7	0.153 3	0.157 3	0.067 5
FDSA[25]	0.277 3	0.123 3	0.226 9	0.099 2	0.338 8	0.155 8	0.165 9	0.074 2
S^3-Rec[26]	0.280 6	0.124 9	0.232 8	0.105 5	0.344 7	0.155 9	0.172 4	0.076 0
HyperRec[29]	0.290 4	0.132 8	0.248 5	0.113 9	0.347 2	0.158 2	0.179 9	0.081 7
DHCN[30]	0.292 1*	0.138 4*	0.251 9*	0.116 7*	0.348 5*	0.164 0*	0.183 7*	0.082 2*
MIN	**0.314 9**	**0.149 5**	**0.264 7**	**0.124 5**	**0.364 8**	**0.172 6**	**0.206 7**	**0.093 1**
Improvement (%)	7.805 5	8.020 2	5.081 3	6.683 8	4.677 1	5.243 9	12.520 0	13.260 0

Note: Bolded numbers are the best performance of each column. The starred numbers represent the best baselines. The last row shows the improvement of our results against the best baseline. The improvement is significant at $p \leqslant 0.05$.

1) For sequential recommenders, we see that deep models GRU4Rec[1], Caser[15], and SASRec[23] achieve good performance. The result is consistent with the previous findings[14, 23].

We also see that SASRec performs better than GRU4Rec and Caser on four datasets. The underlying reason may be that: in the interaction sequence there are items irrelevant to the recommended item. These irrelevant items can be regarded as noises. SASRec uses a self-attention mechanism to assign weights on each element of the interaction sequence. By this way SASRec is capable of selecting the more informative features for sequential recommendation. BERT4Rec performs a little better than SASRec in most cases since it can capture right-to-left patterns in a sequence. However, it does not make a huge improvement. The underlying reason may be that the extra properties are not necessary when predicting the next item. FDSA and S^3-Rec integrate various heterogeneous features of items into feature sequences and can achieve a better performance than BERT4Rec. HyperRec and DHCN achieve better performances. It proves the graph convolutional network has powerful modeling ability in capturing the complex relations.

2) JRL designs a unified recommendation framework that fuses different types of information resources and obtains a better performance than VBPR. This observation is also consistent with our findings in the ablation study. By incorporating the visual information and the textual information with RNN networks, MV-RNN makes recommendations with representations including both latent and unified multimodal information. MV-RNN achieves a better performance than JRL because of its consideration on items' sequential properties.

3) Sequential recommenders perform better than multimedia recommenders in most cases. The underlying reason may be that the sequential pattern is more crucial for sequential recommendation. Fusing extra informative features is more capable of explaining interactions between users and items. The lack of capturing sequential patterns limits the performances of the baselines. This observation also reveals the necessity of mining interactions among modalities from the sequence level.

4) Finally by fusing both item visuals and texts into a unified framework, MIN models interactions among modalities from both the item level and the sequence level and further concatenates informative properties mined for sequential recommendation. MIN achieves the best performance among all the models on four datasets. Taking the Clothing, Shoes & Jewelry dataset as an example, we find when compared with the best baseline the performance improvement of MIN is around 12.52% and 13.26% on Hit-Ratio@20 and NDCG@20 respectively.

5.3 Ablation Study on Model MIN

MIN fuses two kinds of modalities by modeling the complex interactions for sequential recommendation. We conduct experiments to analyze the variants of MIN.

5.3.1 Item Level vs Sequence Level

One advantage of MIN is that it incorporates interactions among modalities from both the item level and the sequence level into ID representations for recommendation. We aim to analyze the performance of MIN when considering the interactions at different levels separately.

We first make some degeneration on MIN. Specifically, in the item-wise interactive layer we remove the encoder-decoder component to ignore modeling the connections across modalities from the item level. In this way we only concern the interactions among modalities from the sequential view and we name the new model MIN-i. In (3) MIN optimizes the loss function including the recommendation loss and the encoder-decoder loss. We remove the encoder-decoder loss L_i. To remove the sequence-wise interactions from MIN, we replace $\boldsymbol{e}^s_{i_n}$, $\boldsymbol{e}^c_{v_n \to t_n}$ and $\boldsymbol{e}^c_{t_n \to v_n}$ with $\boldsymbol{e}_{i_n}$, $\hat{\boldsymbol{e}}_{v_{i_n}}$ and $\hat{\boldsymbol{e}}_{t_{i_n}}$ respectively in (1). By this way we retain the item-wise interactions among modalities, while removing the self-attention over single modality and cross-attention over multiple modalities. We name the new model MIN-s. Fig.4 shows the performance comparisons among MIN and its two degraded models. From the results we have the following observations.

1) MIN-i performs better than MIN-s over all metrics on four datasets. This observation demonstrates the significance and necessity of considering sequential interactions among modalities for sequential recommendation. By capturing the sequential patterns the sequential recommenders can perform significantly better than other recommenders without sequential information.

2) Though less effective, the item-wise interactions among modalities are also valuable. MIN out-

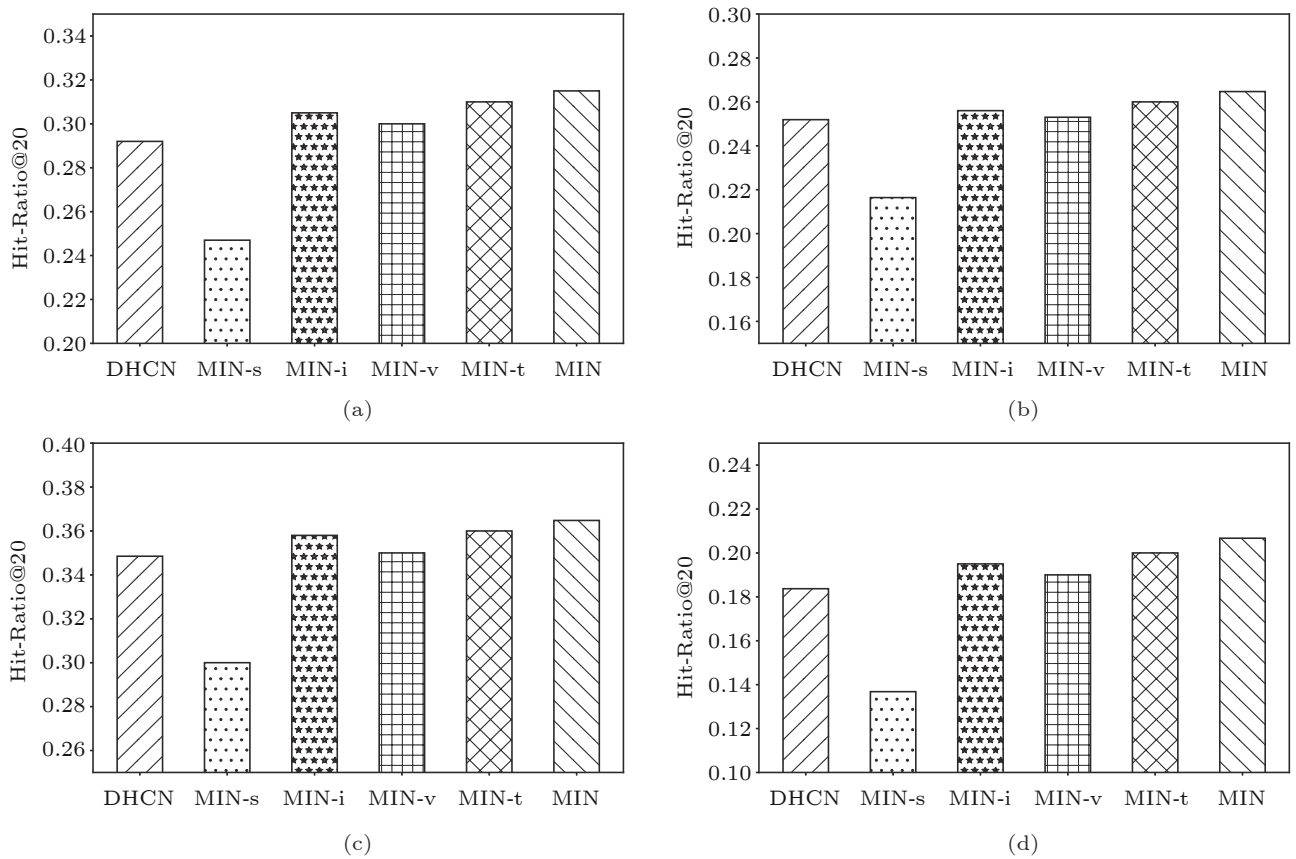

Fig.4. Ablation study on four datasets. MIN-x indicates that the corresponding module x is removed in the whole model. (a) Hit-Ratio@20 on Cell Phones. (b) Hit-Ratio@20 on Sports. (c) Hit-Ratio@20 on MovieLens. (d) Hit-Ratio@20 on Clothing.

performs MIN-i on all evaluation metrics. Taking the Clothing, Shoes & Jewelry dataset as an example, the performance improvement of MIN in terms of the absolute value is around 1.28% on Hit-Ratio@20 compared with MIN-i. It demonstrates the significance of analyzing the interactions among modalities systematically for sequential recommendation.

5.3.2 *Visual Sequence vs Textual Sequence*

As MIN totally utilizes two different informative modalities, we further analyze the benefits when considering item visuals and texts respectively.

For clear comparison we also make some degradation of MIN. Specifically, we remove the visual sequence from MIN. In this way we only concern the impact that the texts brought. We name the new model MIN-v. Similarly, we remove the textual sequence and rename the model MIN-t. We further compare the two single-modality models MIN-v and MIN-t. Fig.4 shows the performance comparisons of these models.

We find that there is no big difference between MIN-v and MIN-t. It indicates that the unique information existing in different modalities may help the recommendation. By fusing both visuals and texts into a unified framework, MIN obtains the best performance on all four datasets. It verifies the effectiveness of considering various types of information sources for sequential recommendation.

5.3.3 *Analysis of Loss Function*

We analyze the contribution of each auxiliary task. As described in (3), we adopt the multi-task training style to train our network. Generally speaking, multi-task training can improve the generalization of models. We conduct experiments on dataset Clothing, Shoes & Jewelry by using different loss functions. The results are shown in Table 3.

From the results in Table 3 we can get the following observations. 1) Multimodal transformation loss L_i and self-supervised learning loss L_{ss} do help improve the performance of the model. 2) Self-supervised learning task contributes a little more than the multimodal transformation task, since it fuses the vi-

Table 3. Performance Comparison for Different Loss Types on Dataset Clothing, Shoes & Jewelry

Loss	Hit-Ratio@20	NDCG@20
L_s^{neg}	0.176 8	0.081 2
$L_s^{neg} + L_i$	0.184 0	0.083 7
$L_s^{neg} + L_{ss}$	0.191 5	0.085 3
$L_s^{neg} + L_i + L_{ss}$	0.206 7	0.093 1

sual information, the textual information and the latent information.

5.4 Variant Methods for Multimodal Feature Interactions

In MIN, multimodal feature interactions are captured by the self-attention mechanism. We compare this mechanism with other methods. In this subsection, we design two variant methods to model sequential feature interactions, which are shown in Fig.5. As RNN is good at dealing with sequential problems, we design two RNN-based variant models to handle the multimodal feature interactions. In Fig.5(a), we name the method Parallel-LSTM, which handles the multimodal feature interactions separately. In Fig.5(b), we name the method Mixed-LSTM, which handles the multimodal feature interactions mixed. We also compare different actions on the final representation in (1) for our model MIN. We use MIN-dot to represent the variant model using dot product operation. We use MIN-add to represent the variant model using addition operation. In the model MIN, we use concatenation action. We conduct this experiment on dataset Clothing, Shoes & Jewelry. The results are shown in Table 4.

From Table 4 we have some conclusions. 1) Mixed-LSTM can capture sequential multimodal features better than Parallel-LSTM. 2) The self-attention mechanism has a strong ability in modeling complex interactions. 3) Besides, the concatenation action on the final preference representation achieves the best performance.

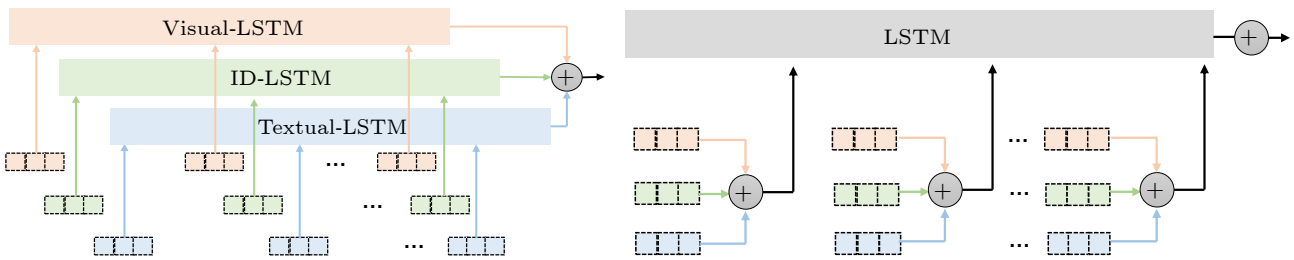

Fig.5. Structure of (a) Parallel-LSTM and (b) Mixed-LSTM.

Table 4. Performance Comparison for Variant Methods on Dataset Clothing, Shoes & Jewelry

Model	Hit-Ratio@20	NDCG@20
Parallel-LSTM	0.132 0	0.058 7
Mixed-LSTM	0.147 5	0.063 8
MIN-dot	0.189 2	0.084 4
MIN-add	0.193 5	0.088 2
MIN	0.206 7	0.093 1

5.5 Embedding Size Analysis

We study the effect of different embedding sizes in MIN. The hidden dimension can affect the modeling capability of the model. Specifically, we tune the embedding size from 10 to 100, and plot the results on Sports and Outdoors in Fig.6. Observations on other datasets are similar.

From the results we find that as the embedding size increases, the test performance in terms of NDCG@20 and Hit-Ratio@20 increases too. The trend is quite consistent over the four datasets. For sequential recommenders we find these models usually obtain stable performance with several tens or at most a hundred of latent factors. If we keep increasing the embedding size, there will be less performance improvement. Larger computational complexity may lead to over-fitting. This observation is quite consistent with the previous findings[11, 23].

Compared with other models, we see that MIN obtains a better performance at the same embedding size. The performance is also consistent with the conclusion in [37]. The reason may be that MIN considers more multimedia data for recommendation. This is why we adopt 50 as the default embedding size for MIN in the previous experiments.

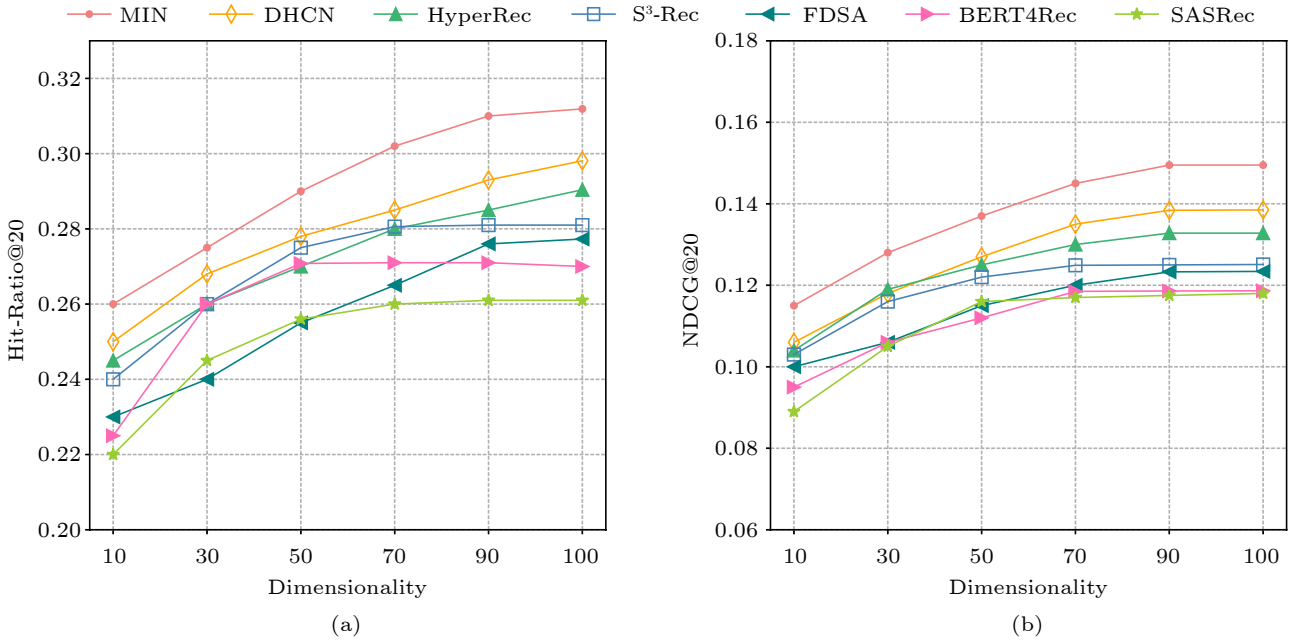

Fig.6. (a) Hit-Ratio@20 and (b) NDCG@20 for different embedding sizes with some methods on dataset Sports and Outdoors.

5.6 Model Complexity Analysis

We count the running time of the model on each dataset to evaluate the time complexity. MIN is a multimodal interactive network designed for capturing sequential feature interactions, which includes the visual modality and textual modality information. We record the training speed (the time taken for one epoch of training) on each dataset, and we conduct all experiments on NVIDIA-2080 Ti under the parameters setting described in Subsection 5.1.4.

According to the results, the training speed is closely related to the dataset. From the perspective of deep learning, the time of model training and prediction is acceptable. The results are listed in Table 5.

Table 5. Statistical Analysis of Training Speed

Dataset	Training Time (s)	Prediction Time (s)
Cell Phones	242	118
Sports	313	153
Clothing	341	167
MovieLens	962	251

5.7 Visualization Analysis

A core point is that we apply the self-attention mechanism to infer informative features from different modalities for improving sequential recommendation in MIN. In order to better understand why it is useful, we further construct qualitative analysis with a case study on dataset Clothing, Shoes & Jewelry in Fig.7.

Specially, we take a snapshot of the interaction sequence for a sample user. The interaction sequence consists of six items. The items are time-ordered. It is interesting to see that both the visual sequence and the textual sequence can offer informative properties for MIN to make a correct recommendation. The textual sequence indicates that the user buys some cuff-links and a tie bar of the same brand "Mrcuff". The user aims to buy a tie and a shirt of the same brand "Kenneth Cole reaction". The visual sequence shows his/her preference to the formal clothing style. By assigning the attention weights on each element of modalities, MIN can well capture these informative properties and recommend the dress shirt correctly.

6 Conclusions

In this work, we proposed Multimodal Interactive Network (MIN) by considering different modalities for sequential recommendation. MIN concerns the complex interactions among modalities from both the item level and the sequence level and further integrates semantic information for recommendation. In contrast to previous multimedia recommenders that mostly focus on the item-wise interactions, our work reveals the significant sequence-wise interactions and improves the recommendation performance. To the best of our knowledge, it is the first time that interac-

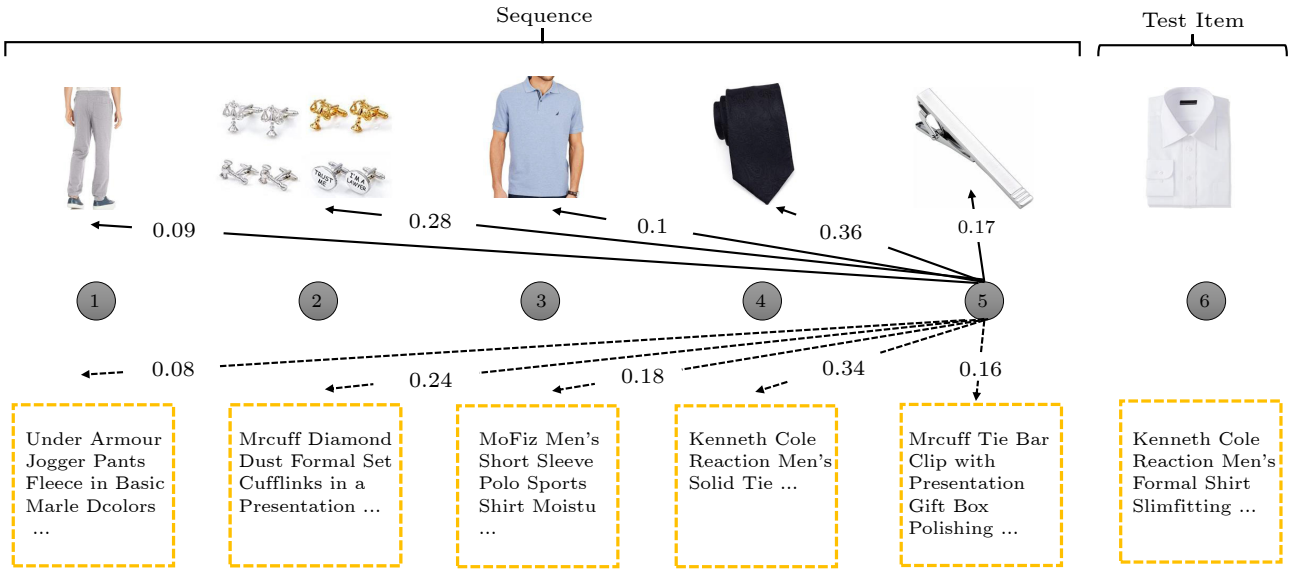

Fig.7. Visualization of attention weights on the visual and textual sequences. The bold fonts indicate the sequence-level preference of the user on each single-modality. Numbers are the attention weights obtained by MIN.

tions among modalities have been explicitly discussed and utilized in sequential recommendation. The multimodal preferences of users in sequence data can be well modeled. In the future, we will consider how to adaptively learn a better modality representation for sequential recommendation.

Conflict of Interest The authors declare that they have no conflict of interest.

References

[1] Hidasi B, Karatzoglou A, Baltrunas L, Tikk D. Session-based recommendations with recurrent neural networks. In *Proc. the 4th International Conference on Learning Representations*, May 2016.

[2] Quadrana M, Karatzoglou A, Hidasi B, Cremonesi P. Personalizing session-based recommendations with hierarchical recurrent neural networks. In *Proc. the 11th ACM Conference on Recommender Systems*, Aug. 2017, pp.130–137. DOI: 10.1145/3109859.3109896.

[3] Li J, Ren P J, Chen Z M, Ren Z C, Lian T, Ma J. Neural attentive session-based recommendation. In *Proc. the 2017 ACM on Conference on Information and Knowledge Management*, Nov. 2017, pp.1419–1428. DOI: 10.1145/3132847.3132926.

[4] He R J, McAuley J J. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proc. the 30th Conference on Artificial Intelligence*, Feb. 2016, pp.144–150.

[5] McAuley J J, Targett C, Shi Q F, Van Den Hengel A. Image-based recommendations on styles and substitutes. In *Proc. the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2015, pp.43–52. DOI: 10.1145/2766462.2767755.

[6] Lin Y J, Ren P J, Chen Z M, Ren Z C, Ma J, De Rijke M. Improving outfit recommendation with co-supervision of fashion generation. In *Proc. the 2019 the World Wide Web Conference*, May 2019, pp.1095–1105.. DOI: 10.1145/3308558.3313614.

[7] Bao Y, Fang H, Zhang J. TopicMF: Simultaneously exploiting ratings and reviews for recommendation. In *Proc. the 28th Conference on Artificial Intelligence*, Jul. 2014, pp.2–8.

[8] Chen X, Chen H X, Xu H T, Zhang Y F, Cao Y X, Qin Z, Zha H. Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation. In *Proc. the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2019, pp.765–774. DOI: 10.1145/3331184.3331254.

[9] Li C L, Niu X C, Luo X Y, Chen Z Z, Quan C. A Review-driven neural model for sequential recommendation. In *Proc. the 28th International Joint Conference on Artificial Intelligence*, Aug. 2019, pp.2866–2872. DOI: 10.24963/ijcai.2019/397.

[10] Chen J Y, Zhang H W, He X N, Nie L Q, Liu W, Chua T S. Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention. In *Proc. the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2017, pp.335–344. DOI: 10.1145/3077136.3080797.

[11] Zhang Y F, Ai Q Y, Chen X, Croft W B. Joint representation learning for top-N recommendation with heterogeneous information sources. In *Proc. the 2017 ACM on Conference on Information and Knowledge Management*, Nov. 2017, pp.1449–1458. DOI: 10.1145/3132847.3132892.

[12] Wang P F, Guo J F, Lan Y Y, Xu J, Wan S X, Cheng X Q. Learning hierarchical representation model for

nextbasket recommendation. In *Proc. the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2015, pp.403–412. DOI: 10.1145/2766462.2767694.

[13] Yu F, Liu Q, Wu S, Wang L, Tan T N. A dynamic recurrent model for next basket recommendation. In *Proc. the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2016, pp.729–732. DOI: 10.1145/2911451.2914683.

[14] Chen X, Xu H T, Zhang Y F, Tang J X, Cao Y X, Qin Z, Zha H Y. Sequential recommendation with user memory networks. In *Proc. the 11th ACM International Conference on Web Search and Data Mining*, Feb. 2018, pp.108–116. DOI: 10.1145/3159652.3159668.

[15] Tang J X, Wang K. Personalized top-N sequential recommendation via convolutional sequence embedding. In *Proc. the 11th ACM International Conference on Web Search and Data Mining*, Feb. 2018, pp.565–573. DOI: 10.1145/3159652.3159656.

[16] Yuan F J, Karatzoglou A, Arapakis I, Jose J M, He X N. A simple convolutional generative network for next item recommendation. In *Proc. the 12th ACM International Conference on Web Search and Data Mining*, Feb. 2019, pp.582–590. DOI: 10.1145/3289600.3290975.

[17] Wu S, Tang Y Y, Zhu Y Q, Wang L, Xie X, Tan T N. Session-based recommendation with graph neural networks. In *Proc. the 33rd Conference on Artificial Intelligence*, Jan. 2019, pp.346–353. DOI: 10.1609/aaai.v33i01.3301346.

[18] Qiu R H, Li J J, Huang Z, Yin H Z. Rethinking the item order in session-based recommendation with graph neural networks. In *Proc. the 28th ACM International Conference on Information and Knowledge Management*, Nov. 2019, pp.579–588. DOI: 10.1145/3357384.3358010.

[19] Qiu R H, Huang Z, Li J J, Yin H Z. Exploiting cross-session information for session-based recommendation with graph neural networks. *ACM Transactions on Information Systems*, 2020, 38(3): Article No. 22. DOI: 10.1145/3382764.

[20] Qiu R H, Yin H Z, Huang Z, Chen T. GAG: Global attributed graph neural network for streaming session-based recommendation. In *Proc. the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp.669–678. DOI: 10.1145/3397271.3401109.

[21] Guo L, Yin H Z, Wang Q Y, Chen T, Zhou A, Hung N Q V. Streaming session-based recommendation. In *Proc. the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2019, pp.1569–1577. DOI: 10.1145/3292500.3330839.

[22] Huang J, Zhao W X, Dou H J, Wen J R, Chang E Y. Improving sequential recommendation with knowledge-enhanced memory networks. In *Proc. the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, Jul. 2018, pp.505–514. DOI: 10.1145/3209978.3210017.

[23] Kang W C, McAuley J J. Self-attentive sequential recommendation. In *Proc. the 2018 IEEE International Conference on Data Mining*, Nov. 2018, pp.197–206. DOI: 10.1109/ICDM.2018.00035.

[24] Sun F, Liu J, Wu J, Pei C H, Lin X, Ou W W, Jiang P. BERT4Rec: Sequential recommendation with bidirectional encoder representations from transformer. In *Proc. the 28th ACM International Conference on Information and Knowledge Management*, Nov. 2019, pp.1441–1450. DOI: 10.1145/3357384.3357895.

[25] Zhang T T, Zhao P P, Liu Y C, Sheng V S, Xu J J, Wang D W, Liu G F, Zhou X F. Feature-level deeper self-attention network for sequential recommendation. In *Proc. the 28th International Joint Conference on Artificial Intelligence*, Aug. 2019, pp.4320–4326. DOI: 10.24963/ijcai.2019/600.

[26] Zhou K, Wang H, Zhao W X, Zhu Y T, Wang S R, Zhang F Z, Wang Z Y, Wen J R. S^3-Rec: Self-supervised learning for sequential recommendation with mutual information maximization. In *Proc. the 29th ACM International Conference on Information & Knowledge Management*, Oct. 2020, pp.1893–1902. DOI: 10.1145/3340531.3411954.

[27] Rendle S, Freudenthaler C, Schmidt-Thieme L. Factorizing personalized Markov chains for next-basket recommendation. In *Proc. the 19th International Conference on World Wide Web*, Apr. 2010, pp.811–820. DOI: 10.1145/1772690.1772773.

[28] He X N, Chua T S. Neural factorization machines for sparse predictive analytics. In *Proc. the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Aug. 2017, pp.355–364. DOI: 10.1145/3077136.3080777.

[29] Wang J L, Ding K Z, Hong L J, Liu H, Caverlee J. Next-item recommendation with sequential hypergraphs. In *Proc. the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, Jul. 2020, pp.1101–1110. DOI: 10.1145/3397271.3401133.

[30] Xia X, Yin H Z, Yu J L, Wang Q Y, Cui L Z, Zhang X L. Self-supervised hypergraph convolutional networks for session-based recommendation. In *Proc. the 35th AAAI Conference on Artificial Intelligence*, Feb. 2021, pp.4503–4511. DOI: 10.1609/aaai.v35i5.16578.

[31] Han X T, Wu Z X, Jiang Y G, Davis L S. Learning fashion compatibility with bidirectional LSTMs. In *Proc. the 25th ACM International Conference on Multimedia*, Oct. 2017, pp.1078–1086. DOI: 10.1145/3123266.3123394.

[32] Song X M, Feng F L, Liu J H, Li Z K, Nie L Q, Ma J. NeuroStylist: Neural compatibility modeling for clothing matching. In *Proc. the 25th ACM International Conference on Multimedia*, Oct. 2017, pp.753–761. DOI: 10.1145/3123266.3123314.

[33] Cheng Z Y, Ding Y, Zhu L, Kankanhalli M S. Aspect-aware latent factor model: Rating prediction with ratings and reviews. In *Proc. the 2018 World Wide Web Conference*, Apr. 2018, pp.639–648. DOI: 10.1145/3178876.3186

145.

[34] He X N, Chen T, Kan M Y, Chen X. TriRank: Review-aware explainable recommendation by modeling aspects. In *Proc. the 24th ACM International on Conference on Information and Knowledge Management*, Oct. 2015, pp. 1661–1670. DOI: 10.1145/2806416.2806504.

[35] Zheng L, Noroozi V, Yu P S. Joint deep modeling of users and items using reviews for recommendation. In *Proc. the 10th ACM International Conference on Web Search and Data Mining*, Feb. 2017, pp.425–434. DOI: 10.1145/3018661.3018665.

[36] Kang W C, Fang C, Wang Z W, McAuley J. Visually-aware fashion recommendation and design with generative image models. In *Proc. the 2017 IEEE International Conference on Data Mining*, Nov. 2017, pp.207–216. DOI: 10.1109/ICDM.2017.30.

[37] Cui Q, Wu S, Liu Q, Zhong W, Wang L. MV-RNN: A multi-view recurrent neural network for sequential recommendation. *IEEE Trans. Knowledge and Data Engineering*, 2020, 32(2): 317-331. DOI: 10.1109/TKDE.2018.2881260.

[38] Van Den Oord A, Li Y Z, Vinyals O. Representation learning with contrastive predictive coding. arXiv: 1807.03748, 2019. https://arxiv.org/abs/1807.03748, Jul. 2023.

[39] Hjelm R D, Fedorov A, Lavoie-Marchildon S, Grewal K, Bachman P, Trischler A, Bengio Y. Learning deep representations by mutual information estimation and maximization. In *Proc. the 7th International Conference on Learning Representations*, May 2019.

[40] Zhang R, Isola P, Efros A A. Colorful image colorization. In *Proc. the 14th European Conference on Computer Vision*, Oct. 2016, pp.649–666. DOI: 10.1007/978-3-319-46487-9_40.

[41] Devlin J, Chang M W, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Jun. 2019, pp.4171–4186. DOI: 10.18653/v1/n19-1423.

[42] Kong L P, D'Autume C D M, Yu L, Ling W, Dai Z H, Yogatama D. A mutual information maximization perspective of language representation learning. In *Proc. the 8th International Conference on Learning Representations*, Apr. 2020.

Teng-Yue Han received her B.S. and M.S. degrees in mathematics and applied mathematics from Hebei Normal University, Shijiazhuang, in 2015 and 2018 respectively. Now she is currently pursuing her Ph.D. degree in computer science and technology from Beijing University of Posts and Telecommunications, Beijing. Her research interests include data mining, deep learning, multi-modality, and recommendation.

Peng-Fei Wang received his B.S. degree in software engineering from the Xidian University, Xi'an, in 2008, his M.S. degree in software engineering from the Beihang University, Beijing, in 2011, and his Ph.D. degree in computer software theory from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, in 2017. He joined the Beijing University of Posts and Telecommunications, Beijing, as an associate professor in 2017. His research interests include data mining, machine learning and recommendation. He has published more than 30 papers in refereed journals and conferences.

Shao-Zhang Niu received his B.S. and M.S. degrees in mathematical science from Beijing Normal University, Beijing, in 1985 and 1988 respectively, and his Ph.D. degree in information science from Beijing University of Posts and Telecommunications, Beijing, in 2004. Now, he is a professor of School of Computer Science, Beijing University of Posts and Telecommunications, Beijing. His research interests include big data analysis, steganography and digital forensics.