IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 21, 2020, accepted January 31, 2020, date of publication March 25, 2020, date of current version April 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2978272

Visual and Textual Jointly Enhanced Interpretable Fashion Recommendation

QIANQIAN WU [1], PENGPENG ZHAO [1], AND ZHIMING CUI [2]

[1]School of Computer Science and Technology, Institute of Artificial Intelligence, Soochow University, Suzhou 215006, China
[2]School of Electronic and Information Engineering, Suzhou University of Science and Technology, Suzhou 215009, China

Corresponding author: Pengpeng Zhao (ppzhao@suda.edu.cn)

This work was supported in part by the NSFC under Grant 61876217, Grant 61872258, and Grant 61728205, and in part by the Suzhou Scienceand Technology Development Program under Grant SYG201803.

ABSTRACT With the rapid development of online shopping, interpretable personalized fashion recommendation using image has attracted increasing attention in recent years. The current work has been able to capture the user's preferences for visible features and provide visual explanations. However, they ignored the invisible features, such as the material and quality of the clothes, and failed to offer textual explanations. To this end, we propose a Visual and Textual Jointly Enhanced Interpretable (VTJEI) model for fashion recommendations based on the product image and historical review. The VTJEI can provide more accurate recommendations and visual and textual explanations through the joint enhancement of textual information and visual information. Specifically, we design a bidirectional two-layer adaptive attention review model to capture the user's visible and invisible preferences to the target product and provide textual explanations by highlighting some words. Moreover, we propose a review-driven visual attention model to get a more personalized image representation driven by the user's preference obtained from the historical review. In this way, we not only realize the joint enhancement of visual information and textual information but also provide a visual explanation by highlighting some regions. Finally, we performed extensive experiments on real datasets to confirm the superiority of our model on Top-N recommendations. We also built a labeled dataset for evaluating our provided visible and invisible explanations quantitatively. The result shows that we can not only provide more accurate recommendations but also can provide both visual and textual explanations.

INDEX TERMS Explainable recommendation, fashion recommendation, visual and textual explanations.

I. INTRODUCTION

Nowadays, when buying fashion products online, the user's decisions are primarily affected by the appearance of products [1]. However, the invisible features that the user cannot observe from the image, such as the material and quality of the clothes, also affect the user's decisions. Therefore, Our work focuses on introducing reviews and capturing the user's visible and invisible preferences together with the image.

Some previous work has made many efforts to exploit product images for fashion recommendations in recent years. Most of the existing methods use pre-trained convolution models to convert the entire fashion image into a fixed-length global image embedding [2]–[5], which ignore visual preference for the specified user and fail to generate reasonable visual explanations. To solve this problem, Chen *et al.* [6] learned an attention model over many pre-segmented image regions to discover fine-grained visual preference, and use user review to supervise the acquisition of more comprehensive user preferences weakly. Besides, considering the semantic attributes of the product, Hou *et al.* [7] defined a readily available interpretable semantic space to extract the pre-segmented image regions corresponding to the attributes and depicted the preferences of different users for different semantic attributes through the attention network. These methods capture the user's visible preferences and provide visual explanations. Still, they ignore the capture of invisible preferences and fail to provide invisible explanations due to the limitation of image information.

User reviews, which contain a lot of auxiliary information, has gained a lot of attention in Top-N interpretable recommendations [8], [9]. In particular, Seo *et al.* [10] proposed a convolutional neural network with both local and global attention to obtaining more complex features of users

The associate editor coordinating the review of this manuscript and approving it for publication was Jeon Gwanggil.

and products from reviews and explaining with attention weight. Zhang *et al.* [11] used phrase-level sentiment analysis to explicit product features and user opinions, which not only keep a high prediction but also generate explainable recommendations. Chen *et al.* [12] proposed a hierarchical co-attention selector and an encoder-selector-decoder architecture to fully exploiting the correlations between the recommendation task and the explanation task. Therefore, considering that reviews contain user's visible and invisible preferences, we can introduce historical review into image-based interpretable recommendations, where the acquisition of invisible preferences can make up for existing methods that can only capture visible preferences and we can also improve recommendation performance by jointly enhancing textual information and visual information.

In order to introduce historical review into an image-based recommendation system, we propose a novel Visual and Textual Jointly Enhanced Interpretable Model (VTJEI). In this model, we propose a bidirectional two-layer adaptive attention review model to capture the user's visible and invisible preferences to the target product and provide textual explanations by highlighting some words. Then, we propose a review-driven visual attention model to get a more personalized image representation driven by the user's preference obtained from the historical review. In this way, we not only realize the joint enhancement of visual information and textual information but also provide a visual explanation by highlighting some regions. Besides, we also take the relative review as a weak supervision signal to highlight image regions. Based on this model, we can provide not only accurate recommendations but also provide corresponding visual and textual explanations for each recommended item.

The main contributions of this model are summarized as follows:

- To our knowledge, we are the first to introduce historical review into an image-based interpretable recommendation system. In this way, we implement visual and textual jointly enhanced interpretable fashion recommendations in the field of fashion recommendation.
- We develop a novel framework, Visual and Textual Jointly Enhanced Interpretable (VTJEI) model for fashion recommendation. VTJEI obtains the user's visible and invisible preferences for the target product by a bidirectional two-layer adaptive attention review model and implement joint enhancement of visual information and textual information by a review-driven image attention model.
- We conduct extensive experiments on four public benchmarks, demonstrating the effectiveness of VTJEI and its interpretability in understanding user's visible and invisible preferences.

II. RELATED WORK

In this section, we introduce some of the fashion recommendations and interpretable recommendations that have been most relevant to our work in recent years. Through the introduction of related work, we will highlight the differences between our method and them.

A. FASHION RECOMMENDATION

In recent years, fashion recommendations have become increasingly popular in industrial and academic circles. In order to efficiently discover user behavior patterns, many effective recommendation models have been proposed. Generally speaking, most of these methods are learning visual image preferences. For example, McAuley *et al.* [13] are devoted to using objects' appearances to discover their implicit relationships, to find complementary products and alternative products. Kang *et al.* [14] learned "fashion-aware" image representations by training the image representation (from the pixel level) and the recommender system jointly to improve recommendation performance. Han *et al.* [15] proposed to jointly learn a visual-semantic embedding and the compatibility relationships among fashion items in an end-to-end fashion. Chen *et al.*. [6] proposed an attention model to depict the user's attention to different pre-segmented image regions, and the user's review information was used to supervise the attention model to ensure the correct positioning of the image by the attention model. Hou *et al.* [7] considered the semantic attributes of the product, a readily available interpretable semantic space was defined to extract the pre-segmented image regions corresponding to the attributes and depict the preferences of different users for different semantic attributes through the attention network.

Virtually, a fashion picture in the above method either ignores the different preferences of users for different parts of the fashion picture because it is transformed into a fixed-length vector. Or, although the user's personalized preferences for fashion pictures are considered, the limited display capability of the image is ignored, then ignore those invisible preferences. In our model, however, we obtain a large number of visible and invisible preferences of users for the target item from historical reviews. We have not only improved recommendation performance but also implemented the collaborative interpretation of image and text.

B. EXPLAINABLE RECOMMENDATION

Because recommendations with relevant explanations can greatly improve user's credibility and shopping experience, researches on explainable recommendations have become more and more popular in recent years [16]–[19]. Existing interpretable models usually explain related recommendations based on textual reviews. Based on this review text information, some of the early methods, such as HFT [20] and RBLT [21] mainly focused on combining latent rating dimensions (such as those of latent-factor recommender systems) with latent review topics (such as those learned by topic models like LDA). Zheng *et al.* [22] Considered the limitation of the "Bag-Of-Words(BOG)" in the topic model when capturing the semantic information of the review, it is believed that this may degrade the recommended performance and the explainable performance. Fortunately, the rise of deep neural networks has brought some light to this problem, and recently there are many related works dedicated to building deep interpretable recommendation models in order to mine more effective semantics from review information. These methods can be divided into two categories. On the one hand, many methods [10], [23]–[25] provide explanation in an "extractive" way. In particular, D-Attn [10] and NARRE [23] use attention networks to identify important parts of reviews under user-item ratings supervision. Since user's previous reviews can reflect a lot of user preference information, taking into account that this information should be closely related to the products that the user will buy, CARL [25] proposed a novel context-aware user-item representation learning model for rating prediction and MPCN [24] proposed a review-by-review pointer-based learning scheme that extracts important reviews from user and item reviews and subsequently matches them in a word-by-word fashion. On the other hand, many models provide explanations in a "generating" manner [26]–[29]. Instead of extracting relevant information from existing reviews for interpretation, these methods automatically generate complete natural language sentences for interpretation. In particular, NRT [26], gC2S [29], and NOR [27] use recurrent neural networks (RNN) and some other variants to generate natural language interpreted sentences. ExpansionNet [28] further combines product "aspects" to provide more diverse sentence interpretations.

Although the goal of the above method and our method is to provide interpretable recommendations, our method uses both user review information and product images to provide joint enhanced recommendations.

III. PROBLEM FORMULATION

In this section, we introduce our problem definition. Suppose there is a user set $U = \{u_1, u_2, \dots u_n\}$ and an item set $V = \{v_1, v_2, \dots, v_m\}$. We collect all the interactions between the user and the item to form the interaction set $O = \{(u,v)|$ user u has interacted with item $v.\}$. Each interaction is accompanied by a corresponding real review. We define the review of user u to item i as: let $W_{uv} = \{w_{uv}^1, w_{uv}^2, \dots, w_{uv}^{l_{uv}}\}(u \in U, v \in V)$, where w_{uv}^t is the t-th word, and l_{uv} is the length of the review. The set of all reviews of interactions is defined as $W = \{W_{uv}|(u, v) \in O\}$. Each item v has a corresponding image. We use deep convolutional neural networks (CNNs) [30] to process the image of item v into the following representation: $F_v = [f_v^1; f_v^2; \dots f_v^h] \in \mathbb{R}^{D*h}$, where$f_v^k \in \mathbb{R}^D$ is a D dimensional vector corresponding to the k-th spatial region of the image, and h is the number of the regions. Accordingly, the set of all item's visual features is represented as $F = \{F_v|v \in V\}$. The historical review of a user u are $r_u = (\mathcal{D}_{u,1}, \dots, \mathcal{D}_{u,l_d})$, where l_d denotes the maximum number of reviews. Each review $\mathcal{D}_{u,1}$ is denoted by a set of words in the review. The set of all user's historical reivew is represented as $R_u = \{r_u|u \in U\}$. Similarly, $r_v = (\mathcal{D}_{v,1}, \dots, \mathcal{D}_{v,l_d})$ represents the historical reviews of item v and the set of all item's historical reivew is represented as $R_v = \{r_v|v \in V\}$.

Formally, given the fashion dataset $\{U, V, O, F, W, R_u, R_v\}$, we need to learn a predictive function f. When given a user-item pair (u, v), the function f can predict the probability that user u likes item v. After ranking the predicted score of all fashion items, the fashion item with the highest score is recommended to the user u and accompany the item with the visual and textual explanation.

IV. THE VTJEI MODEL

In this section, we introduce our model. The overall framework of the model is shown in Figure 1. It consists of two parts: The bidirectional two-layer adaptive attention model, which can obtain the user's preference for the target item; The review-driven attention model with the review enhanced model supervision, which can obtain the personalized representation of the image. In particular, we first introduce a bidirectional two-layer adaptive attention review model, as shown in the gray dotted box on the left side of Figure 1. Then we explain the review-driven attention model, as shown in the gray dotted box on the right side of Figure 1. Finally, we introduce the optimization function objective.

A. BIDIRECTIONAL TWO-LAYER ADAPTIVE ATTENTION REVIEW MODEL

As mentioned earlier, the reviews of the user's previous purchase records contain a large amount of implicit information that reflects the user's preferences, and the relevant reviews of an item contain much attribute information about this item. Intuitively, if a user's reviews and a product's reviews exist some similar expression, the user is likely to buy this product. So, on the basis of user historical reviews and product historical reviews, we can predict whether users like this product.

Based on the above analysis, in order to obtain user preference information to the target product from user historical reviews and historical product reviews, we propose a bidirectional two-layer adaptive attention model upon the user historical reviews and product historical reviews. The bidirectional two-layer adaptive attention model contains two parts: filtering user reviews with the product and filtering item reviews with the user. The relevant details are as follows.

Filtering User Reviews With Product: Formally, assuming user u has N review sentences, they are defined as: $W_u = \{\mathcal{W}_1^u, \mathcal{W}_2^u, \dots, \mathcal{W}_N^u\}$. We user a two-layer adaptive attention network $G(,)$ on user review sentences W_u and target item v to obtain the most relevant information $R_{u,v}$ about target item in user reviews. The formula is designed as follows:

$$R_{u,v} = G(W_u, v) \tag{1}$$

where the two-layer adaptive attention network $G(,)$ will be described in detail below.

Filtering Item Reviews With User: Similarly, assuming item v has M review sentences, they are defined as: $W_v = \{\mathcal{W}_1^v, \mathcal{W}_2^v, \dots, \mathcal{W}_M^v\}$. We also use a two-layer adaptive attention network $G(,)$ on item review sentences W_v and user u

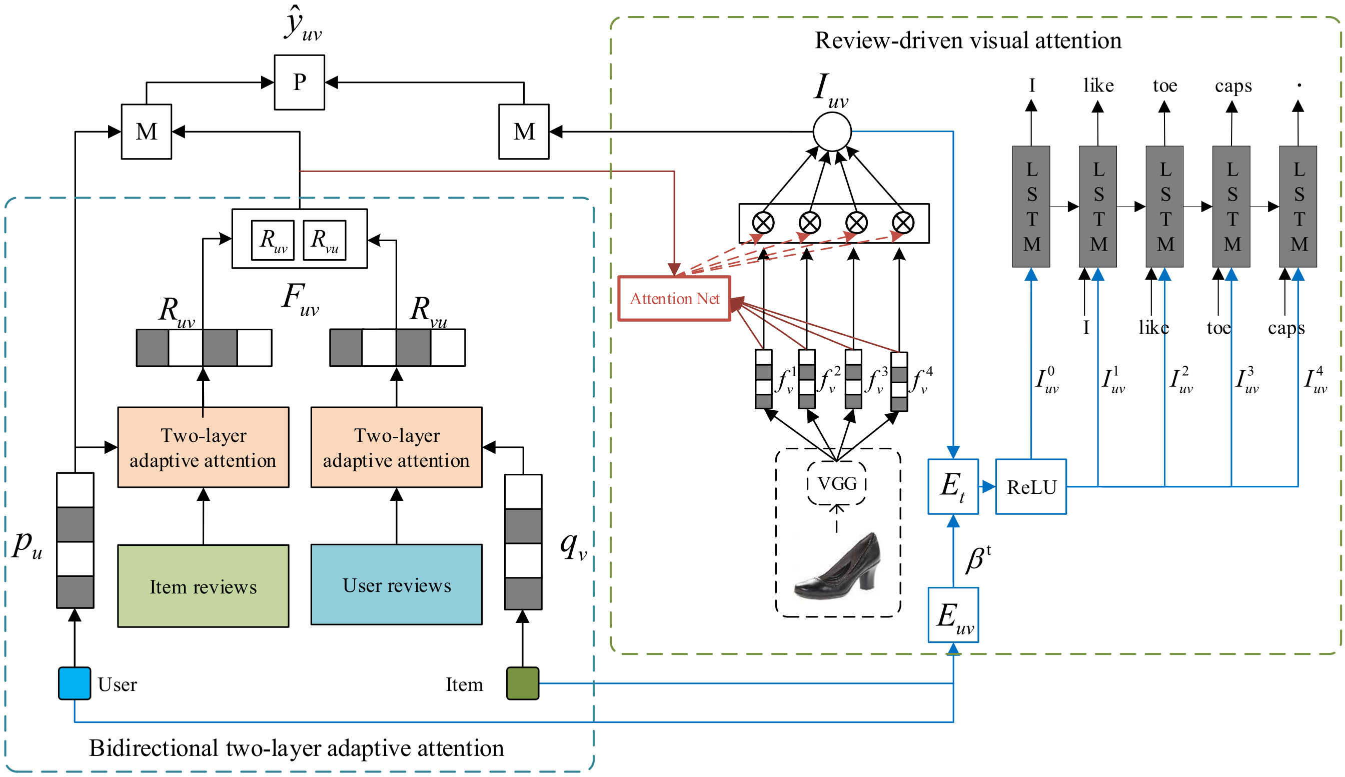


FIGURE 1. The overall framework of the VTJEI model. The dashed box on the left is a bidirectional two-layer adaptive attention model, which can obtain the user's preference for the target item; the dashed box on the right is a review-driven visual attention model, which can obtain a personalized image representation of the item.

to obtain the most relevant information $R_{v,u}$ about the user in item reviews. The formula is designed as follows:

$$R_{v,u} = G(W_v, u) \qquad (2)$$

Two-Layer Adaptive Attention Network: Then we detail our two-layer adaptive attention network $G(,)$ (see Figure 2). In particular, we suppose the input is the user u's reviews $W_u = \{\mathcal{W}_1^u, \mathcal{W}_2^u, \ldots, \mathcal{W}_N^u\}$ and target item v. Let $c^k = \{c_1^k, c_2^k, \ldots, c_T^k\}$ is the word embedding list of $\mathcal{W}_k^u$, where $c_i^k \in \mathbb{R}^d$ is the pre-trained word embedding by Bert for the i-th word in $\mathcal{W}_k^u$. In our model, the word vector of each word has been fused with context information after being processed by a convolutional neural network [30] (CNN). Because different reviews of the user have different relevance to the target item, different words in the same review also have different relevance to the target item. Here we use a two-layer attention network to get the preferences related to the target item in user reviews.

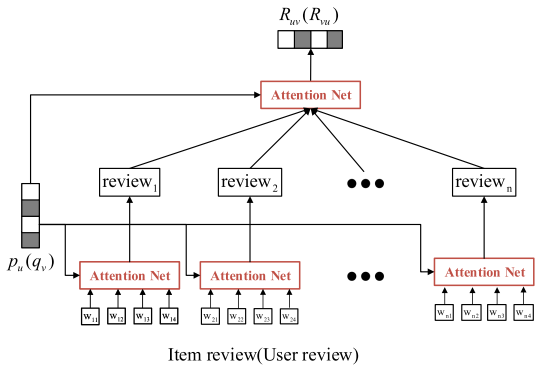


FIGURE 2. The two-layer adaptive attention model.

The word-level attention is designed as follows to get the review embedding under "item-aware" attention weight:

$$R_{uvk} = \sum_{j=1}^{T} \alpha_{vjk} . c_j^k, \qquad (3)$$

where α_{vjk} is j-th word of k-th review weight derived from an attention net as:

$$\begin{aligned} a_{vjk} &= E_2[ReLU(E_1[(W_v p_v) \odot (W_j c_j^k)])] \\ \alpha_{vjk} &= \frac{exp(a_{vjk})}{\sum_{j'=1}^{T} exp(a_{vj'k})} \end{aligned} \qquad (4)$$

where $p_v \in \mathbb{R}^k$ is the embedding of item v, $c_j^k \in \mathbb{R}^d$ is the word embedding of jth word of k-th review, $W_v \in \mathbb{R}^{s*k}$, $W_j \in \mathbb{R}^{s*d}$ are weighting parameters that project $p_v \in \mathbb{R}^k$ and $c_j^k \in \mathbb{R}^d$ into the same space. $E_1(.)$ and $E_2(.)$ are linear conversion operation, $ReLU$ is the Rectified Linear Unit (ReLU) [31]. "$\odot$" is the element-wise multiplication.

By the word-level attention network, we extract the features that most relevant to the target item in the review.

The review-level attention is designed as follows to get the user preference embedding under ''item-aware''attention weight:

$$R_{u,v} = \sum_{k=1}^{N} \alpha_{uvk} . R_{uvk}, \tag{5}$$

where α_{uvk} is k-th review attention weight derived from an attention net as:

$$\begin{aligned} a_{uvk} &= E_2'[ReLU(E_1'[(W_v' p_v) \odot (W_k' R_{uvk})])] \\ \alpha_{uvk} &= \frac{exp(a_{uvk})}{\sum_{k'=1}^{n} exp(a_{uvk'})} \end{aligned} \tag{6}$$

where $p_v \in \mathbb{R}^k$ is the embedding of item v, $R_{uvk} \in \mathbb{R}^d$ is the embedding of k-th review, $W_v' \in \mathbb{R}^{s*k}$, $W_k' \in \mathbb{R}^{s*d}$ are weighting parameters that project $p_v \in \mathbb{R}^k$ and $R_{uvk} \in \mathbb{R}^d$ into the same space. $E_1'(.)$ and $E_2'(.)$are linear conversion operation. By the review-level attention network, we extract the user preference $R_{u,v}$ that most relevant to the target item of all user reviews.

Finally, we will integrate the above two parts to obtain the user's preference for the target item as follows:

$$F_{uv} = (R_{u,v} \odot R_{v,u}) \tag{7}$$

B. REVIEW-DRIVEN VISUAL ATTENTION MODEL

As mentioned earlier, visible features have a great impact on user behavior in the fashion field. From the above work, we have obtained the user's visible and invisible preferences with word description to the target product. Intuitively, these visual preferences correspond to some regions in the product image. So different from previous work [6], which uses user-region aware attention to obtain fine-grained image representations, but no textual description of related visual preferences. We develop a review-driven visual attention model to obtain the more accurate fine-grained image representations, and the image area with higher weight also has relevant word description.

Similar to many previous works [2], [32], [33], we use CNN models to extract the regional features of fashion pictures. In particular, we use a pre-trained VGG-19 model. We input each image into the model and take the 14 * 14 * 512 feature vectors of its conv5 layer as the final representation of the image. 14 * 14 represents that the image is divided into 14 * 14 grid regions, and the 512-dimensional (D = 512) feature corresponds to the representation of each grid region in the image. Therefore, we obtain the image feature matrix $F_v \in \mathbb{R}^{h*D}$ for the item v, where each line $f_v^k \in \mathbb{R}^D$ corresponds to an image region and the total number of areas is $h = 196$.

In order to obtain a user's fine-grained visual preference with word annotation, we design a review-driven visual attention model on all region features of the image, and then calculate the final embedding of the item v by combining the feature matrix F_v and the ''review-region'' perceptual attention weight. The formula is as follows:

$$I_{u,v} = F_v \alpha_{uv} = \sum_{k=1}^{h} \alpha_{uvk} . f_v^k \tag{8}$$

where $\alpha_{uv} = \{\alpha_{uv1}, \alpha_{uv2}, \ldots, \alpha_{uvh}\}$ and α_{uvk} is the attention weights which is calculated as follows:

$$\begin{aligned} a_{uvk} &= E_2[ReLU(E_1[(W_p F_{uv}) \odot (W_f f_v^k)])] \\ \alpha_{uvk} &= \frac{exp(a_{uvk})}{\sum_{k'=1}^{h} exp(a_{uvk'})} \end{aligned} \tag{9}$$

where F_{uv} is the user preference obtained above, f_v^k is the k-th region feature of item v.

Finally, in order to ensure the accuracy of image positioning, we introduce user reviews into our model as weakly supervised signals. In specific, we model word generation based on Vanilla LSTM. Unlike the original LSTM model, we modified the LSTM by introducing the attentive embedding of item's image I_{uv} into the generation of words. Suppose the word list of user u commenting on item i is $w_{uv} = \{w_{uv}^1, w_{uv}^2, \ldots, w_{uv}^{l_{ij}}\}$, where l_{ij} is the length of the review. The calculation rules of the modified LSTM are as follows:

$$\begin{aligned} i_t &= \sigma(E_i[c_{uv}^{t-1}; h_{t-1}; I_{uv}^{t-1}]) \\ f_t &= \sigma(E_f[c_{uv}^{t-1}; h_{t-1}; I_{uv}^{t-1}]) \\ o_t &= \sigma(E_o[c_{uv}^{t-1}; h_{t-1}; I_{uv}^{t-1}]) \\ g_t &= tanh(E_g[c_{uv}^{t-1}; h_{t-1}; I_{uv}^{t-1}]) \\ e_t &= f_t \odot e_{t-1} + i_t \odot g_t \\ h_t &= o_t \odot tanh(e_t) \end{aligned} \tag{10}$$

where [.;.;.] concatenates input vectors, i_t, z_t, o_t and g_t are gate functions, $c_{ij}^t \in \mathbb{R}^d$ is the embedding of the input word w_{ij}^t, $h_t \in \mathbb{R}^z$ is the hidden state, and the I_{uv}^t is the attention image embedding considering different time steps, which is a contextual input determined by the the global embedding of user u and item v, the original attentive image embedding I_{uv}, and the hidden state h_t, that is:

$$I_{uv}^t = ReLU(E_I([\beta^t E_{PQ}([p_u; q_v]); (1 - \beta^t) I_{uv}])), \tag{11}$$

where q_v is the embedding of item v, q_u is the embedding of user u, $E_I(.)$, $E_{PQ}(.)$ are linear transformation, and $\beta^t = \sigma(w_T h_t)$ is a time-awaring gate function, which is used to model whether the current word is generated by the visible features of the image or those implicit features in the embedding of user and item.

C. OPTIMIZATION OBJECTIVE

Finally, the final likeness score for user u to item v is predicted by:

$$\hat{y}_{u,v} = P((p_u \odot F_{u,v}), (q_v \odot (W_I I_{uv}))) \tag{12}$$

where $W_I \in \mathbb{R}^{K*D}$ is the weighting parameter, P(.) is a L-layer neural network, $F_{u,v}$ is the user preference obtained from review, I_{uv} is the visual preference of user u to target item v, we use element-wise multiplication to combine user embedding p_u with its adaptive preference $F_{u,v}$ to obtain new user embedding and combine item embedding q_v with its adaptive visual embedding I_{uv} to obtain new item embedding. By matching new user embedding with the new item embedding, we predict the final likeness form user u to item v.

In the training, we use both score prediction and the review generation to supervise the learning process. Then, the final objective function that need to be maximized is:

$$\mathcal{L} = \sum_{i\in\mathcal{U}} \left(\sum_{j\in\mathcal{V}_+^i} log\sigma(\hat{y}_{ij}) + \sum_{j\in\mathcal{V}/\mathcal{V}_+^i} log\sigma(1-\hat{y}_{ij}) \right) + \beta \sum_{(i,j)\in\mathcal{O}} \sum_{t=1}^{l_{ij}} logp(w_{ij}^t|w_{ij}^{1:t-1}, I_{ij}^{t-1}) - \lambda||\Theta||_2^2 \qquad (13)$$

where β and λ are hyper parameter. Θ is the parameter set that needs regularization adjustment. $\mathcal{V}_+^i$ is the set of items that the user u purchased before. Corresponding to each positive instance, we uniformly sample a negative sample from the set of products that the user has never purchased. In this objective function, the first term is used to ensure that the positive example gets a Large scores and negative examples get a small score. The second term ensures that the currently predicted word is the same as the word in the real review. The last term aims to adjust the parameters to avoid overfitting.

V. EXPERIMENTS

A. EXPERIMENTS SETUP

1) DATASETS

In the public available fashion datasets, we choose the Amazon.com dataset[1] because this dataset provides the reviews and images we need. In particular, we choose the "clothes, shoes and jewelry" fashion data set on Amazon.com. In order to explore the performance of the model in different categories, we divide the data set into four small data sets corresponding to four categories [2] ("men", "woman", "boys and girls", and "baby"). The four data sets are Men, Women, Boys& Girls, and Babies. Table 1 shows the statistics of these datasets. We can see that they cover not only different genders and ages but also significant differences in the amount of data and sparseness.

TABLE 1. Statistics of the datasets in our experiments.

Dataset	User	Item	Word	Interaction	Density
Baby	1082	881	5454	1684	0.18 %
Boys&Girls	1382	1157	6313	1854	0.12 %
Men	24497	4687	30713	48224	0.042 %
Women	36330	15236	89652	146529	0.026%

2) BASELINES

We compare the performance of our method with the following most representative and advanced methods:

- ***BPR***: The visual bayesian personalized ranking [2] model is a well-known hybrid content-aware recommendation based on visual features.
- ***VBPR***: The visual bayesian personalized ranking [2] model is a well-known hybrid content-aware recommendation based on visual features.
- ***NRT***: The neural rating regression model [26]can simultaneously predict precise ratings and generate abstractive tips with good linguistic quality simulating user experience and feelings.
- ***NFM++***: The neural factorization machine (NFM) [34] is a deep network to model the relationships between higher-order features. For comparison, we add review information and global image vectors as contextual features to enhance the NFM.
- ***VECF***: The visually explainable collaborative filtering (VECF) [6] is the first work that not only provided accurate recommendations but also provided the novel personalized image interpretation in the fashion recommendation filed.

3) EXPERIMENTAL DETAILS

We initialize all the trainable parameters with a normalized distribution between [−1, 1], and they are learned by Adam optimizer [35] with a learning rate of 0.01. We set the embedding dimension k of users and items to {50, 100, 150, 200, 250, 300}. The weighting parameter β is searched in $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}\}$. In all experiments, the batch size and regularization parameters λ were fixed at 256 and 0.0001. The number of layers L of the predictive scoring network is set to 4.

4) EVALUATION METHOD

In our experiments, First, we take the last two interaction records of all users to form a set, and then randomly select 70% of the data from this set as the training set, and the rest as the test set. The other purchase records are used to extract historical review information. After our model has completed the learning, for each user, we predict all items and rank the prediction results, and then take the Top-N (10 in our experiment) items as the recommended list. For comparison, we use F_1 [36], Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [37] to evaluate different models.

B. EVALUATION ON RATING PREDICTION

In this section, we show the overall comparison between our VTJEI model and the baselines, and the results are presented in Table 2, we can see that:

- VBPR and NRT performed better than BPR in most cases. The result showed the effectiveness of the user reviews and product images for the task of Top-N recommendation. That's because compared to user/item ID

[1] http://jmcauley.ucsd.edu/data/amazon

[2] This category information comes from the meta information of the dataset, which is provided by http://jmcauley.ucsd.edu/data/amazon

TABLE 2. The results of comparing our model with the baselines. Numbers marked with stars are the best results for baseline. The bold numbers are the best results for all models and all numbers in the table are percentages.

Dataset	Measure	BPR	VBPR	NRT	NFM++	VECF	VTJEI	*Improv.*
Baby	F_1@10	2.326	2.612	2.159	3.185	3.573*	**3.797**	6.27% ↑
	HR@10	12.46	14.08	13.012	18.377	19.753*	**20.988**	6.25% ↑
	NDCG@10	4.552	5.182	4.627	7.283	7.838*	**10.499**	3.39% ↑
Boy&Girls	F_1@10	2.316	2.451	2.712	3.441	3.557*	**3.755**	5.57% ↑
	HR@10	12.34	12.89	14.512	19.076	**20.865***	20.652	-
	NDCG@10	5.861	5.905	6.254	7.329	7.675*	**8.958**	16.72% ↑
Men	F_1@10	2.436	3.215	3.359	3.936	4.443*	**4.758**	7.09% ↑
	HR@10	13.912	19.248	19.368	21.859	22.704*	**24.293**	7.00% ↑
	NDCG@10	5.723	9.739	9.916	10.191	10.218*	**11.543**	12.97% ↑
Women	F_1@10	2.592	2.731	3.167	3.503	3.768*	**4.246**	12.69% ↑
	HR@10	13.962	14.923	17.425	21.519	20.641*	**23.26**	12.69% ↑
	NDCG@10	6.348	7.534	7.528	7.913	8.266*	**10.028**	21.32% ↑

TABLE 3. Visual quantitative evaluation result table.

method	F_1					NDCG				
	M=1	M=2	M=3	M=4	M=5	M=1	M=2	M=3	M=4	M=5
random	0.776	1.352	1.629	2.108	2.527	2.892	2.901	2.823	3.419	3.507
VECF	2.169	3.246	4.325	4.531	4.624	7.628	7.159	7.276	6.315	6.448
VTJEI	2.559	3.668	5.017	4.984	5.133	9.230	8.563	8.513	7.199	7.415
Improv.	18% ↑	13% ↑	16% ↑	10% ↑	11% ↑	21% ↑	19% ↑	17% ↑	14% ↑	15% ↑

TABLE 4. Textual quantitative evaluation result table.

method	F_1					NDCG				
	N=1	N=2	N=3	N=4	N=5	N=1	N=2	N=3	N=4	N=5
random	0.512	0.927	1.246	2.018	2.423	2.415	2.416	2.924	3.157	3.238
VTJEI	0.773	1.492	2.219	3.491	3.974	4.226	3.624	3.624	4.988	5.019
Improv.	84% ↑	62% ↑	78% ↑	73% ↑	64% ↑	75% ↑	50% ↑	67% ↑	58% ↑	55% ↑

information, auxiliary information such as user reviews and product images can provide additional content for user/item profiling, thus increasing the opportunity to understand the true similarity between users and items.

- *T*hen we can see that: The performance of NFM++ is better than NRT and VBPR in most cases. Because NFM++ introduces not only rich auxiliary information from review and image, but also realizes high-level interaction of hidden features.
- Note that VECF was better than any other baseline, the underlying reason is that VECF takes into account the user's region-level preference for the item image, and better capture the user's fine-grained and accurate preference. However, the representation of an image in NFM++ is a fixed-length vector, which ignores that the user may only focus on a particular region of the image.
- Encouragingly, we find that our model was better than VECF across different datasets. As mentioned before, the reviews of previous purchase records of users contain plenty of user's visible and invisible preference information, and the previous reviews of items carry lots of attribute information about the item. This information can help us better understand the relationship between users and items. However, VECF used the attention mechanism to obtain the attention weight of the global user's embedding and item image embedding, which only captured the visual preferences and ignored the capture of invisible preferences. In contrast, our model uses a two-layer adaptive attention mechanism to obtain the user's visible and invisible preferences for the target item from historical review information and ultimately improves the recommendation performance.

C. PARAMETER ANANLYSIS

1) IMPACT OF EMBEDDING DIMENSION SIZE k

We explored how the embedded dimension size affected the recommended effect of our model by adjusting the size of different embedded dimensions k in the experiment. The results are shown in figure 3. We can see that the k value for achieving the best performance in different datasets is different. Moreover, we also observed that too large k could not achieve a good performance, which is consistent with

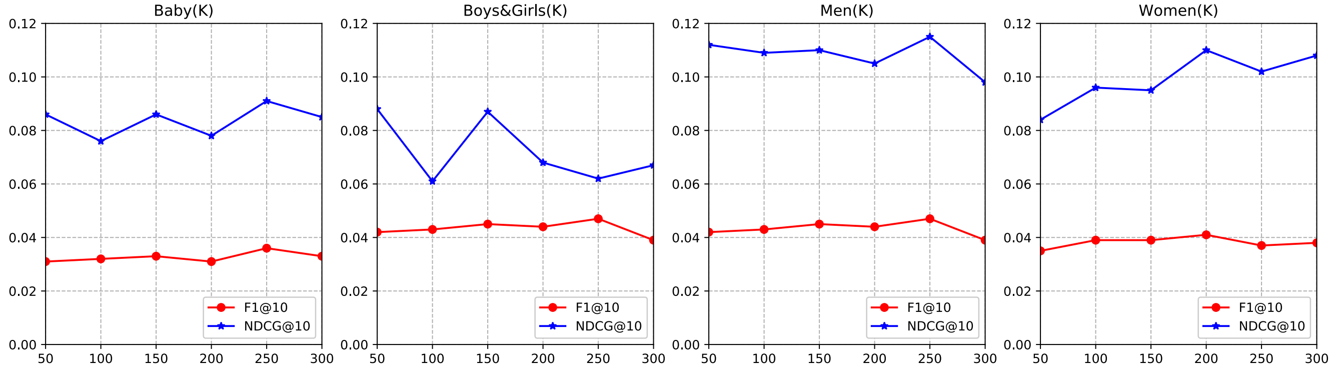


FIGURE 3. Influence of the embedding size *k* for the performance across various datasets.

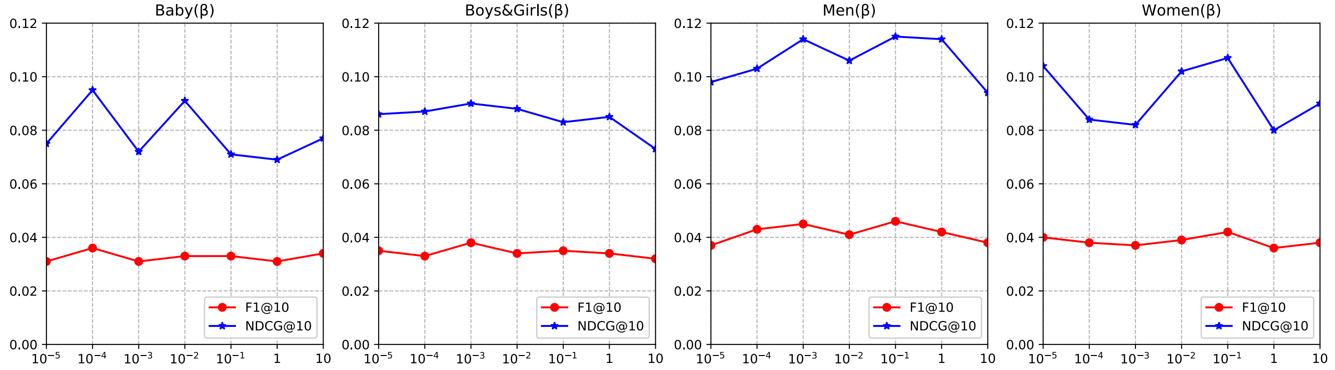


FIGURE 4. Influence of the hyperparameters β for the performance across various datasets.

many previous studies [34], [38]. The reason may be that too large k would increase the complexity of the model and lead to overfitting.

2) INFLUENCE OF THE HYPER PARAMETER β

The super parameter β is used to measure whether ratings predict implicit feedback is more important or user review information is more important. The results are shown in figure 4. We can see that the β value for achieving the best performance in different datasets is different, i.e., $\beta = 0.0001$ for Baby, $\beta = 0.1$ for Women, $\beta = 0.001$ for Boy&Girls and Men. We can see that the β value of all datasets is not too large. That is because if the β value is too large ($\beta = 10$), it means that we have focused too much on the part of review information supervision, but submerge the implicit feedback signal, and therefore, limit the final performance.

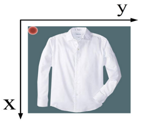
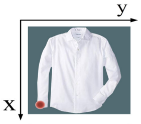
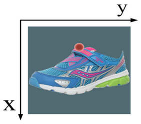
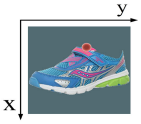
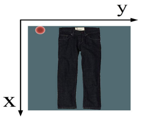
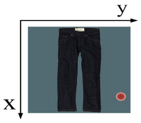
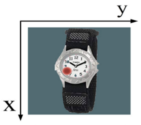
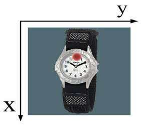
D. EVALUATION ON VISUAL AND TEXTUAL EXPLANATIONS

From the previous introduction, we know that once our model learned, we can provide each recommendation with visual explanations of the user's visible preferences and the textual explanations of the user's invisible preferences. In particular, we highlight some image regions with higher attention weights(i.e., larger α_{uvk}) as the visual explanations and highlight some essential words in history reviews with higher attention weights as the textual explanations. In this section, we evaluate whether the generated visual and textual interpretation can correctly reflect the user's real preference for the target item. First, we set up a real dataset with collective labels for quantitative analysis, and then give some examples for intuitive qualitative analysis.

1) QUANTITATIVE EVALUATION

As far as we know, this is the first effort at visual and textual jointly enhanced interpretable fashion recommendation, and there is no tag data that reflects user's visual and textual preferences in the current real dataset. In order to facilitate evaluation, we build a collectively labeled dataset in a crowd-sourcing manner. We asked volunteers to label 300 randomly selected user-item pairs from the Boy & Girl test set. The history reviews of users and the history reviews of items are also prepared. The image of each item is equally divided into 7*7=49 square regions. After observing the historical reviews of users and items and the real review of the user to this item, some workers need to point out 5 words that best reflect the user's invisible preferences and 5 words that best reflect the user's visible preferences. And identify 5 of the 49 areas of the image that are most relevant to the user's visible preferences.

TABLE 5. Examples of visual-texual explanation, where each row represents a user-item interaction. The second and third columns show the item images and the user review information. The fourth column show the key words or phrase provided by VTJEI, The blue keywords are related to the invisible features mentioned in the real reviews and the red keywords are related to the visible features mentioned in the real reviews. The last two columns present the highlighted image regions predicted by VECF and VTJEI, respectively, the highlighted image regions is related to the visible features mentioned by real review or the key word generated by VTJEI.

	Target Item	Texual Review	key words	Visual Explanation	
				VECF	VTJEI
1		I purchased this shirt for my son as it is always great to have a nice white dress shirt for any occasion you can dress it up or wear it casual with jeans. This shirt is a really nice quality shirt not too thin as some cheaper ones can be.	classy quality sleeve color old	 (x=1 , y=1)	 (x=13 , y=2)
2		My 3 year old daughter plays outside all day long and is really hard on shoes,these shoes look great even after months of hard playing.A couple of things I love about these shoes: they are easy for her to get on and off by herself, the tongue doesn't get messed up, and the insoles stay in place.	comfortable color take off size inside	 (x=4 , y=7)	 (x=4 , y=8)
3		The product is of great quality, color was as expected. Unfortunately, my son could not get into the size 12 as they were too slim. Depending on style he normally fits into 10 regular or 12 slim,but not these this time.	fit material quality price color	 (x=1 , y=2)	 (x=12 , y=13)
4		My son wanted a 'real' watch & not one of those cartoon characters with a display. So I got him this one. The watch seems well constructed and the numbers are easy to read.The movement also seems precise. Will update on the build quality in a few months.	precise velcro number clear fit	 (x=7 , y=6)	 (x=5 , y=7)

Then, other workers distinguished the N word features of each pair of user-item obtained by our model into invisible and visible categories, and determine whether the invisible features fall into the invisible features marked before. Because VECF cannot provide invisible explanations, our ground-truth is a random method, and the results by comparing our predicted invisible features against the ground-truth are presented in Table 4. We can see that the accuracy of the invisible features labeled by our model has been greatly improved compared to random labeling. Next, we identify M regions out of 196 candidates according to the learned attention weights(α_{uvk}), and if the region that has a corresponding word among the N words identified earlier and fell into the human-labeled regions, then we believe that the region is correct. Table 3 shows the results of comparing our predicted regions against the VECF. In Table3, we can see that our model's ability to capture visible features has improved to a certain extent compared to the VECF. That's because the historical reviews we introduce contain a lot of hidden information, which helps us better mine the real visible preferences of users.

2) QUALITATIVE EVALUATION

We quantitatively evaluate our model to provide intuitive understandings on the generated visual explanations. First, we select the region with the largest weight (i.e., α_{uvk}) as a highlighted region in both our model and VECF. Then, we select 5 most important words as the textual explanation according to the weight of the words in our model. Table 5 shows the results of the comparison between our model and VECF. From the result, we can see that:

Our model can not only highlight some region on the product images like VECF but also the words extracted by our model can effectively show the user's invisible preferences. For example, in Case 4, the number of the watch is highlighted by both VECF and our model. But our model differs from VECF is that our model also provides keywords corresponding to the highlighted area, like "number", "clear". Then, we can see in the real review that the watch is precise. For this invisible feature, our model can capture the word "precise" from historical reviews, which means our model can capture invisible features. Besides, when there are no relevant visible feature descriptions in real reviews, VECF cannot be correctly highlighted and cannot provide any valid interpretable. Still, our model can provide descriptions of invisible features related to real reviews, as shown in Case 1: The true description of this shirt is only invisible features such as white, quality. Due to the limited display ability of the image, VECF cannot provide an effective explanation. However, our model produces descriptions similar to real reviews, such as: "quality", "color", which confirms the superiority of our model in generating invisible interpretations. And we captured ''sleeve'' from the historical review and highlighted it on the image.

Also, we noticed some bad cases. For example, in Case 3, We can see that our model can capture invisible features (''quality''), but our model, like VECF, cannot provide meaningful visual explanations, because if the historical comments of users or items contain less meaningful information, our model cannot well describe the user's preference for the target item. In Case 2, VECF and our model both highlighted the tongue of the shoe. And the tongue of the shoe has mentioned in the real review. And our model also produces the phrase ''take-off'' similar to the invisible feature description of real reviews. However, because the reviews contain a lot of noise, the words corresponding to the visible features were not captured. Therefore, in the next step, we will study how to extract more meaningful information from historical reviews and explore better ways to strengthen the joint learning of textual information and visual information.

VI. CONCLUSION

In this article, we propose the use of historical review information to obtain the user's visible and invisible preferences for the target item and use this to drive the generation of visually interpretable images through the attention network. Experiments on real datasets prove that our model not only improves recommendation performance but also provides a more intuitive, vivid, and comprehensive explanation.

This article is the first work to enhance the recommendation effect and provide textual and visual interpretation by jointly learning textual and visual information, but there is still much work to be done in the future. First of all, the user's historical reviews may contain a lot of noise, which may lead to inaccurate extraction of user preferences. Then, since the visual image interpretation is just in its infancy, there is no unified evaluation standard at this stage. In the next step, we will focus on how to further dig more accurate and effective user preference information in the review information, and consider effectively mining the internal connection between user reviews and product images.

REFERENCES

[1] T. Zhang, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, D. Wang, G. Liu, and X. Zhou, ''Feature-level deeper self-attention network for sequential recommendation,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 4320–4326.

[2] R. He and J. McAuley, ''VBPR: Visual Bayesian personalized ranking from implicit feedback,'' in *Proc. AAAI*, Feb. 2016, pp. 144–150.

[3] Y. Hu, X. Yi, and L. S. Davis, ''Collaborative fashion recommendation: A functional tensor factorization approach,'' in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 129–138.

[4] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, and J. Ma, ''NeuroStylist: Neural compatibility modeling for clothing matching,'' in *Proc. ACM Multimedia Conf.*, 2017, pp. 753–761.

[5] C. Xu, P. Zhao, Y. Liu, V. S. Sheng, J. Xu, F. Zhuang, J. Fang, and X. Zhou, ''Graph contextualized self-attention network for session-based recommendation,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3940–3946.

[6] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, ''Personalized fashion recommendation with visual explanations based on multimodal attention network,'' in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2019, pp. 765–774.

[7] M. Hou, L. Wu, E. Chen, Z. Li, V. W. Zheng, and Q. Liu, ''Explainable fashion recommendation: A semantic attribute region guided approach,'' 2019, *arXiv:1905.12862*. [Online]. Available: http://arxiv.org/abs/1905.12862

[8] C. Xu, P. Zhao, Y. Liu, J. Xu, V. S. S. S. Sheng, Z. Cui, X. Zhou, and H. Xiong, ''Recurrent convolutional neural network for sequential recommendation,'' in *Proc. World Wide Web Conf. (WWW)*, 2019, pp. 3398–3404.

[9] P. Zhao, H. Zhu, Y. Liu, J. Xu, Z. Li, F. Zhuang, V. S. Sheng, and X. Zhou, ''Where to go next: A spatio-temporal gated network for next POI recommendation,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 5877–5884.

[10] S. Seo, J. Huang, H. Yang, and Y. Liu, ''Interpretable convolutional neural networks with dual local and global attention for review rating prediction,'' in *Proc. 11th ACM Conf. Recommender Syst. (RecSys)*, 2017, pp. 297–305.

[11] Y. Zhang, G. Lai, M. Zhang, Y. Zhang, Y. Liu, and S. Ma, ''Explicit factor models for explainable recommendation based on phrase-level sentiment analysis,'' in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2014, pp. 83–92.

[12] Z. Chen, X. Wang, X. Xie, T. Wu, G. Bu, Y. Wang, and E. Chen, ''Co-attentive multi-task learning for explainable recommendation,'' in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 2137–2143.

[13] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, ''Image-based recommendations on styles and substitutes,'' in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2015, pp. 43–52.

[14] W.-C. Kang, C. Fang, Z. Wang, and J. McAuley, ''Visually-aware fashion recommendation and design with generative image models,'' in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2017, pp. 207–216.

[15] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, ''Learning fashion compatibility with bidirectional LSTMs,'' in *Proc. ACM Multimedia Conf.*, 2017, pp. 1078–1086.

[16] X. Chen, H. Xu, Y. Zhang, J. Tang, Y. Cao, Z. Qin, and H. Zha, ''Sequential recommendation with and data mining,'' in *Proc. 18th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 108–116.

[17] Y. Zhang, ''Explainable recommendation: Theory and applications,'' 2017, *arXiv:1708.06409*. [Online]. Available: http://arxiv.org/abs/1708.06409

[18] Y. Zhang and X. Chen, ''Explainable recommendation: A survey and new perspectives,'' 2018, *arXiv:1804.11192*. [Online]. Available: http://arxiv.org/abs/1804.11192

[19] Y. Zhang, Y. Zhang, and M. Zhang, ''SIGIR 2018 workshop on Explain-Able recommendation and search (EARS 2018),'' in *Proc. 41st Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2018, pp. 1411–1413.

[20] J. McAuley and J. Leskovec, ''Hidden factors and hidden topics: Understanding rating dimensions with review text,'' in *Proc. 7th ACM Conf. Recommender Syst.*, 2013, pp. 165–172.

[21] Y. Tan, M. Zhang, Y. Liu, and S. Ma, ''Rating-boosted latent topics: Understanding users and items with ratings and reviews,'' in *Proc. IJCAI*, vol. 16, 2016, pp. 2640–2646.

[22] L. Zheng, V. Noroozi, and P. S. Yu, ''Joint deep modeling of users and items using reviews for recommendation,'' in *Proc. 10th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2017, pp. 425–434.

[23] C. Chen, M. Zhang, Y. Liu, and S. Ma, ''Neural attentional rating regression with review-level explanations,'' in *Proc. WWW*, 2018, pp. 1583–1592.

[24] Y. Tay, A. T. Luu, and S. C. Hui, ''Multi-pointer co-attention networks for recommendation,'' in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 2309–2318.

[25] L. Wu, C. Quan, C. Li, Q. Wang, B. Zheng, and X. Luo, ''A context-aware user-item representation learning for item recommendation,'' *ACM Trans. Inf. Syst.*, vol. 37, no. 2, pp. 1–29, Mar. 2019.

[26] P. Li, Z. Wang, Z. Ren, L. Bing, and W. Lam, ''Neural rating regression with abstractive tips generation for recommendation,'' in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 345–354.

[27] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, and M. De Rijke, ''Explainable outfit recommendation with joint outfit matching and comment generation,'' *IEEE Trans. Knowl. Data Eng.*, early access, Mar. 19, 2019, doi: 10.1109/TKDE.2019.2906190.

[28] J. Ni and J. McAuley, ''Personalized review generation by expanding phrases and attending on aspect-aware representations,'' in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 706–711.

[29] J. Tang, Y. Yang, S. Carton, M. Zhang, and Q. Mei, ''Context-aware natural language generation with recurrent neural networks,'' 2016, *arXiv:1611.09900*. [Online]. Available: http://arxiv.org/abs/1611.09900

[30] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[31] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, no. 1, 2013, p. 3.

[32] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with Item- and component-level attention," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 335–344.

[33] Q. Liu, S. Wu, and L. Wang, "DeepStyle: Learning user preferences for visual recommendation," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 841–844.

[34] X. He and T.-S. Chua, "Neural factorization machines for sparse predictive analytics," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. (SIGIR)*, 2017, pp. 355–361.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[36] G. Karypis, "Evaluation of item-based top-n recommendation algorithms," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, 2001, pp. 247–254.

[37] K. Järvelin and J. Kekäläinen, "IR evaluation methods for retrieving highly relevant documents," *ACM SIGIR Forum*, vol. 51, no. 2, pp. 243–250, Aug. 2017.

[38] Y. Zhang, Q. Ai, X. Chen, and W. B. Croft, "Joint representation learning for Top-N recommendation with heterogeneous information sources," in *Proc. ACM Conf. Inf. Knowl. Manage. (CIKM)*, 2017, pp. 1449–1458.

QIANQIAN WU received the bachelor's degree in engineering from the Shandong University of Science and Technology, in 2018. She is currently pursuing the master's degree with the School of Computer Science and Technology, Soochow University, Suzhou. Her main research interests are spatial data processing, recommendation systems, and data mining.

PENGPENG ZHAO received the Ph.D. degree in computer science from Soochow University, in 2008. He is currently a Professor with the School of Computer Science and Technology, Soochow University. From 2016 to 2017, he was a Visiting Scholar, working at the Data Mining and Business Analysis Laboratory, Rutgers University. He has published more than 60 articles in prestigious international conferences and journals, including ACM MM, AAAI, IJCAI, ICDM, CIKM, DASFAA, and ICME. He was a Program Committee Member of international conferences, such as AAAI, IJCAI, CIKM, and PAKDD. His current research interests include data mining, deep learning, big data analysis, and recommender systems.

ZHIMING CUI is currently a Professor with the Institute of Intelligent Information Processing and Application at Soochow University, China. He is currently an outstanding expert of Jiangsu Province, China. He presided four National Natural Science Foundation of China. He has published several articles in computer vision, data mining, image processing, and pattern recognition. His research interests include deep web, computer vision, image processing, and pattern recognition.