applied sciences

MDPI

*Article*

# Auxiliary Information-Enhanced Recommendations

**Shoujin Wang, Wanggen Wan *, Tong Qu * and Yanqiu Dong**

School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China; jimmwangshu@163.com (S.W.); yanqiu_dong@shu.edu.cn (Y.D.)
* Correspondence: wanwg@staff.shu.edu.cn (W.W.); qu_tong@shu.edu.cn (T.Q.)

**Abstract:** Sequential recommendations have attracted increasing attention from both academia and industry in recent years. They predict a given user's next choice of items by mainly modeling the sequential relations over a sequence of the user's interactions with the items. However, most of the existing sequential recommendation algorithms mainly focus on the sequential dependencies between item IDs within sequences, while ignoring the rich and complex relations embedded in the auxiliary information, such as items' image information and textual information. Such complex relations can help us better understand users' preferences towards items, and thus benefit from the recommendations. To bridge this gap, we propose an auxiliary information-enhanced sequential recommendation algorithm called memory fusion network for recommendation (MFN4Rec) to incorporate both items' image and textual information for sequential recommendations. Accordingly, item IDs, item image information and item textual information are regarded as three modalities. By comprehensively modelling the sequential relations within modalities and interaction relations across modalities, MFN4Rec can learn a more informative representation of users' preferences for more accurate recommendations. Extensive experiments on two real-world datasets demonstrate the superiority of MFN4Rec over state-of-the-art sequential recommendation algorithms.

**Keywords:** recommendations; sequential recommendations; recommender systems; auxiliary information



**Citation:** Wang, S.; Wan, W.; Qu, T.; Dong, Y. Auxiliary Information-Enhanced Recommendations. *Appl. Sci.* **2021**, *11*, 8830. https://doi.org/10.3390/app11198830

Academic Editor: Vincent A. Cicirello

Received: 17 July 2021
Accepted: 20 September 2021
Published: 23 September 2021



## 1. Introduction

Recommender systems have had an ever-increasingly important role in our daily life to help users effectively and efficiently find the items of their interest from a large amount of choices. Sequential recommender systems, as a relatively new type of recommender system, have attracted much more attention in recent years. A sequential recommender system (SRS) aims at providing sequential recommendations, namely recommending the next item to an user by learning the user's preference from their recent historical interactions (e.g., clicks, purchases) with items. By effectively modeling the user's recent interactions, an SRS is able to capture a user's latest preference and thus generate accurate sequential recommendations.

Although effective, existing SRSs still have some drawbacks. One typical case is the ignorance of auxiliary information. Specifically, in real-world e-commence cases, in addition to the explicit or implicit user–item interactions which are mainly indicated by item IDs, there are other types of information which can also reveal users' preferences, such as item attributes, appearance images, and description texts. In practice, item ID information and the corresponding various types of auxiliary information can be treated as multi-modal information where each type of information serves as one modality. Some conventional recommendation algorithms including collaborative filtering and content-based filtering have utilized this auxiliary information to better characterize items and to complement the user–item interaction information. As a result, more specific user preferences towards items can be extracted for improving recommendation performance.

In the sequential recommendation scenarios, there are more complex relations embedded in the aforementioned multi-modal information. To be specific, there are not only *sequential relations within modalities*, e.g., a user's implicit interactions (clicks) with items usually being sequentially dependent, but also *interaction relations between different modalities*, e.g., the correlations between the item description texts and item appearance images. However, most of the existing SRS algorithms either ignore such auxiliary multi-model information or simply model a single type of relation embedded within such auxiliary information. For example, the visual content-enhanced sequential recommender system (VCSRS) first learns an attentive item visual content representation and then incorporates it into an LSTM-based recurrent neural network (RNN) for next-item recommendations [1]. However, VCSRS not only ignores the richer textual description information of items (e.g., reviews), but also fails to model the sequential dependencies within each modality as well as the interaction relations across different modalities. The parallel recurrent neural network (p-RNN) first utilizes multi recurrent neural networks to model the sequential dependencies over items embedded in user–item interactions, i.e., clicks, item description texts and item images, separately and then integrate the modeled sequential dependencies from different modalities together for the downstream recommendations [2]. However, p-RNN only models the sequential dependencies within each modality while ignoring the complex interaction relations across different modalities. Multi-view RNN (MV-RNN) also employs both text and image information for sequential recommendations [3]. In MV-RNN, an auto-encoder-based multi-modal representation fusion module is designed to generate a compound representation for each item by integrating the item-related information from multiple modalities. The compound representation of a given item is then input into a gated recurrent unit (GRU) of the corresponding time step of an RNN to model the sequential dependencies among items. Finally, the final hidden state of the RNN is regarded as the user's preference for generating recommendations. Although effective, such a method mainly considers the interaction relations across different modalities, while the sequential dependencies within each modality are weakened.

To bridge the aforementioned drawbacks of existing works, in this paper, we aim at developing an accurate sequential recommendation algorithm by effectively extracting and aggregating useful information from multi-modal auxiliary information, as well as modeling the complex interaction relations embedded in them. To be specific, we devise a memory fusion network for recommendation (MFN4Rec) by effectively integrating the relevant information from three modalities, i.e., item IDs, item images and item description texts, and modelling the complex relations between and within modalities. MFN4Rec is built on a typical work in multi-modal sequence representation learning, i.e., a memory fusion network for multi-view sequential learning [4], for multi-modal representation learning for sequential recommendations. To be specific, MFN4Rec contains a multi-GRU layer, a multi-view gated memory network (MGMN), and a prediction module. The multi-GRU layer contains three GRU-based RNNs, while each RNN models the sequential dependencies by taking the modal-specific representation of each item as the input of each step. MGMN is designed to model and extract the interaction relations across different modalities. The outputs from both the multi-GRU layer and MGMN are combined together as the input of the prediction module for the next-item prediction. Benefiting from the information memory and spreading mechanism of the memory network, MFN4Rec is able to not only effectively handle the relations within and across modalities, but also effectively model the dynamic sequential dependencies in sequences, and thus make the multi-modal auxiliary information contribute more to the sequential recommendations.

The contributions of this work are summarized below:

- We propose a memory fusion network for recommendation (MFN4Rec) to effectively model auxiliary multi-modal information for accurate sequential recommendations.
- A multi-GRU layer is designed to effectively model the sequential dependencies with each modality.

- A multi-view gated memory network (MGMN) is particularly devised to effectively model the complex interaction relations across different modalities.

Extensive experiments have been conducted on two real-world e-commerce transaction datasets. The results have demonstrated the superiority of our proposed SRS algorithm over the state-of-the-art ones when performing sequential recommendations.

## 2. Related Work

In this section, we first review the existing work on conventional sequential recommendations and then review the existing work on auxiliary information-enhanced sequential recommendations.

### *2.1. Sequential Recommendation Algorithms*

Generally speaking, according to the employed techniques, sequential recommendation algorithms can be roughly divided into traditional sequential recommendation algorithms and deep-learning-based sequential recommendation algorithms.

Traditional sequential recommendation algorithms are built on traditional data mining or machine learning techniques, including sequential pattern mining, Markov chain models, matrix factorization, and neighborhood models. Yap et al. [5] introduced a personalized sequential pattern mining algorithm to first mine personalized sequential patterns and then utilize the mined patterns for guiding the downstream recommendations. Pattern mining-based algorithms are simple and sometimes effective, but they easily lose those infrequent, but important, items and patterns, and thus reduce the recommendation accuracy. Feng et al. [6] proposed a Markov chain-based SRS algorithm called the Personalized Ranking Metric Embedding (PRME) model for the next POI recommendations. Markov chain-based algorithms can only model the first-order dependencies while ignoring the high-order dependencies, and thus reduce the recommendation accuracy. Rendle et al. [7] proposed a classic matrix factorization model called the Factorized Personalized Markov Chains (FPMC) model to factorize the transition matrix over items from adjacent baskets into the latent factors of items. The latent factors are then utilized for next-basket recommendation. However, matrix factorization methods easily suffer from data sparsity issues.

In recent years, deep learning models including RNN and CNN have shown great potential to capture the complex relations in sequence, and thus have been widely employed into sequential recommendations. Due to its powerful capability to model sequence data, RNN is the prominent deep model for sequential recommendations. Hidasi et al. [8] proposed an Gated Recurrent Units (GRU)-equipped RNN-based model called GRU4Rec for the next-item prediction. GRU4Rec was further improved by introducing a novel and tailored ranking loss function [9]. Some other similar works include Long Short Term Memory (LSTM)-based SRS algorithms [10]. Later, hierarchical RNN was employed in sequential recommendations to model both intra-sequence dependencies and inter-sequence dependencies for next-item recommendations [11]. However, the rigid order assumption over any two adjacent interactions employed in RNN may lead to generating false sequential dependencies [12]. In addition to RNN, CNN are also applied into sequential recommendations to build CNN-based SRS algorithms. Tang et al. [13] developed a convolutional sequence embedding recommendation model called Caser. Caser employs horizontal and vertical convolutional filters to learn the item-level and feature-level dependencies, respectively, for sequential recommendations. Further, a 3D CNN model was developed for jointly modeling the sequential relations and item content features for next-item recommendations [14]. However, CNN-based SRSs may not be able to effectively capture the long-range dependencies due to the limited perceptive fieldof CNN. Most recently, graph neural networks (GNN), as an advanced deep architecture, have been applied into sequential recommendations. Typical GNN-based sequential recommendation algorithms include memory augmented graph neural networks (MA-GNN) [15] and RetaGNN [16]. Some other researchers employed an attention mechanism into sequential recommendations for improving the recommendation performance. Wang et al. [12] utilized the attention model

to learn attentive item and session representations for next-item recommendations. Later, a self-attention mechanism was introduced to better capture those heterogeneous relations embedded in a sequence of interactions for accurate sequential recommendations [17–19]. Although these deep models have shown great potential in achieving good recommendation performance, they usually ignore the rich auxiliary information. This limits the further improvement of the recommendation performance.

### 2.2. Auxiliary Information-Enhanced Sequential Recommendations

In the real-world scenarios, in addition to the commonly used item ID information, there is rich auxiliary information related to items, users and interactions. Such auxiliary information can provide more contextual information for an in-depth understanding of users' sequential behaviors, and thus can benefit the subsequent sequential recommendations. For instance, Wang et al. [20] take both the item ID and the corresponding item attributes in a session as the input of a shallow neural networks to learn a compound embedding for each item for the downstream next-item recommendations. A 3D convolutional neural network was proposed by Tuan et al. [14] to learn informative item representations from both item IDs and content features of items for next-item prediction. With the introduction of a neighborhood model, Garg et al. [21] incorporated the readily available position information of items within sequences for more accurate sequential recommendations. The occurrenceof timestamps of users' interactions in sequences was explored by Li et al. [22] and Ye et al. [23] for next-item recommendations. These works have taken a step forward to incorporate more auxiliary information to enhance sequential recommendations, but they ignore the important and representative item image and textual information.

Only quite limited works on sequential recommendations have taken item image and/or textual information into account. An RNN-based sequential recommendation model called VCSRS was proposed by Qu et al. [1]. VCSRS first utilizes an attention-based visual feature representation learning component to learn a task-specific item visual representation and then effectively incorporate it into a single LSTM-based RNN to complement the item ID information for next-item recommendations. Although effective, on one hand, VCSRS ignores another important piece of auxiliary information, i.e., item textual information; on the other hand, VCSRS does not model different types of information as different modalities, and thus it fails to effectively capture the sequential dependencies within each type of information, namely each modality, (e.g., item ID and visual feature), respectively, and the interactions between different types of information, namely different modalities. Therefore, VCSRS are different from this work in terms of the input data, the solution and the model architecture. A parallel RNN-based model called p-RNN was developed by Hidasi et al. [2] to take item IDs, item images and item textual features as the input to learn informative item representations for sequential recommendations. In the parallel RNN model, three RNNs are utilized to model the aforementioned three parts of information, respectively, and the outputs of all RNNs are combined together for the prediction task. Although p-RNN can improve the recommendation performance to some degree, it ignores the interaction relations between different modalities and thus cannot fully model the complex relations embedded in users' interaction sequence data. Another similar work is multi-view RNN (MV-RNN), which also employs both text and image information for sequential recommendations [3]. In MV-RNN, an auto-encoder-based multi-modal representation fusion module is developed to generate a compound representation for each item by integrating both the item image and textual information. The compound representation is then input to a GRU-based RNN for predicting the next item. Although effective, such a method mainly considers the relations across different modalities, while the sequential dependencies within each modality are weakened.

In summary, although some works have tried to integrate multi-modal auxiliary information into sequential recommendations, they either ignore the interaction relations across different modalities or weaken the sequential dependencies within modalities.

This has limited the further improvement of the recommendation performance. A more effective and reliable algorithm which can effectively incorporate different types of auxiliary information for sequential recommendations is in need, which motivates our work in this paper.

## 3. The Proposed SRS Algorithm

As shown in Figure 1, our proposed memory fusion network for recommendation (MFN4Rec) mainly contains three stages. (1) First, it extracts the feature embedding from each modality, i.e., item IDs, item images, and item description texts, and then imports these extracted feature embeddings into the multi-GRU layer including three GRU-based RNNs. The feature embedding from each modality is imported into the corresponding modal-specific GRU-based RNN for modelling the sequential dependencies within the modality. (2) Second, the output from each RNN is then imported into the multi-view gated memory network (MGMN) to learn the interaction relations across modalities. (3) Finally, the output from both the multi-GRU layer and MGMN are taken as the input of the prediction layer for the next-item prediction. Next, we introduce each stage of MFN4Rec algorithm, respectively.

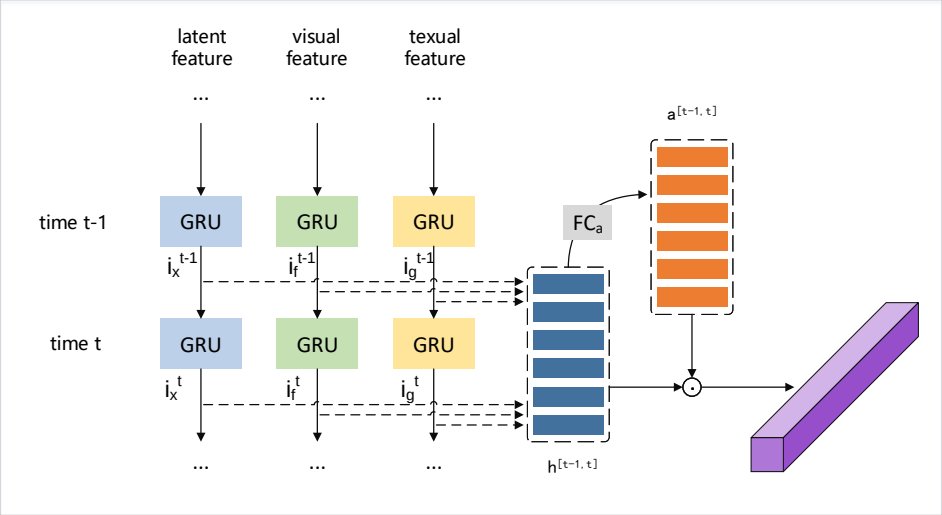


**Figure 1.** The framework of the proposed MFN4Rec algorithm.

### *3.1. Multi-GRU Layer*

Given a sequence of items interacted by a user, $s = v_1, v_2, \ldots, v_{|s|}$, we first extract the multi-modal features for each item in $s$. Particularly, for $v_t \in s$, we extract its ID embedding $\mathbf{m}^t \in \mathbf{R}^{d_m}$, its image feature embedding $\mathbf{f}^t \in \mathbf{R}^{d_f}$ and text feature embedding $\mathbf{g}^t \in \mathbf{R}^{d_g}$. Item ID embedding is obtained from a learnable ID-embedding matrix. Item image embedding is extracted via a 16-layer convolutional neural network (CNN) named VGGNet (shortened to VGG-16) [24] that was pre-trained on ImageNet [25]. The text feature embedding is obtained via the commonly used word-embedding algorithm called GloVe [26]. We first use the pre-trained GloVe to obtain the word-embedding vector of each word in the item description texts, and then we introduce the commonly used TF-IDF algorithm to calculate a weight for each word in the item's text. The final item text embedding with a dimension of 100 is calculated as a weighted sum of the embeddings of words in the text. The dimensions of item ID embedding and image embedding are 25 and 1000, respectively. After the embeddings in three modalities are ready for each item in a sequence, they are imported into the modal-specific GRU-based RNN to model the sequential dependencies within each modality.

Given a sequence $s$, the embedding vectors from each modality of all the items can form a modal-specific embedding sequence. For the $t$th item $v_t$ in $s$, its ID embedding $\mathbf{m}^t$, image feature embedding $\mathbf{f}^t$ and text feature embedding $\mathbf{g}^t$ are taken as the input of the

GRU of the $t$th step in the corresponding GRU-based RNN to output the corresponding hidden state. Accordingly, the multi-GRU layer conducts the following operations in the $t$th step:

$$\mathbf{h}_m^t = GRU_m(\mathbf{m}^t, \mathbf{h}_m^{t-1}) \tag{1}$$

$$\mathbf{h}_f^t = GRU_f(\mathbf{f}^t, \mathbf{h}_f^{t-1}) \tag{2}$$

$$\mathbf{h}_g^t = GRU_g(\mathbf{g}^t, \mathbf{h}_g^{t-1}) \tag{3}$$

where $GRU$ indicates the operations in a normal GRU cell [27]. $GRU_m$, $GRU_f$ and $GRU_g$ are the corresponding modal-specific GRU cells. $\mathbf{h}_m^t$, $\mathbf{h}_f^t$ and $\mathbf{h}_g^t$ are the corresponding hidden state of the current step and they keep the modal-specific sequential information in the sequence $s$. The dimension of the hidden state in each modal is equal to the dimension of the corresponding input in the same modality.

### *3.2. Differentiated Attention Layer*

The differentiated attention layer is designed to extract the cross-modal interaction relations from the three different modal-specific hidden states at each step. Specifically, at the $t$th step, we want to extract the interactions over $\mathbf{h}_m^t$, $\mathbf{h}_f^t$ and $\mathbf{h}_g^t$. Since the different subspace in the hidden state may have different cross-modal interaction strength with the hidden states from other modalities, we need to differentiate the importance of the dimensions of the hidden state when extracting cross-modal interaction relations. For each hidden state, we devised an attention mechanism to emphasize the dimensions which have more interactions with other modalities. The hidden state of multi-GRU at the $t$th step can be represented as $\mathbf{h}^t = [\mathbf{h}_m^t, \mathbf{h}_f^t, \mathbf{h}_g^t]$, where $[,]$ indicates the concatenated operation of vectors.

The differentiated attention layer takes the hidden states from any two adjacent steps as the input to extract the cross-modal relations. At the $t$th step, the input is the concatenation of $\mathbf{h}^{t-1}$ and $\mathbf{h}^t$, denoted as $\mathbf{h}^{[t-1,t]} \in \mathbf{R}^{6*d_h}$. Such input is imported into a full-connected (FC) layer with softmax as the activation function to output the attention weights:

$$\mathbf{a}^{[t-1,t]} = FC_a(\mathbf{h}^{[t-1,t]}) \tag{4}$$

$\mathbf{a}^{[t-1,t]} \in \mathbf{R}^{6*d_h}$ is the element-wise weight vector with the same dimension as $\mathbf{h}^{[t-1,t]}$. Then, $\mathbf{a}^{[t-1,t]}$ performs an element-wise multiplication with $\mathbf{h}^{[t-1,t]}$ to emphasize those important dimensions in the hidden states.

$$\hat{\mathbf{h}}^{[t-1,t]} = \mathbf{a}^{[t-1,t]} \odot \mathbf{h}^{[t-1,t]} \tag{5}$$

The obtained $\hat{\mathbf{h}}^{[t-1,t]}$ can be seen as the latent representation of the cross-modal relations at the current $t$th step.

In Equation (4), by comparing the information embedded in two adjacent hidden states, the attention mechanism can assign the weights accordingly when the hidden state changes from step $t-1$ to step $t$.

The differentiated attention layer takes the hidden state as the input, which usually contains information from the past steps. Therefore, it can capture the interaction relations of different modalities across multiple time steps. This can help our model discover the complex relations embedded in the sequence data.

### *3.3. Gated Multi-Modal Memory Network*

Once the cross-modal interaction relations are extracted at each time step, we utilize a memory network to handle such relations recurrently along with the time steps to achieve the final multi-modal compound memory representation $\mathbf{u}^{\in}\mathbf{R}^{d_m e}$. Specifically, the candidate memory in the current step $t$ is obtained by:

$$\hat{\mathbf{u}}^t = FC_u(\hat{\mathbf{h}}^{[t-1,t]}) \tag{6}$$

where $FC_u$ is a fully connected layer. Then, in the memory network, we use the gate mechanism to control the remains and update the candidate memory. Particularly, two gates, i.e., a remaining gate $\mathbf{g}_1$ and update gate $\mathbf{g}_2$ are introduced as below:

$$\mathbf{g}_1^t = FC_{g_1}(\hat{\mathbf{h}}^{[t-1,t]}) \tag{7}$$

$$\mathbf{g}_2^t = FC_{g_2}(\hat{\mathbf{h}}^{[t-1,t]}) \tag{8}$$

$\mathbf{g}_1$ and $\mathbf{g}_2$ determine how much information in the candidate memory should remain and be updated, respectively. The final memory is obtained as below:

$$\mathbf{u}^t = \mathbf{g}_1^t \odot \mathbf{u}^{t-1} + \mathbf{g}_2^t \odot tanh(\hat{\mathbf{u}}^t) \tag{9}$$

### *3.4. Prediction and Optimization*

The final multi-modal representation of the sequence $s$ of length $T$ is calculated based on the outputs of the multi-GRU layer, i.e., $\mathbf{h}_m^T$, $\mathbf{h}_f^T$, $\mathbf{h}_g^T$, and the output of the gated multi-modal memory network, i.e., $\mathbf{u}^T$. Specifically,

$$\mathbf{h}_{out} = FC_{out}([\mathbf{h}_m^T, \mathbf{h}_f^T, \mathbf{h}_g^T, \mathbf{u}^T]) \tag{10}$$

where $FC_{out}$ is a fully connected layer.

In the prediction layer, a softmax layer is used to map the $\mathbf{h}_{out}$ into the probability distribution over all the candidate items. Then the candidate items will be ranked based on their probability and the top-ranked ones will form the recommendation list. To be specific, the probability is computed as:

$$p^T = \frac{exp(\mathbf{W}\mathbf{h}_{out})}{\sum(exp(\mathbf{W}\mathbf{h}_{out})} \tag{11}$$

The cross-entropy loss is used as the loss function during the training of the model.

$$L = -\sum(\mathbf{p} log(p^T)) \tag{12}$$

where $p^T$ is the predicted probability distribution and $\mathbf{p}$ is the one-hot vector of the ground-truth item to be predicted.

In the model training, Adam optimizer and batch gradient descent are used to optimize the model parameters. Dropout strategy is used to avoid the overfitting of model parameters. We utilize grid-search and cross-validation to adjust the hyper parameters of the algorithm and the used hyper parameters are listed in Table 1.

**Table 1.** The hyper parameters of fully connected layers in the model.

| Fully Connected Layer | Input Dimension | Hidden State Dimension | Output Dimension | Dropout Rate |
|---|---|---|---|---|
| $FC_a$ | $6 * 20$ | no | $6 * 20$ | 0.3 |
| $FC_u$ | $6 * 20$ | 50 | 20 | 0.3 |
| $FC_{g_1}$ | $6 * 20$ | 50 | 20 | 0.3 |
| $FC_{g_2}$ | $6 * 20$ | 50 | 20 | 0.3 |
| $FC_{out}$ | $4 * 20$ | no | 20 | 0.3 |

## 4. Experiments

### *4.1. Data Preparation and Experiment Set Up*

The two subsets of "Clothing, Shoes and Jewelry" and the subset "Phone" in the Amazon dataset https://jmcauley.ucsd.edu/data/amazon/ (accessed on 16 July 2021) are used for our experiments, denoted as the Amazon Clothing dataset and Amazon Phone dataset, respectively, in this work. Intuitively, for these two categories of items, the item images may play a more important role in users' choices of items. Both datasets contain

users' reviews from Amazon.com with timestamps. Following existing works [2,3], we regard each user's review of an item as their interaction with the item. The reviewed items by each user are sorted in chronological order to build the user's sequence of interactions with items. In addition to such interactions, each item has a corresponding image describing the item appearance and review texts from users [28]. We removed sequences whose length is larger than 100 to avoid the lengths of the sequences to be too varying. To avoid the dataset being too sparse, we only keep the interactions which happened in the last two years. The statistics of the dataset are shown in Table 2.

**Table 2.** The statistics of experimental datasets.

| Dataset | # Items | # Users | # Interactions | # Items per User on Average |
|---|---|---|---|---|
| Amazon Clothing | 38,840 | 22,586 | 272,949 | 12.08 |
| Amazon Phone | 27,879 | 10,429 | 175,645 | 16.84 |

For each sequence, we take the last item as the target item to be predicted and use all the other items as the corresponding given context to predict the target item. For each user–item interaction sequence, we first rank them according to the occurrence time. Then, we take the first 60% of interacted items as the training data, the following 20% as the validation data and the last 20% as the test data. Similar to [29,30], we tune the hyperparameters according to the performance on the validation set to obtain the optimal hyperparameters. Then, we use the whole training set to re-train the model. Finally, we test the model on the test set. The dimensions of the hidden state $d_h$ and $d_{mem}$ are 20; the batch size is 64. The initial learning rate is 0.01.

### 4.2. Performance Comparison with Baselines

In the experiments, the representative and state-of-the-art sequential recommendation algorithms are selected as the baselines. We compare our proposed algorithm with these baseline algorithms to evaluate the performance of our algorithm. Specifically, the baseline algorithms include two representative sequential recommendation algorithms, BPR [31] and LSTM [32], and four representative and/or state-of-the-art sequential recommendation algorithms which also incorporate auxiliary information such as item images, namely VBPR [33], p-RNN [2], MV-RNN [3] and VCSRS [1]. Within the same setting of this work, p-RNN, MV-RNN and VCSRS incorporate multi-modal auxiliary information, i.e., item images and text, to improve sequential recommendations. VCSRS is adapted to incorporate both item images and text information by concatenating image representation vectors and text representation vectors to form a unified item auxiliary information representation as the input of the feature-level attention module (FAM). All the baseline algorithms and our proposed algorithm are tested on the aforementioned experimental datasets for recommendation performance comparison. Two representative ranking-based measures, recall and mean average precision (MAP), are used as the evaluation metrics. They are commonly used to evaluate the performance of sequential recommendations. The experimental results are shown in Tables 3 and 4, where the values are percentages and the best ones are marked in bold.

According to Tables 3 and 4, it is clear that our proposed algorithm MFN4Rec achieved the best performance w.r.t all the evaluation metrics, which demonstrate the effectiveness of our proposed algorithm. BPR and LSTM only take the item ID as the input to model the single-modal sequential dependencies among user–item interactions, and they thus perform the worst. Based on BPR, VBPR adds the item image information as auxiliary information. Based on LSTM, p-RNN adds both the item image and text information for sequential recommendation. The performance improvement of both VBPR and p-RNN is limited; this is because they only model the sequential relations within modalities while ignoring the interaction relations across different modalities. Due to the effective learning of item visual content representation and the careful incorporation of it into the LSTM-

based RNN dominated by item ID information, VCSRS can improve the recommendation performance. However, it fails to model the unique sequential dependencies within each modality and the interaction relations between modalities, e.g., item ID and item visual content. The reason is that only one RNN is utilized to model the sequential dependencies over items without a particularly designed component to model the interactions between different modalities. Therefore, the aforementioned sequential dependencies within different modalities are mixed together, and also the interaction relations between different modalities cannot be modelled effectively. Out of all the baseline algorithms, MV-RNN performs the best, demonstrating that the utilization of an auto-encoder can effectively capture the multi-modal information, especially the cross-modal interaction relations. However, it still fails to consider the sequential relations within modalities.

**Table 3.** The performance comparison with baseline algorithms on the Amazon Clothing dataset.

| Algorithm | Recall@20 | MAP@20 | Recall@30 | MAP@30 |
|---|---|---|---|---|
| BPR | 0.641 | 0.168 | 0.812 | 0.176 |
| VBPR | 0.700 | 0.181 | 0.922 | 0.190 |
| LSTM | 1.443 | 0.283 | 1.982 | 0.301 |
| p-RNN | 1.484 | 0.301 | 1.939 | 0.320 |
| MV-RNN(Con) | 2.113 | 0.522 | 2.827 | 0.554 |
| MV-RNN(Fus) | 2.157 | 0.508 | 2.867 | 0.538 |
| MV-RNN(3mDAE) | 2.243 | 0.541 | 2.995 | 0.570 |
| VCSRS | 2.238 | 0.537 | 2.992 | 0.574 |
| MFN4Rec(ours) | **2.384** | **0.572** | **3.111** | **0.609** |

**Table 4.** The performance comparison with baseline algorithms on the Amazon Phone dataset.

| Algorithm | Recall@20 | MAP@20 | Recall@30 | MAP@30 |
|---|---|---|---|---|
| BPR | 4.6398 | 1.5384 | 7.6127 | 1.7624 |
| VBPR | 4.6410 | 1.5438 | 7.6136 | 1.7723 |
| LSTM | 6.7616 | 1.7227 | 8.9707 | 1.8208 |
| p-RNN | 6.8645 | 1.9688 | 8.8463 | 1.8347 |
| MV-RNN(Con) | 6.5580 | 1.7509 | 8.0625 | 1.8739 |
| MV-RNN(Fus) | 6.8783 | 1.8133 | 8.2881 | 1.8764 |
| MV-RNN(3mDAE) | 6.1330 | 1.4085 | 8.0270 | 1.8469 |
| VCSRS | 6.8742 | 1.9904 | 9.0214 | 1.8726 |
| MFN4Rec(ours) | **7.6629** | **2.2697** | **9.7815** | **2.3709** |

In comparison, our proposed MFN4Rec is able to effectively capture the sequential dependencies embedded in multi-modal sequence data by simultaneously capturing both the sequential dependencies within modalities and interaction relations across modalities. As a result, our algorithm can achieve the best performance.

### *4.3. Ablation Analysis*

To verify the effectiveness of each module in our proposed MFN4Rec algorithm, we conduct ablation analysis to measure the contributions of each module to the performance improvement. To be specific, three simplified versions of MFN4Rec are designed to: (1) only keep the multi-GRU layer and combine the final hidden states of all GRU-based RNNs as the input for next-item prediction, denoted as MFN4Rec-g; (2) only keep the gated multi-modal memory network and take its output as the input for prediction, denoted as MFN4Rec-m; (3) remove the gated multi-modal memory network and add the final hidden states of all modalities as the multi-modal memory representation $\mathbf{u}^t$, while others remain the same as MFN4Rec, denoted as MFN4Rec-add. We compare the performance of these three simplified versions with that of MFN4Rec under the same experimental setting. The results are shown in Figure 2.

The experimental results show that MFN4Rec-m performs the worst; it only considers the interaction relations across modalities while ignoring the sequential dependencies within modalities. Particularly, the lack of item ID information significantly reduces the recommendation performance. MFN4Rec-g and MFN4Rec-add consider the sequential dependencies with modalities, while not fully capturing the complex interaction relations across modalities, and thus they cannot perform very well. In summary, such observations demonstrate the significance of effectively capturing both sequential dependencies within modalities and interaction relations across modalities.

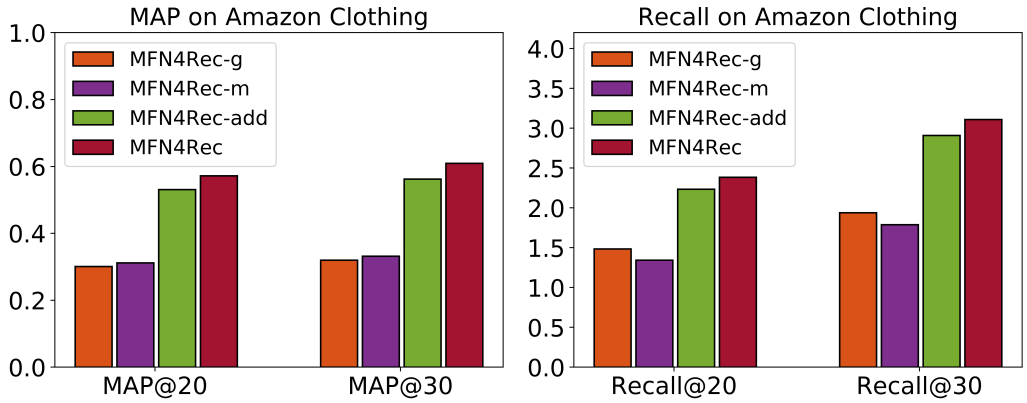


**Figure 2.** MAP and Recall of MFN4Rec and its variants.

## 5. Conclusions

Most of the existing sequential recommendation algorithms are not able to effectively utilize the multi-modal auxiliary information to capture the complex dependencies and interaction relations embedded in users' sequential behaviours. Aiming at this problem, we proposed a novel multi-modal sequential recommendation algorithm called MFN4Rec to effectively incorporate the item's images and text description information. Thanks to the particular design, MFN4Rec can effectively model both sequential dependencies within modalities and the interaction relations across modalities for more accurate sequential recommendations. The experiments on real-world e-commerce datasets demonstrate the effectiveness of MFN4Rec and the significance of modeling both sequential dependencies within modalities and interaction relations across modalities.

**Author Contributions:** Conceptualization, methodology and writing, S.W.; software, experiments and validation, T.Q.; supervision, W.W.; revision, Y.D. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was supported by China Postdoctoral Science Foundation No. 2020M681264, Anhui Natural Science Foundation Grant 1908085MF178 and Anhui Key Research and Development Plan Project Grant 202104b11020031.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The used experimental datasets are available at https://jmcauley.ucsd.edu/data/amazon/ accessed on 17 July 2021.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Qu, T.; Wan, W.; Wang, S. Visual content-enhanced sequential recommendation with feature-level attention. *Neurocomputing* **2021**, *443*, 262–271. [CrossRef]
2. Hidasi, B.; Quadrana, M.; Karatzoglou, A.; Tikk, D. Parallel recurrent neural network architectures for feature-rich session-based recommendations. In Proceedings of the Tenth ACM Conference on Recommender Systems, Boston, MA, USA, 15–19 September 2016; pp. 241–248.

3. Cui, Q.; Wu, S.; Liu, Q.; Zhong, W.; Wang, L. MV-RNN: A multi-view recurrent neural network for sequential recommendation. *IEEE Trans. Knowl. Data Eng.* **2020**, *32*, 317–331. [CrossRef]
4. Zadeh, A.; Liang, P.P.; Mazumder, N.; Poria, S.; Cambria, E.; Morency, L.P. Memory fusion network for multi-view sequential learning. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
5. Yap, G.E.; Li, X.L.; Philip, S.Y. Effective next-items recommendation via personalized sequential pattern mining. In Proceedings of the Seventeenth International Conference on Database Systems for Advanced Applications, Busan, Korea, 15–19 April 2012; pp. 48–64.
6. Feng, S.; Li, X.; Zeng, Y.; Cong, G.; Chee, Y.M.; Yuan, Q. Personalized ranking metric embedding for next new POI recommendation. In Proceedings of the Twenty Fourth International Conference on Artificial Intelligence, Buenos Aires, Argentina, 25–31 July 2015; pp. 2069–2075.
7. Rendle, S.; Freudenthaler, C.; Schmidt-Thieme, L. Factorizing personalized markov chains for next-basket recommendation. In Proceedings of the Nineteenth International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 811–820.
8. Hidasi, B.; Karatzoglou, A.; Baltrunas, L.; Tikk, D. Session-based recommendations with recurrent neural networks. In Proceedings of the Fourth International Conference on Learning Representations, San Juan, Puerto Rico, 2–4 May 2016; pp. 1–10.
9. Hidasi, B.; Karatzoglou, A. Recurrent neural networks with top-k gains for session-based recommendations. In Proceedings of the 27th ACM International Conference on Information and Knowledge Management, Turin, Italy, 22–26 October 2018; pp. 843–852.
10. Wu, C.Y.; Ahmed, A.; Beutel, A.; Smola, A.J.; Jing, H. Recurrent recommender networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining, Cambridge, UK, 6–10 February 2017; pp. 495–503.
11. Quadrana, M.; Karatzoglou, A.; Hidasi, B.; Cremonesi, P. Personalizing session-based recommendations with hierarchical recurrent neural networks. In Proceedings of the Eleventh ACM Conference on Recommender Systems, Como, Italy, 27–31 August 2017; pp. 130–137.
12. Wang, S.; Hu, L.; Cao, L.; Huang, X.; Lian, D.; Liu, W. Attention-based transactional context embedding for next-item recommendation. In Proceedings of the Thirty Second AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 2532–2539.
13. Tang, J.; Wang, K. Personalized top-n sequential recommendation via convolutional sequence embedding. In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, Marina Del Rey, CA, USA, 5–9 February 2018; pp. 565–573.
14. Tuan, T.X.; Phuong, T.M. 3D convolutional networks for session-based recommendation with content features. In Proceedings of the Eleventh ACM Conference on Recommender Systems, Como, Italy, 27–31 August 2017; pp. 138–146.
15. Ma, C.; Ma, L.; Zhang, Y.; Sun, J.; Liu, X.; Coates, M. Memory augmented graph neural networks for sequential recommendation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 5045–5052.
16. Hsu, C.; Li, C.T. RetaGNN: Relational Temporal Attentive Graph Neural Networks for Holistic Sequential Recommendation. In Proceedings of the Web Conference 2021, Virtual, 19–23 April 2021; pp. 2968–2979.
17. Kang, W.C.; McAuley, J. Self-attentive sequential recommendation. In Proceedings of the 2018 IEEE International Conference on Data Mining (ICDM), Singapore, 17–20 November 2018; pp. 197–206.
18. Zhang, T.; Zhao, P.; Liu, Y.; Sheng, V.S.; Xu, J.; Wang, D.; Liu, G.; Zhou, X. Feature-level Deeper Self-Attention Network for Sequential Recommendation. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, Macao, China, 10–16 August 2019; pp. 4320–4326.
19. Ren, R.; Liu, Z.; Li, Y.; Zhao, W.X.; Wang, H.; Ding, B.; Wen, J.R. Sequential recommendation with self-attentive multi-adversarial network. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 89–98.
20. Wang, S.; Hu, L.; Cao, L. Perceiving the next choice with comprehensive transaction embeddings for online recommendation. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Skopje, Macedonia, 18–22 September 2017; pp. 285–302.
21. Garg, D.; Gupta, P.; Malhotra, P.; Vig, L.; Shroff, G. Sequence and time aware neighborhood for session-based recommendations: Stan. In Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, Paris, France, 21–25 July 2019; pp. 1069–1072.
22. Li, J.; Wang, Y.; McAuley, J. Time interval aware self-attention for sequential recommendation. In Proceedings of the 13th International Conference on Web Search and Data Mining, Houston, TX, USA, 3–7 February 2020; pp. 322–330.
23. Ye, W.; Wang, S.; Chen, X.; Wang, X.; Qin, Z.; Yin, D. Time Matters: Sequential Recommendation with Complex Temporal Information. In Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, Xi'an, China, 25–30 July 2020; pp. 1459–1468.
24. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
25. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
26. Pennington, J.; Socher, R.; Manning, C.D. Glove: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1532–1543.

27. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
28. Gatys, L.A.; Ecker, A.S.; Bethge, M. Image style transfer using convolutional neural networks. In Proceedings of the Twenty Eighth IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 2414–2423.
29. Tang, J.; Belletti, F.; Jain, S.; Chen, M.; Beutel, A.; Xu, C.; H. Chi, E. Towards neural mixture recommender for long range dependent user sequences. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 1782–1793.
30. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 649–657.
31. Rendle, S.; Freudenthaler, C.; Gantner, Z.; Schmidt-Thieme, L. BPR: Bayesian personalized ranking from implicit feedback. In Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, Montreal, QC, Canada, 18–21 June 2009; pp. 452–461.
32. Yu, F.; Liu, Q.; Wu, S.; Wang, L.; Tan, T. A dynamic recurrent model for next basket recommendation. In Proceedings of the Thirty Ninth International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 729–732.
33. He, R.; McAuley, J. VBPR: Visual bayesian personalized ranking from implicit feedback. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; pp. 144–150.