Pattern Recognition Letters 152 (2021) 10–17

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Multimodal-adaptive hierarchical network for multimedia sequential recommendation

Tengyue Han[a], Shaozhang Niu[a,*], Pengfei Wang[b]

[a] Beijing Key Lab of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing, 100876, China
[b] School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, 100876, China

ARTICLE INFO

Article history:
Received 1 December 2020
Revised 24 June 2021
Accepted 23 August 2021
Available online 14 September 2021

Edited by Prof. S. Sarkar

2008 MSC:
41A05
41A10
65D05
65D17

Keywords:
Multimedia
Multimodal
Sequential recommendation
Multimodal-adaptive

ABSTRACT

Recommender system has a pivotal role in electronic economy especially for the online shopping platforms. Studies over the past two decades have proved that exploiting the inherent properties of items contributes a lot to the accuracy of multimedia sequential recommendation. There is no doubt that multimedia information including images and texts of a product have an impact on user's purchase decision. However, modeling user's dynamic preferences for multimodal (visual and textual in this paper) information over time is still a challenging problem. To solve this problem, we propose a Multimodal-Adaptive Hierarchical Network (MAHN for short) for multimedia sequential recommendation, which includes a hierarchical recurrent neural network and an information modulation module between the hierarchical structure. Specifically, the hierarchical recurrent neural network achieves the re-selection of multimodal information from the first layer to the second layer, the information modulation module realizes the selection of each modal information at time step t based on the previous time steps. Finally, to improve the generalization ability of our model, we adopt the multi-task training style to jointly optimize BPR loss and reconstruction loss of multimodal information. Experiments are conducted on two real world public datasets, and the results demonstrate that our model outperforms the other methods.

1. Introduction

There is a growing body of literature that recognizes the importance of multimedia recommendation especially for online shopping platforms in recent years. Giving the user's historical behavior records, recommender system aims to predict the next commodity that the user will interact with. It has been proved that incorporating external knowledge to recommender system can contribute to improving the performance of recommendation ([14]). With the significant achievements made in multimodal machine learning domain, recent trends in cross-domain learning have led to a proliferation of studies that integrate different kinds of information from each modality to improve the performance in image and video captioning ([15,18]), visual QA ([7,26]), text-to-image generation ([20,25]) and recommender systems ([6]). It is promising to analyse the interactions in multimodal information for sequential recommendation. On the one hand, incorporating multimodal information of items can alleviate the sparse problem which is caused by the limited explicit interactions; on the other hand, capturing sequential pattern of multimodal information can model the user's dynamic modality-specific preference, for example, color and style (visual-modality), brand and price (textual-modality).

Many studies have explored the effects of incorporating multimedia information such as images and texts to recommender system. Prior works in recommendation can be divided into three main categories coarsely, visual-based methods ([8,19]), textual-based methods ([1,5]) and hybrid-based methods ([6]). Some of these works try to enrich the representations of items by extracting their image features or summarizing the salient properties from reviews. Some other works try to model the user's multiple preference including visual preference or textual preference by generalizing all items that the user has interacted with. There is no doubt that these works are helpful to improve the performance of the recommender system. Despite the success of the previous studies, capturing user's dynamic multimodal preference in sequential recommendation is still a challenging problem. Intuitively, multimodal information have different influence on a user at time step t according to his (or her) historical behavior records before time t. To explain this, we present an illustrative example in Fig. 1, where a sequence of items purchased by a user. At the second to last time step, this user may pay more attention to the item's textual-modality information (Nike) since he purchased a short sleeve of

* Corresponding author.
E-mail address: szniu@bupt.edu.cn (S. Niu).

https://doi.org/10.1016/j.patrec.2021.08.023
0167-8655/

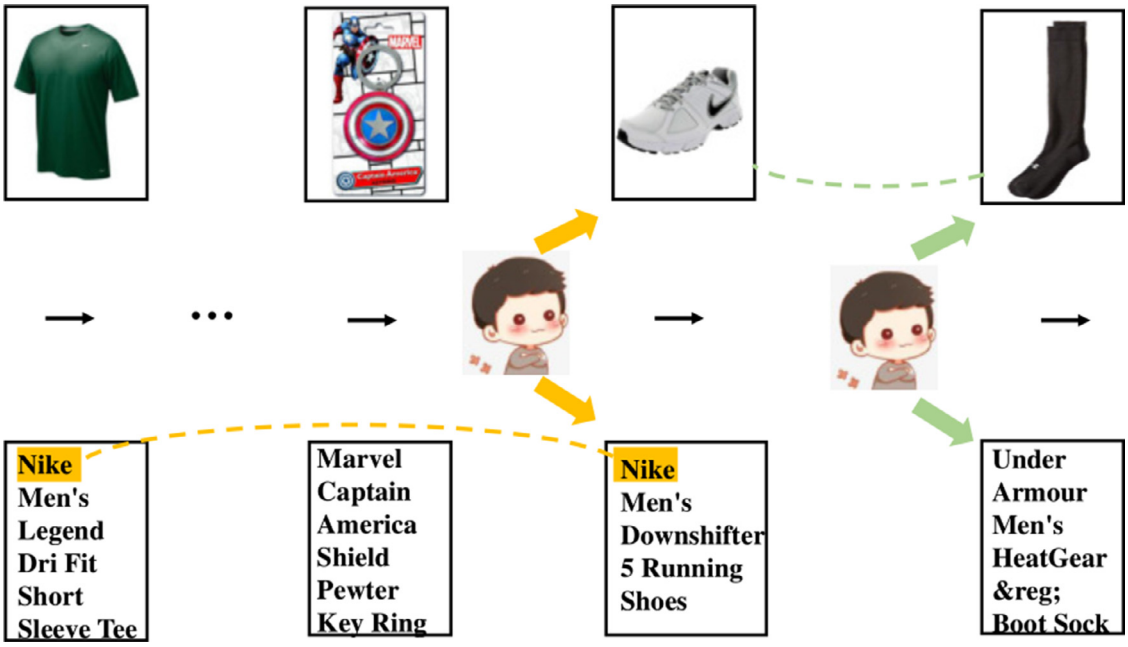

Fig. 1. A sequence of items purchased by a user. Each item is accompanied by an image and its title information.

Nike. At the last time step, the user may care more about the color of socks since they match well with a pair of white sports shoes.

Although incorporating multimodal information with sequential recommendation has great importance, it's non-trivial to model the sequential multimodal interactions and capture dynamic multimodal preference for each user. To address the above problem, we design a multimodal-adaptive hierarchical network (MAHN for short), which can effectively model the sequential dynamic multimodal interactions including visual- and textual-modality. Specifically, the hierarchical recurrent neural network is designed to realize the re-selection of multimodal information from the first layer to the second layer. The information modulation module is designed to select each modal information at time step t according to historical multimodal information. Finally, to improve the generalization ability of our model, we adopt the multi-task training style to jointly optimize BPR loss and reconstruction loss of multimodal information.

In this paper, we propose a new framework incorporating mulitmodal information to sequential recommendation. Our main contributions are listed as follows:

- — We design a multimodal-adaptive hierarchical network to model the sequential patterns of multi-modalities and consider different influences of various modalities at each time step in a sequence. The hierarchical RNN-based network can capture features from the bottom layer to the top layer, which models the dynamic impact on users from multimodal information. Extensive experiments are conducted on two public datasets, which show that our method outperforms the other methods on Top-K sequential recommendation task.
- — We adopt multi-task learning style to train our network, including BPR loss and multimodal information reconstruction loss, which can narrow the gap between modalities and enhance the generalization ability of the model.

2. Related works

In this section, we review some related fields including sequential recommendation, multimedia recommendation.

2.1. Sequential recommendation

A large and growing body of literature has investigated about how to capture the user's dynamic preference over time. Much of the current literature on sequential recommendation pays attention to deep neural networks which have powerful modeling capabilities. According to the basic framework of these models, they can be broadly classified into three main approaches, RNN-based methods, CNN-based methods and attention-based methods. Among the RNN-based methods, one key milestone is model GRU4Rec ([12]), which is the first one that applies RNN to sequential recommendation. Based on it, several improved methods are proposed to model user's behavior sequence by RNN networks ([2,4,11,23]). For example, [13] propose a parallel RNN architecture to model behaviors based on click actions and additional features of the clicked items. [14] adopt a GRU component for capturing sequential dependency and further incorporate KG for enhancing the modeling of attribute-level user preference. Among the CNN-based models, [24] apply convolution on the 2-dimensional latent matrix which is the embedding matrix of L consequent items and capture both point-level and union-level features by a horizontal convolutional layer and a vertical convolutional layer respectively. Based on [21], [10] integrate future data into model training to fill the gap between historical and future data. Among the attention-based models, [17] propose a sequential recommendation model based on self-attention mechanism to select the more relevant item according to historical behaviors. [22] apply the bidirectional self-attention network to sequential recommendation to capture both the two directions sequential interactive features.

2.2. Multimedia recommendation

There are a number of studies which aim to improve recommendation performance by leveraging the different multimodal information of items. According to the specific modality involved by these methods, they can be divided into three main approaches, visual-modality based methods, textual-modality based methods and hybrid-modality based methods. Among the models integrated with visual information, [19] make suit recommendations accord-

ing to style information extracted from images. [8] improve the expressive ability of the network by modeling both the item's visual representation and the user's visual preference representation. [16] learn the fashion-aware image representations and can generate new item images satisfying user's personal taste. [3] apply an attention layer on the fine-grained visual regions to give out visual explanations. Among the models integrated with textual information, [1] propose a novel matrix factorization model which can consider the ratings and corresponding review texts together. [9] make an explainable recommendation by exploring the aspects extracted from reviews. [5] introduce an aspect-aware topic model to analysis review text, and then evaluate user preferences on different aspects of items by importance score. Among the hybrid-modality methods, [27] propose a joint representation learning framework which can leverage different modalities to learn user and item representations. [6] encode multimodal information with autoencoders to enrich the representations of items, which aims to solve the cold start problem.

3. Preliminary

In this section, we introduce the symbols used in this paper and give a formalization of the sequential recommendation. We also briefly introduce some preliminary knowledge about recurrent neural Network (RNN).

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

Task Definition. Based on these notations, our task of sequential recommendation aims 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 $t^u_{1:n}$:

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

RNN. RNN-based models such as GRU4Rec ([12]) are good at capturing sequential user behaviors. LSTM and GRU are variant forms of Recurrent Neural Network. In this paper, we briefly introduce the network structure of LSTM, which will be used in our network. It includes three gate mechanisms, forget gate, input gate and output gate. The basic operations are listed as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \hat{c}_t &= tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \\ c_t &= f_t \circ c_{t-1} + i_t \circ \hat{c}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t \circ tanh(c_t) \end{aligned} \tag{2}$$

where h_t, c_t represent hidden state and cell state at time t, x_t represents the input at time t. $W_l, b_l, l \in \{f, i, c, o\}$ are parameters. For simplicity, we use function $lstm(x)$ to represent the LSTM network in the following paper.

4. Our approach

In this section, we introduce the proposed multimodal-adaptive hierarchical network (MAHN) in detail, the structure of MAHN is shown in Fig. 2. For simplicity, We drop the superscript of u in the notations for ease of reading. MAHN is a hierarchical network structure, and each layer has its special role in the overall framework.

The First Layer. To get the knowledge included in multi-modalities, we first do some work to handle the images and texts. For each item i, the pre-trained network VGG16 is used to extract the image feature vector following the work [3], denoted by $e^v_i \in \mathbb{R}^{4096}$, the pre-trained Bert-Encoder is used to extract the text feature vector, denoted by $e^t_i \in \mathbb{R}^{768}$. Since the difference of e^v_i and e^t_i, we set two parameter matrices W and E to map the vectors from high-dimensional space to low-dimensional space. And e_i represents the latent factors of item i. The operations are listed as follows:

$$\begin{aligned} \hat{e}^v_i &= We^v_i \\ \hat{e}^t_i &= Ee^t_i \end{aligned} \tag{3}$$

After we get the knowledge of visual modality and textual modality, which denoted by $\hat{e}^v_i$ and $\hat{e}^t_i$ in Eq. (3), we take them and item latent representation as all inputs of item i. Thus, we can derive the input of item i in the first layer by concatenating them:

$$h_t = lstm(\hat{e}^v_i \oplus e_i \oplus \hat{e}^t_i) \tag{4}$$

where $\oplus$ represents concatenation. Here we must point out that other operations on these three vectors are allowed. We will explore different influences on recommendation performance of some other operations in the experiment part.

The Second Layer. In this layer, we design an information modulation module to select the useful information for each modality based on historical multimodal information, which is captured by the first layer. Based on the first layer, h_t can be considered as a summary of multimodal historical information at time step t. Then h_t is used to select the valuable information carrying by visual vector and textual vector. For simplicity, we use the visual modality as an example. The modulation operation are listed as follows:

$$\begin{aligned} \alpha &= \sigma((M_v \cdot h_t + b_v) \odot (N_v \cdot \hat{e}^v_i + q_v)) \\ \bar{e}^v_i &= \alpha \odot \hat{e}^v_i \end{aligned} \tag{5}$$

where M_v, N_v, b_v, q_v are training parameters. σ is the sigmoid function. We can get $\bar{e}^t_i$ by the same operation on $\hat{e}^t_i$. Then the input of the second layer is:

$$\tilde{h}_t = lstm(\bar{e}^v_i \oplus h_t \oplus \bar{e}^t_i) \tag{6}$$

where $\oplus$ represents concatenation. In the experiment part, we will explore the operations used in Eq. (6) too.

Learning and Prediction. To improve the generalization ability of the model, we adopt multi-task training style to learn all parameters. The multimodal information reconstructed loss is used to regular the representations of each modality. The specific operations are as follows:

$$loss_i = \|W^T(\hat{e}^v_i + \hat{e}^t_i) - e^v_i\|^2 + \|E^T(\hat{e}^v_i + \hat{e}^t_i) - e^t_i\|^2 \tag{7}$$

In this paper, we use BPR rank loss to optimize the recommendation loss. We use the triple set $S = \{(u, i, j) : u \in \mathcal{U}, i, j \in \mathcal{I}\}$, where i is target item, and j is negative item. the user u's preference score on item i is computed as follows:

$$s_{u,i} = h_t(\hat{e}^v_i \oplus e_i \oplus \hat{e}^t_i) \tag{8}$$

where $\hat{e}^v_i \oplus e_i \oplus \hat{e}^t_i$ is the representation of item i. In the evaluation phase, $\hat{e}^v_i$ and $\hat{e}^t_i$ can be derived since parameters have been trained. Finally, we optimize the overall loss function by Adam Optimizer.

$$loss = \sum_{(u,i,j)\in S} -\log(\sigma(s_{u,i} - s_{u,j})) + loss_i + loss_j \tag{9}$$

5. Experiments

In this section, we evaluate the proposed model MAHN by conducting experiments on two public datasets. we describe the datasets and baselines used in experiments.

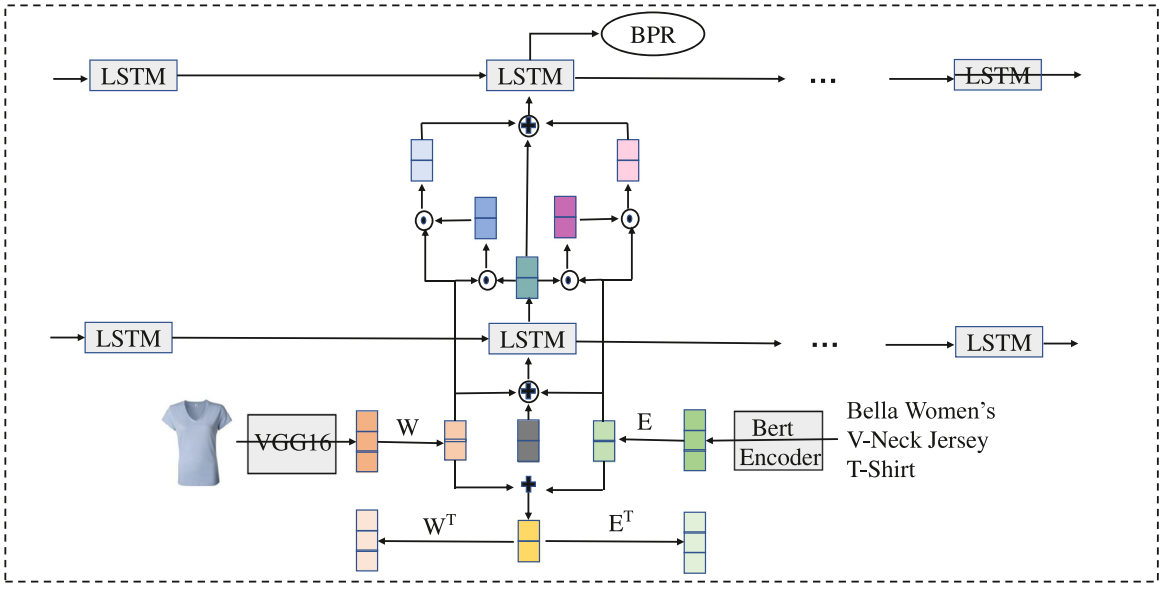

Fig. 2. The overview of model MAHN. It contains two layers, which achieve the re-combination of multimodal information from the bottom layer to the top layer. The information modulation module between the two layers realizes the selection of each modal information at time step t based on the previous time steps.

Table 1
Statistics of datasets for experiments.

Datasets	# *Cell Phones*	# *Clothing*
# Users	27,804	39,386
# Items	10,192	23,010
# Interactions	191,396	278,406

Datasets and Evaluation Metrics. We conduct our experiments on two public datasets, *Cell Phones & Accessories* and *Clothing, Shoes & Jewelry*, which are two domains from Amazon Datasets [19]. For simplicity, we use *Cell Phones*, *Clothing* to represent them respectively. We filter the item without title and image. We only keep the users and items whose times of occurrences are not lower than five records. We choose the last one interaction of each record as the test data, and the second last interaction as the validation data, the remaining interactions as the training data. The statistics of two datasets are shown in Table 1. For evaluation, we choose Top-N recommendation list for each user, where N=10, 20. For each user, we randomly sample 100 negatives and rank them with the target item following the strategy in [17] to avoid heavy computation.

Baselines. We compare with the following recommendation models to justify the effectiveness of our approaches.

- **VBPR** [8]: A visual Bayesian Personalized Ranking model.
- **JRL** [27]: Joint Representation Learning framework that incorporates heterogeneous information sources for recommendation.
- **MV-RNN** [6] : A RNN-based model which utilizes visual and textual information to enhance the representations of items.
- **GRU4Rec** [12]: An RNN-based model, which uses GRU units and utilizes session-parallel minibatches to make session-based recommendation.
- **Caser** [24]: Caser captures the stream-level patterns by utilizing CNN on the adjacent items
- **SASRec** [17]: A self-attention based sequential model that captures long-term semantics for recommendation.
- **BERT4Rec** [22]: A bidirectional self-attention network which learns users sequential patterns to make recommendations.

Parameter Settings. As to the baselines, we utilize the recommended setting by their original work. For SASRec, Caser, we use the codes released by their authors. The rest models are implemented in PyTorch. For MAHN, the learning rate is set to 10^{-3}, and the regularization coefficient is 10^{-4}. During training, we use dynamic scheduler learning rate by *lr_scheduler_StepLR*, where *step_size* is set to 10, and *gamma* is set to 0.5. The embedding size of all modalities is set to 50. Batch size is 5. The number of layer for lstm is set to 1. We conduct all experiments on NVIDIA 2080Ti.

5.1. Comparison against baselines

We compare our MAHN model against several competitive baseline methods on next-one recommendation task. We present the comparison results in Table 2 and 3. From the experimental results, we have the following observations.

- Firstly, all evaluation metrics have achieved the best results on the two datasets, which can prove that the proposed model MAHN can effectively model the interactions of multimodal information in sequential data. The overall network structure can achieve the re-selection of valuable information by modulation module, which relies on the output of the first layer.
- Secondly, the improved performance is more when $N = 20$ than when $N = 10$. The improvements are 3.16% and 2.28% at Hit-Ration@10 and NDCG@10 respectively, 5.59% and 5.18% at Hit-Ration@20 and NDCG@20 respectively. This phenomenon proves that our model has higher accuracy when recommending more products to each user.
- Finally, the improvements on dataset *Clothing* is more than that on *Cell Phones*. This result is consistent with our intuition that the visual and textual characteristics of items have a greater impact on clothing products than electronic products.

5.2. Variants of MAHN model

In this section, we design some variants of MAHN model to explore that how different treatments on multimodal information influence the final recommendation accuracy. We use *add* and *cat* to represent addition operation and concatenation operation between multimodal information. We design four variant MAHN of different combinations of *add* and *cat*, denoted by *add-add, add-cat, cat-add* and *cat-cat*.

Table 2
Performance comparison for baselines and our model MAHN, where N is set to 10 for Top-N Recommendation. Bolded numbers are the best performance of each column.

Recommendation Models	*Cell Phones & Accessories*		*Clothing Shoes & Jewelry*	
	Hit-Ratio@10	NDCG@10	Hit-Ratio@10	NDCG@10
VBPR	0.2778	0.1562	0.1573	0.0869
JRL	0.3427	0.1991	0.2230	0.1189
MV-RNN	0.5349	0.3323	0.3422	0.1996
GRU4Rec	0.4414	0.2703	0.2755	0.1557
Caser	0.4973	0.3172	0.2827	0.1619
SASRec	0.5659	0.3607	0.3818	0.2227
BERT4Rec	0.5783	0.3674	0.3879	0.2269
MAHN	**0.5966**	**0.3758**	**0.4262**	**0.2467**
Improv.	3.16%	2.28%	9.87%	8.72%

Table 3
Performance comparison for baselines and our model MAHN, where N is set to 20 for Top-N Recommendation. Bolded numbers are the best performance of each column.

Recommendation Models	*Cell Phones & Accessories*		*Clothing Shoes & Jewelry*	
	Hit-Ratio@20	NDCG@20	Hit-Ratio@20	NDCG@20
VBPR	0.3689	0.1726	0.2643	0.1157
JRL	0.4525	0.2286	0.3549	0.1531
MV-RNN	0.6701	0.3483	0.4562	0.2159
GRU4Rec	0.5596	0.2984	0.3940	0.1856
Caser	0.6009	0.3372	0.4064	0.1881
SASRec	0.6929	0.3835	0.4968	0.2429
BERT4Rec	0.7057	0.3918	0.5028	0.2491
MAHN	**0.7452**	**0.4121**	**0.5570**	**0.2721**
Improv.	5.59%	5.18%	10.7%	9.23%

Table 4
Performance comparison of Variants about MAHN, where N is set to 10 for Top-N Recommendation. *add-add* means addition operation in the first and the second layer, *add-cat* means addition operation in the first layer and concatenation operation in the second layer, *cat-add* means concatenation operation in the first layer and addition operation in the second layer and *cat-cat* means concatenation operation in the first and the second layer.

Recommendation Models	*Cell Phones & Accessories*		*Clothing Shoes & Jewelry*	
	Hit-Ratio@10	NDCG@10	Hit-Ratio@10	NDCG@10
add-add	0.5886	0.3675	0.4185	0.2348
add-cat	0.5961	0.3746	0.4219	0.2423
cat-add	0.5873	0.3667	0.4154	0.2377
cat-cat	0.5966	0.3758	0.4262	0.2467

Table 5
Performance comparison of Variants about MAHN, where N is set to 20 for Top-N Recommendation. *add-add* means addition operation in the first and the second layer, *add-cat* means addition operation in the first layer and concatenation operation in the second layer, *cat-add* means concatenation operation in the first layer and addition operation in the second layer and *cat-cat* means concatenation operation in the first and the second layer.

Recommendation Models	*Cell Phones & Accessories*		*Clothing Shoes & Jewelry*	
	Hit-Ratio@20	NDCG@20	Hit-Ratio@20	NDCG@20
add-add	0.7399	0.4056	0.5421	0.2659
add-cat	0.7447	0.4129	0.5562	0.2715
cat-add	0.7393	0.4050	0.5446	0.2632
cat-cat	0.7452	0.4121	0.5570	0.2721

We conduct experiments on the two datasets using the same hyperparameters setting with model MAHN. We also evaluate the performance for Top-N recommendation, where N is 10, 20. Results are shown in Table 4 and Table 5. From the experimental results, we can drop the following conclusions.

- The best performance is achieved when concatenation operation is adopted by both the two layers. Furthermore, the performance difference between variants *add-cat* and *cat-cat* is very small.
- The performance difference between variants *add-add* and *cat-add* is very small. It means that when the addition operation is applied in the second layer, what operation is adopted by the first layer has little impact.
- We can find that when concatenation operation is adopted by the second layer, the performance is higher about one percent than when addition operation is adopted by the second layer no

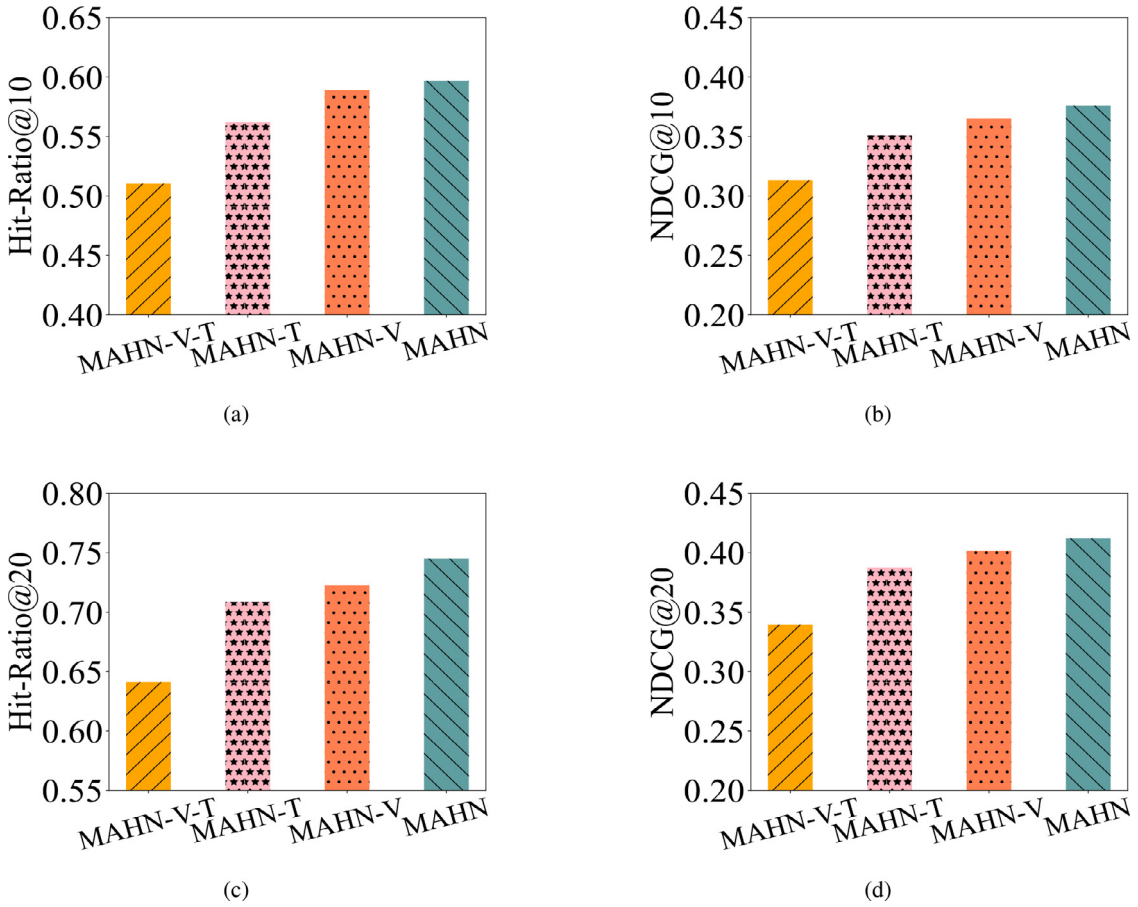

Fig. 3. Three ablation studies on four evaluation metrics of our approach on dataset *Cell Phones*.

matter what operation is adopted by the first layer. Therefore, we can drop a conclusion that the operation adopted by the second layer has more influence on the final performance and the concatenation operation can achieves better performance.

5.3. Ablation studies on MAHN model

To evaluate each module of MAHN, we design some ablation studies on our model. To assess the impact of visual information, we remove the visual-related modules, namely MAHN-V. To evaluate the influence of textual information, we remove the textual-related modules, denoted by MAHN-T. Finally, we remove visual- and textual-related modules together, namely, MAHN-V-T. Fig. 3 shows the performance comparison among MAHN and its three degraded models. From the results of three ablation studies, we can get the following observations.

- MAHN-V-T is worst among all the three ablation studies, which proves that integrating multimodal information of item contributes to the accuracy definitely. MAHN-V-T only contains the network structure of RNN, so the performance is similar to GRU4Rec, which proves that multimodal information have great contribution to sequence recommendation. Comparing with the baseline MV-RNN, which is a model including sequential recommendation modeling and multi-modal modeling, our model MAHN improves about 12% and 10% for Hit-Ration@10 and NDCG@10 respectively on *Cell Phones & Accessories* dataset. It proves that multimodal information modulation module can capture multi-modal sequence interactions efficiently.
- We find that MAHN-T and MAHN-V performs better than MAHN-V-T, but worse than MAHN on two evaluation metrics. This phenomenon indicates that each modality may help the recommendation in their own way. Textual modality can provide some information such as categories and characteristics of commodities, and visual modality provide the images of goods. MAHN-V is better than MAHN-T by 2% and 1% for Hit-Ration@10 and NDCG@10, which means that textual information contributes a little more than visual information.

5.4. Parameter setting analysis

To analyze the performance of the model, we study the effect of hyper-parameters to MAHN in this section. We study the effect of different embedding sizes to MAHN. Specifically, we tune the embedding size from 10 to 50, and plot the results on *Cell Phones & Accessories* in Fig. 4. Observations on the other dataset are similar.

From the results we find that as the embedding size increases, the test performance in terms of Hit-Ratio@10 and NDCG@10 increases too. The trending is quite consistent over the other dataset. We find that if we keep increasing the embedding size, there will be less performance improvement but larger computational complexity and may lead to the over-fitting problem. By observing the experimental results, we found that when the dimension is larger than 30, the improvement of performance becomes smaller. It is worth noting that in this experiment, we don't compare the baseline JRL. Because JRL achieves the best performance when the embedding size is 300 since of its assembled structure. By comparing the experimental results, our model MAHN can achieve better results than these baselines, which proves that our model can capture sequential and multi-modal features well.

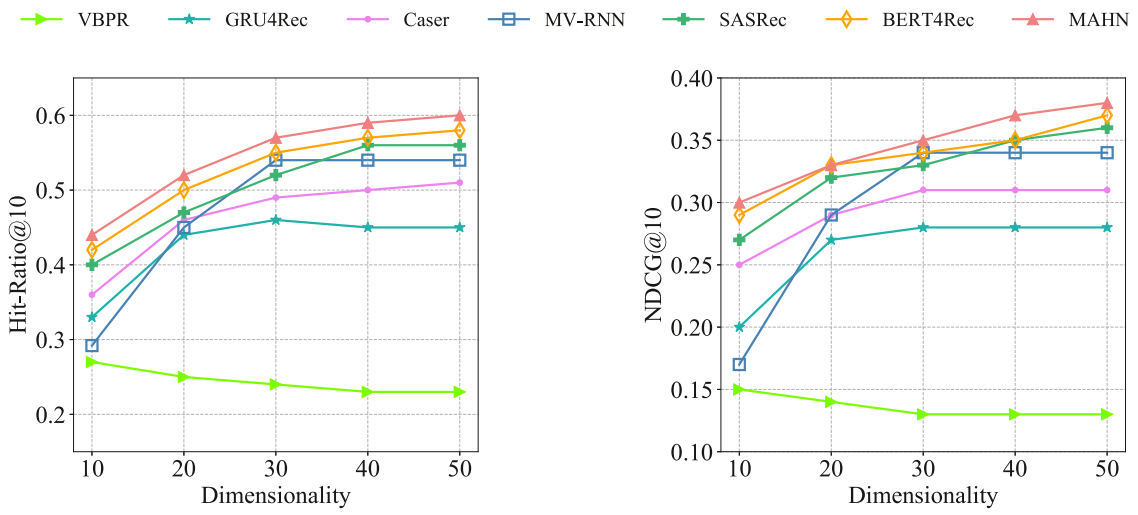

Fig. 4. Performance variation in terms of Hit-Ratio@10 and NDCG@10 against embedding size on Cell Phones and Accessories dataset. The number of embedding seize is increased from 10 to 50.

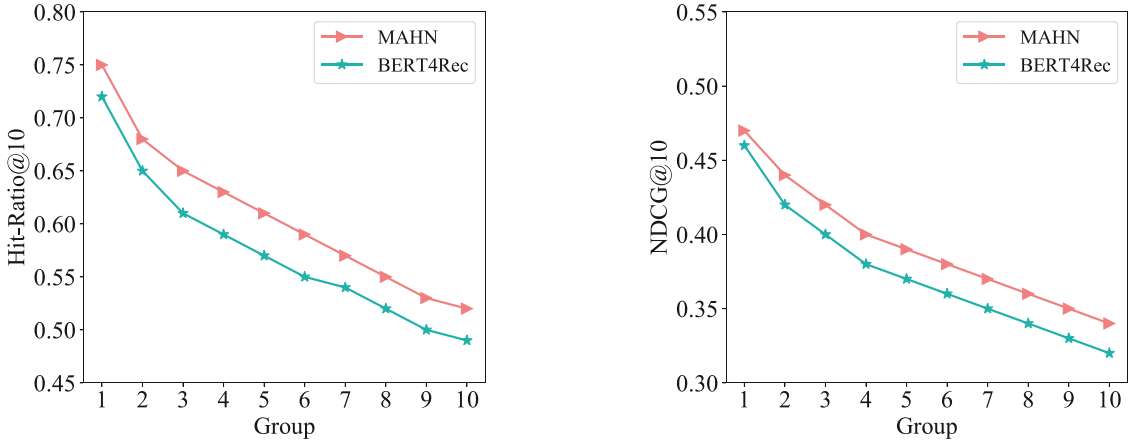

Fig. 5. Performance comparison among different item groups on Cell Phones and Accessories dataset. The x-axis represents different item groups, y-axis represents the performance in terms of Hit-Ratio@10 and NDCG@10 respectively.

5.5. Sparse analysis

To obtain a better understanding whether MAHN can alleviate the sparse problem, in this section, we conduct the case study to compare MAHN the second-best model BERT4Rec in Fig. 5 qualitatively.

Take *Cell Phones & Accessories* dataset as an example, we first sort all the items according their frequencies in our dataset, then we split the sorted items into 10 groups. In this way, the first group contains the most frequent items, while the $10-th$ group contains the sparsest items. Given this, we compare the performance of these two models on all groups. The performance of MAHN and BERT4Rec decrease when modelling on sparse item groups, and this is consistent with the expectation that sparse items will degrade the performance. Comparing with BERT4Rec, MAHN shows better performance on all item groups in all evaluation metrics. An interesting observation is that performance gain between MAHN and BERT4Rec is increasing when applying these models on sparser item groups.

6. Conclusions

In this paper, we propose a Multimodal-Adaptive Hierarchical Network for multimedia recommendation, which model the influence of historical multimodal information on the current moment by a modulation module between the two *lstm* layers. The first layer of our framework is to capture the historical multimodal information, and the second layer is to re-select the valuable multimodal information based on the output of the first layer by an information modulation module. In the future, we will explore three modalities sequential interactions such as video data. There is still a lot of do to solve the challenges in multimedia recommendation field.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by The National Natural Science Foundation of China (No.U1536121, 61370195).

References

[1] Y. Bao, H. Fang, J. Zhang, Topicmf: Simultaneously exploiting ratings and reviews for recommendation, in: AAAI, AAAI Press, 2014, pp. 2–8.
[2] V. Bogina, T. Kuflik, Incorporating dwell time in session-based recommendations with recurrent neural networks, in: RecTemp@RecSys, in: CEUR Workshop Proceedings, volume 1922, CEUR-WS.org, 2017, pp. 57–59.

[3] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: SIGIR, ACM, 2019, pp. 765–774.
[4] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, H. Zha, Sequential recommendation with user memory networks, in: WSDM, ACM, 2018, pp. 108–116.
[5] Z. Cheng, Y. Ding, L. Zhu, M.S. Kankanhalli, Aspect-aware latent factor model: Rating prediction with ratings and reviews, in: WWW, ACM, 2018, pp. 639–648.
[6] Q. Cui, S. Wu, Q. Liu, W. Zhong, L. Wang, MV-RNN: A multi-view recurrent neural network for sequential recommendation, IEEE Trans. Knowl. Data Eng. 32 (2) (2020) 317–331.
[7] D. Geman, S. Geman, N. Hallonquist, L. Younes, Visual turing test for computer vision systems, Proc. Natl. Acad. Sci. USA 112 (12) (2015) 3618–3623.
[8] R. He, J.J. McAuley, VBPR: visual bayesian personalized ranking from implicit feedback, in: AAAI, AAAI Press, 2016, pp. 144–150.
[9] X. He, T. Chen, M. Kan, X. Chen, Trirank: Review-aware explainable recommendation by modeling aspects, in: CIKM, ACM, 2015, pp. 1661–1670.
[10] X. He, T. Chua, Neural factorization machines for sparse predictive analytics, in: SIGIR, ACM, 2017, pp. 355–364.
[11] B. Hidasi, A. Karatzoglou, Recurrent neural networks with top-k gains for session-based recommendations, in: CIKM, ACM, 2018, pp. 843–852.
[12] B. Hidasi, A. Karatzoglou, L. Baltrunas, D. Tikk, Session-based recommendations with recurrent neural networks, ICLR (Poster), 2016.
[13] B. Hidasi, M. Quadrana, A. Karatzoglou, D. Tikk, Parallel recurrent neural network architectures for feature-rich session-based recommendations, in: RecSys, ACM, 2016, pp. 241–248.
[14] J. Huang, W.X. Zhao, H. Dou, J. Wen, E.Y. Chang, Improving sequential recommendation with knowledge-enhanced memory networks, in: SIGIR, ACM, 2018, pp. 505–514.
[15] J. Johnson, A. Karpathy, L. Fei-Fei, Densecap: Fully convolutional localization networks for dense captioning, in: CVPR, IEEE Computer Society, 2016, pp. 4565–4574.
[16] W. Kang, C. Fang, Z. Wang, J.J. McAuley, Visually-aware fashion recommendation and design with generative image models, in: ICDM, IEEE Computer Society, 2017, pp. 207–216.
[17] W. Kang, J.J. McAuley, Self-attentive sequential recommendation, in: ICDM, IEEE Computer Society, 2018, pp. 197–206.
[18] A. Karpathy, F. Li, Deep visual-semantic alignments for generating image descriptions, in: CVPR, IEEE Computer Society, 2015, pp. 3128–3137.
[19] J.J. McAuley, C. Targett, Q. Shi, A. van den Hengel, Image-based recommendations on styles and substitutes, in: SIGIR, ACM, 2015, pp. 43–52.
[20] S.E. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, H. Lee, Generative adversarial text to image synthesis, in: ICML, in: JMLR Workshop and Conference Proceedings, volume 48, JMLR.org, 2016, pp. 1060–1069.
[21] S. Rendle, C. Freudenthaler, L. Schmidt-Thieme, Factorizing personalized markov chains for next-basket recommendation, in: WWW, ACM, 2010, pp. 811–820.
[22] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: CIKM, ACM, 2019, pp. 1441–1450.
[23] Y.K. Tan, X. Xu, Y. Liu, Improved recurrent neural networks for session-based recommendations, in: DLRS@RecSys, ACM, 2016, pp. 17–22.
[24] J. Tang, K. Wang, Personalized top-n sequential recommendation via convolutional sequence embedding, in: WSDM, ACM, 2018, pp. 565–573.
[25] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: Fine–grained text to image generation with attentional generative adversarial networks, in: CVPR, IEEE Computer Society, 2018, pp. 1316–1324.
[26] L. Yu, E. Park, A.C. Berg, T.L. Berg, Visual madlibs: Fill in the blank description generation and question answering, in: ICCV, IEEE Computer Society, 2015, pp. 2461–2469.
[27] Y. Zhang, Q. Ai, X. Chen, W.B. Croft, Joint representation learning for top-n recommendation with heterogeneous information sources, in: CIKM, ACM, 2017, pp. 1449–1458.