Personalized Fashion Recommendation with Visual Explanations based on Multimodal Attention Network

Towards Visually Explainable Recommendation

Xu Chen[1], Hanxiong Chen[2], Hongteng Xu[3], Yongfeng Zhang[2]
Yixin Cao[4], Zheng Qin[1**], Hongyuan Zha[5]
[1]Tsinghua University, [2]Rutgers University, [3]Duke University & InfiniaML, Inc.
[4]National University of Singapore, [5]Georgia Institute of Technology
[1]{xu-ch14, qingzh}@mails.tsinghua.edu.cn, [2]{hanxiong.chen, yongfeng.zhang}@rutgers.edu
[3]hongteng.xu@infiniaml.com, [4]caoyixin2011@gmail.com, [5]zha@cc.gatech.edu

ABSTRACT

Fashion recommendation has attracted increasing attention from both industry and academic communities. This paper proposes a novel neural architecture for fashion recommendation based on both image region-level features and user review information. Our basic intuition is that: for a fashion image, not all the regions are equally important for the users, *i.e.*, people usually care about a few parts of the fashion image. To model such human sense, we learn an attention model over many pre-segmented image regions, based on which we can understand where a user is really interested in on the image, and correspondingly, represent the image in a more accurate manner. In addition, by discovering such fine-grained visual preference, we can visually explain a recommendation by highlighting some regions of its image. For better learning the attention model, we also introduce user review information as a weak supervision signal to collect more comprehensive user preference. In our final framework, the visual and textual features are seamlessly coupled by a multimodal attention network. Based on this architecture, we can not only provide accurate recommendation, but also can accompany each recommended item with novel visual explanations. We conduct extensive experiments to demonstrate the superiority of our proposed model in terms of Top-N recommendation, and also we build a collectively labeled dataset for evaluating our provided visual explanations in a quantitative manner.

∗∗ Corresponding author.

---

1 INTRODUCTION

With the ever prospering of on-line shopping for fashion products, fashion recommendation has attracted increasing attention from both industry and academic communities. Different from other fields, user decisions in the fashion domain are highly dependent on the product appearance [12, 18], for example, people usually purchase a clothing only after browsing its images when shopping on the Internet. Following this character, recent years have witnessed many efforts on exploiting product images for fashion recommendation [12, 15, 33]. Despite effectiveness, existing methods mostly transform a whole fashion image into a fixed-length vector, which may limit themselves in three aspects: (i) intuitively, people may only care about a few regions of a fashion image, and different users may have their individual preferences. As exampled in Figure 1, according to the review information, the focus of user A mainly lies on the neck area, while user B may be more interested in the pocket region. Apparently, such fine-grained visual preference is important for understanding different users, and can derive more accurate user similarities for enhancing the utility of collaborative filtering. However, the global image embedding in existing methods is hard to discover and exploit user local preferences, which may arouse negative impacts on the final recommendation performance. (ii) Many irrelevant image regions, such as the white pants and shoes in Figure 1, are indiscriminately encoded into the global representation of the fashion image, which may introduce too much noise into the model learning process. (iii) Recommendation explainability is important for enhancing user shopping experience [45]. In the fashion domain, item appearance is significant to user behaviors, thus providing visual explanations can be both intuitive and effective. However, the global image embedding prohibits existing methods from discovering user specific visual preference, and therefore, fail to generate reasonable visual explanations.

For closing these gaps, we propose a visually explainable collaborative filtering (called **VECF** for short) method for more effective fashion recommendation. Our general idea is to represent a fashion image by learning an attention model over many pre-segmented regions. With the supervision of collaborative filtering information, the attention mechanism is expected to highlight valuable image regions, while degrade the impact of the noisy parts. By discovering such fine-grained user preference, our recommended items can also be explained visually, which is more vivid, convenient, attractive, and comprehensively-efficient as compared with traditional textual explanations [7, 26, 31, 46] for on line users in the fashion domain.

Figure 1: Example of users with different fine-grained visual preferences. User reviews may have partial correspondences to the fashion image. The pink italic and green bold fonts indicate the review information that can and cannot be aligned with some specific image regions.

Although this seems to be a promising direction, it is non-trivial due to the following challenges: **(i) Less informative supervision signal.** Previous fashion recommender models mostly base their supervision on the binary user implicit feedback. However, such signal is sparse and less informative in revealing user fine-grained visual preference, *i.e.*, discovering where a user is interested in on a fashion image. **(ii) Difficulties in selecting appropriate image segmentation method.** For a fashion image, an ideal segmentation strategy is to utilize image processing techniques (*e.g.*, object detection) to divide it semantically into regions, such as the neck, sleeve and body for a clothing. However, it is time-consuming and unscalable to define semantic labels and obtain their annotations for training in each fashion category. What's worse, it's hard to set a unified segmentation granularity, since user preferences are usually diverse and changeable, *e.g.*, some people may only care about the cuff of a clothing, while others may tend to take the sleeve as a whole. **(iii) Lack of evaluation dataset.** Last but not least, once the model learned, there is no publicly available dataset to quantitatively evaluate whether the provided visual explanations (*i.e.*, the learned attention weights) are reasonable.

For addressing these issues, we propose a multimodal attention network with fixed region proposals for fine-grained visual preferences modeling. To begin with, we introduce user review information for enhancing the model supervision signal. Comparing with implicit interaction data, user review is much more powerful in reflecting user opinions and sentiments (as shown in Figure 1), which may provide more thorough and constraint supervision for better learning the visual attention weights. From the model design perspective, the review information is modeled by a customized LSTM model, and to effectively couple different modalities, we seamlessly infuse visual features into the word generation process, which allows us to combine multimodal information in a unified framework. For practicability, we directly divide each fashion image into many small grids, such that they can be flexibly assembled into different preference granularities via the attention mechanism. Comparing with existing methods, our approach is not only able to improve the recommendation performance, but also can generate intuitive visual explanations for the recommended fashion products. We conduct extensive experiments on real-world datasets to verify

the effectiveness of our proposed models in terms of Top-N recommendation. And also, we build a collectively labeled dataset to evaluate our generated visual explanations from both quantitative and qualitative perspectives.

2 PROBLEM FORMULATION

For easy understanding, we formally define our problem in this section. Suppose we have a user set $\mathcal{U} = \{u_1, u_2, ..., u_N\}$ and an item set $\mathcal{V} = \{v_1, v_2, ..., v_M\}$. The interaction set between the users and items is defined as: $O = \{(i, j)|$user i has interacted with item $j.\}$. Each fashion item $j \in \mathcal{V}$ is attached with an image, which can be seen as the visual content of this item. By existing feature extraction methods, such as deep convolutional neural networks (CNNs) [32], we represent the image of item j as $\boldsymbol{F}_j = [\boldsymbol{f}_j^1; \boldsymbol{f}_j^2; ...; \boldsymbol{f}_j^h] \in \mathbb{R}^{D \times h}$, where $\boldsymbol{f}_j^k \in \mathbb{R}^D$ 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 items' visual features is denoted as $\mathcal{F} = \{\boldsymbol{F}_j | j \in \mathcal{V}\}$. In addition, we also have review information, from which we can collect more comprehensive user preferences and item properties. Let $\boldsymbol{w}_{ij} = \{w_{ij}^1, w_{ij}^2, ..., w_{ij}^{l_{ij}}\}$ $(i \in \mathcal{U}, j \in \mathcal{V})$ be the textual review of user i on item j, where w_{ij}^t is the t-th word, and l_{ij} is length of the review. We define the set of all user reviews as $\mathcal{W} = \{\boldsymbol{w}_{ij} | (i, j) \in O)\}$.

Formally, given a multimodal fashion dataset $\{\mathcal{U}, \mathcal{V}, O, \mathcal{F}, \mathcal{W}\}$, our task is to learn a predictive function f, such that for each user, it can accurately rank all the fashion items according to his/her preference. And further, the internal parameters or intermediate outputs should provide visual explanations (from $\mathcal{F}$) for the final recommended items.

3 THE VECF MODEL

In this section, we detail our model principle (see Figure 2). Specifically, we first describe the visual attention mechanism for fine-grained fashion image modeling. Then, we introduce user review information as a weak supervision signal to enhance the model learning process. At last, we present the overall optimization objective.

3.1 Fine-grained Visual Preference Modeling

As mentioned before, visual features are important factors that influence user behaviors in the fashion domain. It's intuitive that a user is unlikely to pay exactly the same attention to different regions of a fashion image. So different from previous work [12, 18], which mostly transform the whole image into a fixed vector and ignore the discrepancies of user preference upon different image regions, we derive an item image's embedding by attentively combining its pre-extracted region features, and utilize it to enhance the item representation for computing the final prediction.

Similar to many previous work [6, 12, 24], we extract regional features of a fashion image by CNN models. In specific, we feed each fashion image into the pre-trained VGG-19 [32] model, and use the $14 \times 14 \times 512$ feature map obtained from its $conv5$ layer as the final representation of the image. For each spatial point in the 14×14 grid, its 512-dimensional ($D = 512$) feature corresponds to the representation of a potential region of interest (ROI) in the
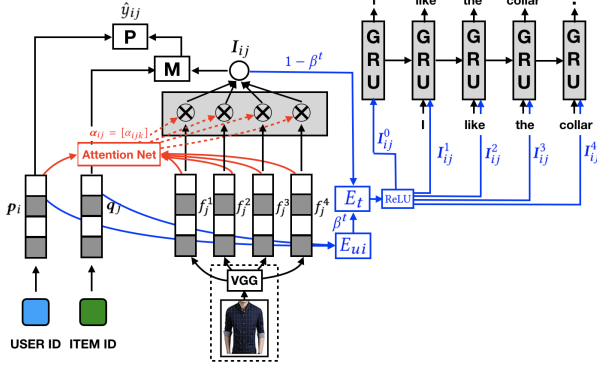
Figure 2: The proposed VECF model. The red lines indicate the attention mechanism designed for fashion image modeling. The blue lines highlight the modeling of user reviews.

image. Accordingly, for the image of item j, we obtain a feature matrix $\mathbf{F}_j \in \mathbb{R}^{D \times h}$, where each column $\mathbf{f}_j^k \in \mathbb{R}^D$ corresponds to an image region, and $h = 196$ is the number of total regions[1].

For representing a fashion image by taking user fine-grained visual preference into consideration, we design a visual attention mechanism upon the extracted region-level features (as shown in Figure 2). Mathematically, the final embedding of item j's image is computed by pooling feature matrix $\mathbf{F}_j$ under "user-region" aware attention weights as:

$$\mathbf{I}_{ij} = \mathbf{F}_j \boldsymbol{\alpha}_{ij} = \sum\nolimits_{k=1}^{h} \alpha_{ijk} \cdot \mathbf{f}_j^k, \tag{1}$$

where $\boldsymbol{\alpha}_{ij} = [\alpha_{ijk}] \in \mathbb{R}^h$ contains the attention weights jointly determined by the embedding of user i and the region feature $\mathbf{f}_j^k$, that is:

$$a_{ijk} = E_2[\text{ReLU}(E_1[(\mathbf{W}_u \mathbf{p}_i) \odot (\mathbf{W}_f \mathbf{f}_j^k)])]$$
$$\alpha_{ijk} = \frac{\exp(a_{ijk})}{\sum_{k'=1}^{h} \exp(a_{ijk'})} \tag{2}$$

where $\mathbf{p}_i \in \mathbb{R}^K$ is the embedding of user i, $\mathbf{W}_u \in \mathbb{R}^{s \times K}$, $\mathbf{W}_f \in \mathbb{R}^{s \times D}$ are weighting parameters that project $\mathbf{p}_i \in \mathbb{R}^K$ and $\mathbf{f}_j^k \in \mathbb{R}^D$ into the same space, $E_1(\cdot)$ and $E_2(\cdot)$ are linear transformations, ReLU is the Rectified Linear Unit (ReLU) [25]. "$\odot$" is the element-wise multiplication (or called Hadamard multiplication). The learned attention weights can reflect user fine-grained visual preference, which will be leveraged to provide visual explanations in our experiments.

3.2 Review enhanced Model Supervision

As mentioned before, on many fashion shopping platforms, people often express their opinions in the form of textual reviews. Comparing with the simple implicit feedback, such review information is typically more informative, pooling an extensive wealth of knowledge in revealing user preference. See the example in Figure 1, although both user A and B bought a same top (both users exhibited positive implicit feedback for the item), the aspects they cared about were quite different according to their posted reviews, *i.e.*, user A cared more about the fitting and the neck opening, while user B

---

[1] If more computational resources are available, we can also free the VGG component and fine-tune $\mathbf{F}_j$ to explore better performance.

was more interested in the clothing's quality and pocket. From the model learning perspective, the simply user implicit feedback will back propagate exactly the same gradient for these two training samples, while the additional modeling of review information can bring us with opportunities to bias the embeddings of user A and user B toward different directions to capture their ground truth preferences.

Based the above analysis, we introduce user review as a weak supervision signal into our model to help enhance the recommendation performance and explainability. A major challenge here is how to integrate heterogeneous information (*i.e.*, user review and product image) into a unified framework. We base the review information modeling on the well-known LSTM unit [42], and for coupling different feature modalities, we revise Vanilla LSTM by seamlessly infusing the attentive image embedding $\mathbf{I}_{ij}$ into the word generation process. In specific, given the review of user i on item j, *i.e.*, the word list $\mathbf{w}_{ij} = \{w_{ij}^1, w_{ij}^2, ..., w_{ij}^{l_{ij}}\}$, the computational rules of the modified LSTM are presented as follows:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(E_i([\mathbf{c}_{ij}^{t-1}; \mathbf{h}_{t-1}; \mathbf{I}_{ij}^{t-1}])) \\
\mathbf{f}_t &= \sigma(E_f([\mathbf{c}_{ij}^{t-1}; \mathbf{h}_{t-1}; \mathbf{I}_{ij}^{t-1}])) \\
\mathbf{o}_t &= \sigma(E_o([\mathbf{c}_{ij}^{t-1}; \mathbf{h}_{t-1}; \mathbf{I}_{ij}^{t-1}])) \\
\mathbf{g}_t &= \tanh(E_g([\mathbf{c}_{ij}^{t-1}; \mathbf{h}_{t-1}; \mathbf{I}_{ij}^{t-1}])) \\
\mathbf{e}_t &= \mathbf{f}_t \odot \mathbf{e}_{t-1} + \mathbf{i}_t \odot \mathbf{g}_t \\
\mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{e}_t)
\end{aligned} \tag{3}$$

where $[\cdot; \cdot; \cdot]$ concatenates input vectors, $\mathbf{i}_t$, $\mathbf{z}_t$, $\mathbf{o}_t$ and $\mathbf{g}_t$ are gate functions, each of which is obtained by applying a composite function (*i.e.*, sigmoid function $\sigma(\cdot)$ or hyperbolic tangent function $\tanh(\cdot)$ +linear embedding $E(\cdot)$) to the concatenated input. $\mathbf{c}_{ij}^t \in \mathbb{R}^O$ is the embedding of the input word w_{ij}^t, and $\mathbf{h}_t \in \mathbb{R}^Z$ is the hidden state.

The core of the revised LSTM unit is the temporal attentive image embedding $\mathbf{I}_{ij}^t$, which is a contextual input derived based on the original attentive image embedding $\mathbf{I}_{ij}$, the global embedding of user i and item j, and the hidden state $\mathbf{h}_t$, that is,

$$\mathbf{I}_{ij}^t = \text{ReLU}(E_I([\beta^t E_{PQ}([\mathbf{p}_i; \mathbf{q}_j]); (1 - \beta^t)\mathbf{I}_{ij}])), \tag{4}$$

where $\mathbf{q}_j$ is the embedding of item j, $E_I(\cdot)$, $E_{PQ}(\cdot)$ are linear transformations, and $\beta^t = \sigma(\mathbf{w}^\top \mathbf{h}_t)$ is a time-varying *gate function* to model whether the current word is generated from the image features or the user/item embeddings in a soft manner. Ideally, to the words having explicit correspondences to the image (*e.g.*, the "neck" and "pocket" in the review shown in Figure 1), β^t will be small and the attentive embedding of image $\mathbf{I}_{ij}$ contributes more to the temporal attentive embedding. On the contrary, to the words not having correspondences to the image (*e.g.*, the "nice quality" in user B's review in Figure 1), β^t will be large and the temporal attentive embedding relies more on the global user embedding $\mathbf{p}_i$ and item embedding $\mathbf{q}_j$.

Given the hidden state $\mathbf{h}_t$, we can predict the probability of the word at time step t by: $p(w_{ij}^t | w_{ij}^{1:t-1}, \mathbf{I}_{ij}^{t-1}) = \text{softmax}(\mathbf{W}_v \mathbf{h}_t)$, where $\mathbf{W}_v \in \mathbb{R}^{V \times Z}$ is a weighting matrix that project the hidden state into a "vocabulary-sized"(V) vector.

3.3 Optimization Objective

Based on the above components, the final likeness score in our model from user i to item j is predicted by:

$$\hat{y}_{ij} = P(\boldsymbol{p}_i,\ \boldsymbol{q}_j \odot (\boldsymbol{W}_I \boldsymbol{I}_{ij})), \tag{5}$$

where $\boldsymbol{W}_I \in \mathbb{R}^{K \times D}$ is the weighting parameter. $P(\cdot)$ is empirically specified as a L-layer neural network upon the concatenate operation due to its higher effectiveness on our datasets, that is, $P(\boldsymbol{x}, \boldsymbol{y}) = \phi_L(\phi_{L-1}(...\phi_1([\boldsymbol{x}; \boldsymbol{y}])))$, where ϕ_i is the sigmoid active function. In this predictive network, the pooling result $\boldsymbol{I}_{ij}$ reflects the visual preference from user i to different image regions of item j. The element-wise multiplication, which has been demonstrated to be efficient [2] for feature interaction modeling, is leveraged to combine item embedding $\boldsymbol{q}_j$ with its adaptive visual embedding $\boldsymbol{I}_{ij}$. By matching $\boldsymbol{q}_j \odot (\boldsymbol{W}_I \boldsymbol{I}_{ij})$ with the user's global embedding, the final likeness from user i to item j is predicted by taking user fine-grained visual preferences into consideration.

In the training phase, we supervise the learning process by both user implicit feedback and review information. The final objective function to be maximized is:

$$
\begin{aligned}
\mathcal{L} = \sum_{i \in \mathcal{U}} \left(\sum_{j \in \mathcal{V}_+^i} \log \sigma(\hat{y}_{ij}) + \sum_{j \in \mathcal{V}/\mathcal{V}_+^i} \log(1 - \sigma(\hat{y}_{ij})) \right) + \\
\beta \sum_{(i,j) \in O} \sum_{t=1}^{l_{ij}} \log\ p(w_{ij}^t | \boldsymbol{w}_{ij}^{1:t-1}, \boldsymbol{I}_{ij}^{t-1}) - \lambda \|\Theta\|_2^2,
\end{aligned}
\tag{6}
$$

where β and λ are hyper parameters. Θ is the set of parameters need to be regularized. $\mathcal{V}_+^i$ is the set of items that user i has bought before. Corresponding to each positive instance, we uniformly sample one item from the unpurchased item set $\mathcal{V}/\mathcal{V}_+^i$ as the negative instance. In this objective function, the first term is used to maximize the likelihood of user implicit feedback, the second term corresponds to the loss function that predicts current words from historical observations, and the last term aims to regularize the parameters to avoid over fitting. In our model, the parameters can be easily learned in an end-to-end manner, and once the framework converged, we are not only able to provide personalized recommendation for a target user according to the predicted score (*i.e.* $\hat{y}_{ij}$), but also can accompany each recommended item with visual explanations by highlighting some regions of its image according to the learned attention weights (*i.e.*, $\boldsymbol{\alpha}_{ij}$ for $\boldsymbol{I}_{ij}$). In our architecture, because the heavy neural language model is designed as an output component, it will not influence our model's runtime efficiency at test time, which is important for a practical recommender system. Specifically, suppose E_2 projects $\mathbb{R}^s$ to $\mathbb{R}^l$ and E_1 projects $\mathbb{R}^l$ to $\mathbb{R}$, the total complexity of the non-linear layers in the predictive function P is Q. At test time, our model's complexities for making prediction and providing visual explanation (for a user-item pair) only depend on equation (5) and (2), which is, $O(hs(K+D+l)+KD+Q)$ and $O(hs(K+D+l))$, respectively. In the following experiments, we will show that by jointly capturing the preferences from items' images and users' reviews, we can achieve both improved recommendation accuracy and reasonable visual explanations.

Table 1: Statistics of the datasets in our experiments. Different datasets cover various data characters, and the density ranges from 0.035% to 0.47%.

Dataset	#User	#Item	#Word	#Interaction	Density
Baby	2211	380	8636	3927	0.47%
Boys&Girls	6999	1026	17510	10780	0.15%
Men	23139	5690	77182	70949	0.054%
Women	35330	14383	107504	177611	0.035%

Table 2: Summary of the models in our experiments. We compare the specific information used in each model as well as the model depth.

Reference	Model	Information	Depth
[30]	BPR	-	shallow
[12]	VBPR	image	shallow
[21]	NRT	text	deep
[14]	NFM++	image+text	deep
-	VECF	image+text	deep

4 EXPERIMENTS

In this section, we evaluate our proposed model focusing on four research questions, that is,

- **RQ 1**: Whether our model can enhance the performance of fashion recommendation as compared with other state-of-the-art methods?

- **RQ 2**: How do different hyper-parameters in our model influence the final recommendation performance?

- **RQ 3**: What are the effects of different model components in our framework for the eventual results?

- **RQ 4**: Whether our generated visual explanations (*e.g.*, the learned visual attention weights) are reasonable for the recommended items?

We begin by introducing the experimental setup, and then report and analyze the experimental results to answer these research questions.

4.1 Experimental Setup

4.1.1 Datasets. There are many public available fashion datasets, including FashionCV [33], Amazon.com² [11], Tradesy.com [12], etc. Among these datasets, Amazon.com suits our problem best, because it is the only one that simultaneously provides us with both user review and product image information. To explore our model's capability on different categories, we split the fashion dataset (*i.e.*, "Clothing, Shoes and Jewelry") of Amazon.com into four different subsets related to Men, Women, Boys&Girls and Baby, respectively.³ The final statistics of the these datasets are summarized in Table 1. We can see they cover different genders and ages, and the data characters vary in both size and sparsity, e.g., "Baby" is a small and dense dataset, while "Women" is much larger, but sparser.

4.1.2 Evaluation methods. In our experiments, each user's 70% interactions are used for model training, while the remaining is left

---

Table 3: Summary of the performance for baselines and our model. The starred numbers are the best baseline results. The bolded numbers are the best performance of each column, and all numbers in the table are percentage numbers with '%' omitted.

Dataset	Baby			Boys&Girls			Men			Women		
Measure@10(%)	F_1	HR	NDCG	F_1	HR	NDCG	F_1	HR	NDCG	F_1	HR	NDCG
BPR	2.426	13.53	4.719	2.424	13.37	5.979	2.541	14.06	5.810	2.596	14.32	6.258
VBPR	2.728	15.29	6.584	2.543	13.99	5.961	3.524	19.16	9.566	2.779	15.08	7.110
NRT	2.513	13.89	5.514	2.977	16.43	7.522^*	3.635	19.80	9.799	3.167	17.33	7.568^*
NFM++	3.174^*	17.64^*	7.057^*	3.125^*	17.48^*	6.833	3.819^*	21.09^*	11.53^*	3.276^*	**19.34***	7.512
VECF	**3.244**	**17.94**	**8.036**	**3.253**	**17.97**	**8.862**	**4.716**	**25.75**	**12.16**	**3.421**	18.84	**8.987**

for testing. Once a model learned, we first rank all the items for each user, and then truncate the ranking list at position N (which is set as 10 in our experiments) to investigate the Top-N recommendation problem. For higher evaluation efficiency, we randomly sample 100 items for performance ranking the testing. For comparison, the widely used metrics, including F_1 [19], Hit Ratio (HR) and Normalized Discounted Cumulative Gain (NDCG) [17], are utilized to evaluate different models. In general, F_1 and HR measure the recommendation accuracy based on the overlapping between the recommended items and the actually interacted ones, while NDCG aims to evaluate the performance by taking the ranking positions of the correct items into consideration.

4.1.3 ***Baselines***. We adopt the following representative and state-of-the-art methods as baselines for performance comparison:

• **BPR:** The bayesian personalized ranking [30] model is a popular method for Top-N recommendation. We adopt matrix factorization as the prediction component.

• **VBPR:** The visual bayesian personalized ranking [12] model is a well known method for recommendation based on visual features. This model has been demonstrated to be very competitive in the field of fashion recommendation.

• **NRT:** The neural rating regression model [21] is a deep recommender based on user review information, where the textual features are incorporated as an output component. In the experiments, the original objective function designed for rating prediction is revised as the pairwise ranking loss of BPR to model user implicit feedback.

• **NFM++:** The neural factorization machine (NFM) [14] is a deep architecture for effective feature interaction modeling. For comparison, we enhance original NFM by inputting the review information as well as the global image vectors as contextual features. Finally, our model (see Figure 2) of visually explainable collaborative filtering is denoted as **VECF**. As our algorithm aims to model the relationship between users and items, we mainly compare our method with user-centered (user to item) models. We leave out the comparison with item-centered (item to item) models, such as IBR [27] and BPR-DAE [33], because performance discrepancies may be caused by the user models for personalization. For easy understanding, we summarize all the models compared in our experiments in Table 2.

4.1.4 ***Implementation details***. We initialize all the trainable parameters according to a uniform distribution in the range of $[-1, 1]$. And then, the parameters are learned by the Adam optimizer [20]

with a learning rate of 0.01. We tune the dimension of user/item embedding K in the range of $\{50, 100, 150, 200, 250, 300\}$. The weighting parameter β is searched in $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0, 10^1\}$. For all the experiments, the batch size and the regularization coefficient λ are empirically set as 256 and 0.0001, respectively. The number of predictive layers L is set as 4. For the review information, we first pre-process it by the Natural Language Toolkit[4], and then the word embeddings for each dataset are pre-trained based on the Skip-gram model[5]. Our experiments are conducted on a sever with 1 TITAN X GPU, 256G memory and 40 cores.

4.2 Top-N Recommendation (RQ1)

The overall comparison between our VECF model and the baselines are presented in Table 3, we can see:

• By integrating user reviews or product images, VBPR and NRT obtained better performance than BPR in most cases, which verifies the effectiveness of these information for the task of Top-N recommendation. The underlying reason can be that: comparing with user/item ID information, external knowledge, such as user review or product image, can provide additional contents for user/item profiling, which may bring additional opportunities to understand the real similarities between users or items for better collaborative filtering.

• NFM++ achieved better performance than the other baselines in most cases. This is not surprising, because NFM++ leveraged more information to assist user profiling, and the nested feature interaction modeling of NFM++ can effectively couple different modalities for more powerful expressiveness.

• Encouragingly, we find that our VECF model was better than NFM++ across different datasets in most cases. This result ascertains the effectiveness of our proposed model and positively answers **RQ1**. As mentioned before, profiling visual features is important for fashion recommendation. However, NFM++ learns a fixed vector to represent the whole fashion image, even though a user may be only interested in some particular regions. In contrast, our model utilize attention mechanism to focus on user favored image regions, which helps to better capture user preferences and eventually improve the recommendation performance.

4.3 Parameter analysis (RQ2)

In this section, we study how the embedding dimension K and the hyper parameters β influence our model's performance. Specifically,
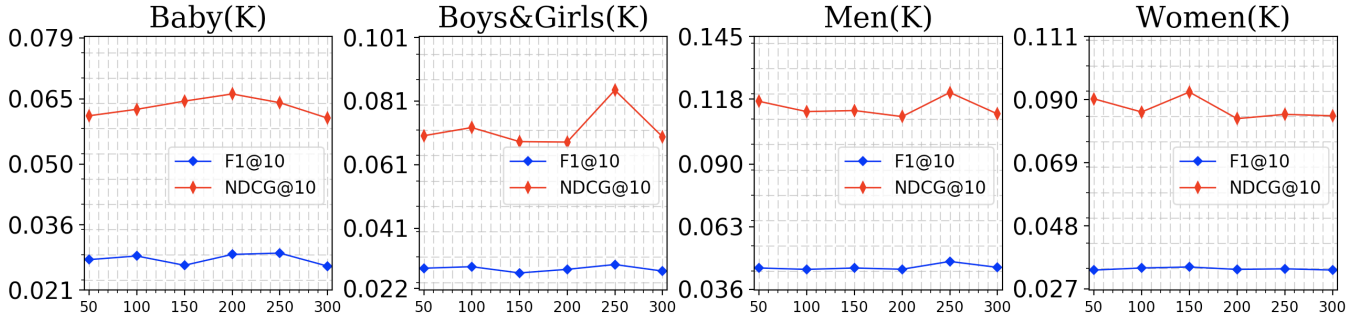
---

[4]https://www.nltk.org/
[5]http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/

Figure 3: Performance of our model with different Ks across various datasets.
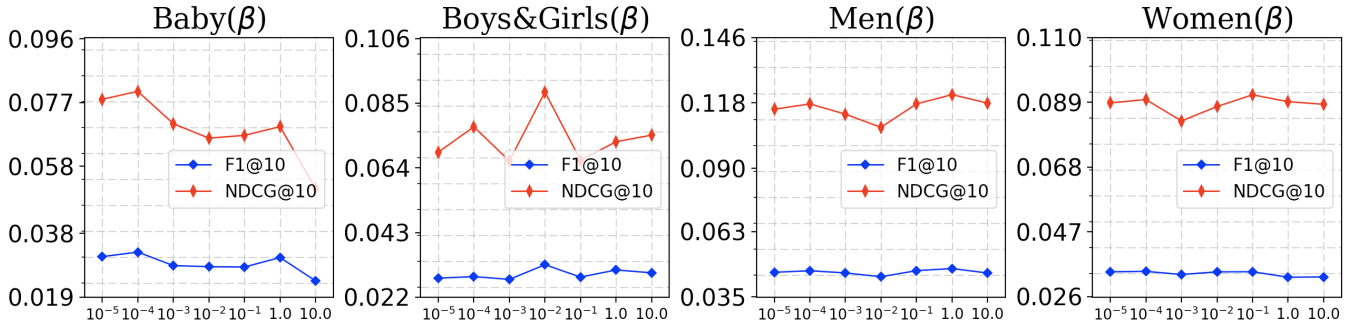

Figure 4: Performance of our model with different βs across various datasets.

we first tune K by setting β as 0.0001. Then we fix K as the optimal value, and observe the model performance by exploring different βs. Due to the space limitation, unless specified, we only report F1@10 and NDCG@10, while the results on HR@10 are similar and omitted.

• **Influence of the embedding dimension K.** The results of tuning embedding dimension K are presented in Figure 3, from which we can see: in most cases, the best performances of our model across different datasets were achieved when K is moderate, while too large embedding dimension (*e.g.*, $K = 300$) didn't help to further promote the results. This observation is consistent with many previous studies [14, 44], and the reasons can be that: the redundant parameters will increase the model complexity and may lead to the over-fitting problem, which will weaken the model generalization ability.

• **Influence of the hyper parameter β.** In our model, β determines the trade-off between the modeling of implicit feedback and user review information. The performance of our model with different βs are presented in Figure 4. We found that the optimal value of β varies across different datasets, *i.e.*, $\beta = 0.0001$ for Baby and Women, $\beta = 0.01$ for Boys&Girls, $\beta = 1.0$ for Men. It seems that β is more of a domain-dependent setting, since we were not able to find any correlation with dataset size, sparsity and etc. However, one consistent fact is that too large β (such as $\beta = 10$), which means we focus too much on the review information, didn't perform well on all the datasets. We speculate that too large β may submerge the implicit feedback signal, which is important for propagating

collaborative filtering information, and therefore, will limit the final performance.

4.4 Model Ablation Study (RQ3)

There are many components in our model, for better understanding their impacts on the final performance, in this section, we conduct ablation study by comparing our VECF model with its two variations. The first compared method is called VECF(-rev), where we remove the user review information from the outputs (*i.e.*, $\beta = 0$). The other variation is named as VECF(-att), where we drop the attention mechanism and directly average all the regional features[6]. The comparison results on F_1 and NDCG are presented in Figure 5. We found that the relative performance ranking between two variants are always interchanging across different datasets, while by incorporating review information and attention mechanism together, the final VECF model consistently performed better than both of its variations. The improvement from VECF(-att) to VECF justify our intuition in section 3.1, that is, discriminating user fine-grained visual preference may have positive effect on the performance of fashion recommendation. The superiority of VECF against VECF(-rev) highlights the effectiveness of review information, which verifies our analysis in the beginning of section 3.2. At last, the attention mechanism and review information may play different roles for the improved results, and our designed architecture

---

[6]For different datasets, K and β are set as the optimal values according to section 4.3.
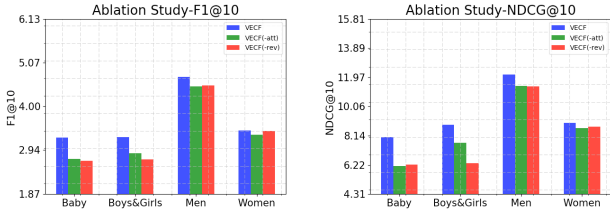
Figure 5: Model ablation study. The final VECF Model is compared with its two variations, where VECF(-rev) doesn't involve review information, while VECF(-att) directly average all the regional visual features.

provides a reasonable integration between them for more effective user preference modeling.

4.5 Evaluation of Visual Explanations (RQ4)

As mention before, once our model learned, we can provide each recommendation with visual explanations by highlighting the image regions with highest attention weights (*i.e.*, larger α_{ijk}). In this section, we evaluate whether the provided visual explanations are reasonable, i.e., whether the highlighted regions of the image can reveal a user's potential interests on the recommended item. We first conduct quantitative analysis based on a dataset with collectively labeled ground-truth. And then, to provide better intuitions for the highlighted image regions, we present and analyse several examples learned by our models in a qualitative manner.

4.5.1 Quantitative evaluation. To the best of our knowledge, this is the first work on visually explainable fashion recommendation, and there is no publicly available dataset with labeled ground-truth to evaluate whether the visual explanations (i.e., the visual highlights) generated by our model are reasonable or not. To tackle the problem, we build a collectively labeled dataset in a crowd sourcing manner. The workers are asked to identify the image regions that may explain why a user bought a particular item, based on the user's purchase records and her review written on the target item.

More specifically, we randomly select 500 user-item pairs in the testing set of Men for the workers to label. The image of the target item is equally divided into $7 \times 7 = 49$ square regions. A label task for a worker is to identify 5 out of the 49 regions that the worker believes are most relevant to the user's preference. For each label task, we provide the following two information sources to the worker for reference:

- Images and the corresponding reviews of the products that the user interacted in the training set.
- The user's review on the target item to be labeled.

In a label task, a worker is first required to read the image-review pairs of the user's interacted items. After that, the worker will be shown the target image as well as the corresponding review, and be asked to identify 5 regions of the image. In this labeling process, the worker is expected to fully understand the user's preferences according to his behaviors in the training set. Then the worker should simulate himself as the real user, and identify the most relevant image regions based on the corresponding review information.

As far as we know, few work can provide visual explanations for the recommended items. We compare VECF with its variation

Table 4: Performance comparison between VECF and VECF(-rev) on visual explanation task by identifying Top-M out of 196 candidate regions. All numbers in the table are percentage numbers with '%' omitted. Bolded numbers are used to label the best performance, and the relative improvement against the second best model is presented in the bracket.
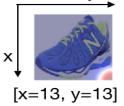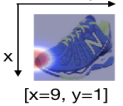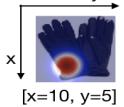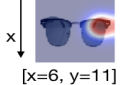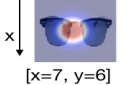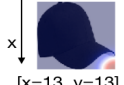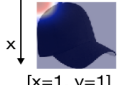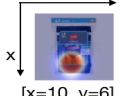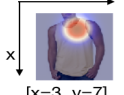
Method		Random	VECF(-rev)	VECF
F_1	M=1	0.777	1.220	**2.273** *(86.3% ↑)*
	M=2	1.430	2.012	**3.180** *(58.1% ↑)*
	M=3	1.968	2.516	**4.513** *(79.4% ↑)*
	M=4	2.281	2.857	**4.514** *(58.0% ↑)*
	M=5	2.749	3.350	**4.774** *(42.5% ↑)*
NDCG	M=1	2.975	4.348	**7.551** *(73.7% ↑)*
	M=2	2.975	4.436	**6.666** *(50.3% ↑)*
	M=3	3.458	4.254	**7.089** *(66.6% ↑)*
	M=4	2.882	4.039	**6.320** *(56.5% ↑)*
	M=5	3.501	4.284	**6.455** *(50.7% ↑)*

VECF(-rev), aiming to study whether the review information is helpful for learning better visual attention weights. We set the dimension of user/item embedding K as 250, and the weighting parameter β as 1.0, based on which we can achieve the best Top-N recommendation performance on the dataset of Men. The metrics of F_1 and NDCG are utilized as the evaluation methods. Recall that both VECF and VECF(-rev) models work on $14 \times 14 = 196$ image regional features. In our experiment, we use each model to identify M regions out of the 196 candidates according to the learned attention weights ($\alpha_{i,j,k}$), and an identified region is considered correct if it falls into the human-labeled regions. The results by comparing our predicted regions against the ground-truth are presented in Table 4. It should be noted that selecting a few regions out of 196 candidates itself poses a difficult problem as a ranking task, which is shown by the inferior performance of a randomized selection. By attentively learning the importances of different image regions based on user implicit feedback, the VECF(-rev) model obtained better performance than the random strategy. When we further introduce the supervision of review information, the final VECF model generated much more accurate visual explanations, which verifies the effectiveness of user reviews for enhancing the learning of visual attention weights.

4.5.2 Qualitative evaluation. Recommendation explainability is often assessed qualitatively [13, 29, 37, 39]. In this section, we evaluate our model in a similar manner to provide intuitive understandings on the generated visual explanations. We compare VECF with VECF(-rev), and present their provided explanations on the same product. We highlight one region for each image according to the learned attention weights (*i.e.*, $\alpha_{i,j,k}$'s), and the examples are presented in Table 5. From the results, we have the following observations:

Both VECF and VECF(-rev) can highlight some fashion elements on the product images. In Case 1, for example, the toe of the shoes were highlighted by VECF, and in Case 2, VECF(-rev) labeled the wrist of the gloves. By comparing the highlighted regions with the user review information, we can see: VECF can highlight more accurate image regions than VECF(-rev). In Case 3, for example, the

Table 5: Examples of both satisfied and failed visual explanations, where each row represents a user-item interaction. The second and third columns show the item images and the user review information. The last two columns present the highlighted image regions (as well as their coordinates) predicted by VECF and VECF(-rev), respectively. Bolded italic font indicates correspondences between the user review information and the visual explanations provided by VECF.

| | Target Item | Textual Review | Visual Explanation | |
			VECF(-rev)	VECF
1 ☺		I loved about the previous generation and ***expanded the toe box a little to improve the fit.*** great buy, highly recommended.	[x=13, y=13]	[x=9, y=1]
2 ☺		***They fit my stubby fingered hand pretty well.*** I bought the large and my hand measured 9.25&34 at the knuckles.	[x=10, y=5]	[x=3, y=1]
3 ☺		These sunglasses fit well and ***I like the design around the nose***; they sit rather than dig like most other glasses can. The included pouch is great for keeping your glasses safe and scratch free.	[x=6, y=11]	[x=7, y=6]
4 ☹		The cap, which is made of a fairly heavy fabric, makes the head feel hot when worn for several hours in a warm gym or outside on a warm day. I, therefore, tend to wear it only when it is cold outside . -bi	[x=13, y=13]	[x=1, y=1]
5 ☹		These are comfortable and are a great value. I like the waist band and they are so so so (more words) comfortable....; -)-bi	[x=10, y=6]	[x=4, y=4]
6 ☹		The fabric is amazingly soft and the fit is perfect. I own several items from next level and will continue to add to my collection with different colors and styles. Amazing company, Amazing product.	[x=3, y=7]	[x=1, y=5]

user praised the nose of the glasses by "...*I like the design around the nose*...", and VECF successfully labeled the *nose* regions as a visual explanation, while VECF(-rev) highlighted the lens of the glasses. Case 1 and 2 also imply similar superiorities of VECF against VECF(-rev) in terms of visual explanation. These observations further demonstrate that the supervision of user reviews can provide informative user preference for constraining the visual attention learning in a more reasonable manner.

In addition to many favorable results, we have also noticed some bad cases. In general, there are three types of undesirable results, and the representative examples are listed as the last three cases in Table 5: In case 4, both VECF and VECF(-rev) highlighted the background regions, which are meaningless for providing visual explanations. In practice, this problem can be alleviated by constraining the candidate area to the regions containing the target products, which can be easily accomplished based on some objective detection technologies. In case 5, some specific features (*e.g.*, the waist band) of the fashion product (*i.e.*, the pants) is discussed in the user review information. However, as the image provided in the dataset is a *packing box* of the pants, both VECF and VECF(-rev) fail to provide desirable visual explanations. In case 6, although our models highlighted some meaningful image regions, the contents

(such as fabric, fit, etc) described in the review information can hardly be reflected in the product image, so the provided visual explanation didn't agree with the user review information. This manifests that although user review information can be helpful for providing better visual explanations, it also contains many noises which may bias the attention learning process. We leave the review denoising problem as a future work, which may further improve the quality of our provided visual explanations.

5 RELATED WORK

In this section, we briefly review the recent progress in the areas of fashion recommendation and recommendation explainability, which are highly relevant with our work. By illustrating the nature of existing methods, we'd like to highlight the key differences between our work and the previous ones.

5.1 Fashion recommendation

Recent years have witnessed the increasing popular of fashion recommendation in both industry and academic communities. For effectively discovering user behavior patterns in the fashion domain, many promising recommender models have been proposed.

Generally speaking, these methods mostly based themselves on the learning of user visual preferences. For example, McAuley et al. [27] studied the relationships between different products based on their appearances in the field of e-commerce, and released a large dataset for promoting this research direction. He et al. [12] represented each product image as a fixed length vector, and infused it into the bayesian personalized ranking (BPR) framwork [30] to improve the performance of Top-N recommendation. Kang et al. [18] attempted to jointly train the image representation as well as the parameters in a recommender model, and used the learned embedding to generate fashion images, which provided inspiring insights on the relationship between different ways of leveraging content information (*i.e.*, using it as an input feature or outputting it as a target.). Lin et al. [23] incorporated generation loss for better visual understanding in the fashion recommendation domain. Han et al. [10] jointly learned a visual-semantic embedding and the compatibility relationships among fashion items in an end-to-end fashion. Song et al. [33] proposed a content-based neural scheme to model the compatibility between fashion items based on the Bayesian personalized ranking (BPR) framework. Hu et al. [15] proposed a functional tensor factorization method to model the interactions between user and fashion items.

In essence, a fashion image in the above methods is usually transformed into a fixed-length vector to combine with various personalization models. While in our model, we made a further step to discover user fine-grained visual preferences, that is, modeling user various attentions on different image regions. Our model is not only able to improve the fashion recommendation performance, but also can generate visual explanations for the recommended items.

5.2 Explainable Recommendation

Explainable recommendation essentially aims to solve the problem of "why an item is recommended to a user", which is important in a practical system due to its benefits on enhancing the recommendation perverseness as well as the users' satisfaction [8, 43, 45, 47]. Existing explainable models usually interpret a recommendation based on external knowledges, among which user textual review is a mostly adopted one. Based on this information, early methods, such as HFT [26] and RBLT [34], mainly focused on combining topic models (*e.g.*, LDA [3]) with matrix factorization (MF) for collective user review and rating modeling. The core idea of these methods is aligning each dimension in the user latent factors with a topic in the review information, and leveraging the top words of the learned topic to explain the user preference represented by the latent factors. Despite effectiveness, the "Bag-Of-Words (BOG)" assumption held by topic models is limited in capturing review semantic information, which may degrade the performance as well as the interpretability of these recommender models [48]. Fortunately, the prospering of deep (representation) learning technology shed some lights on this problem, and many recent efforts have been devoted to build deep explainable recommender models for more accurate semantic mining on the review information. According to the manners of providing explanations, these models can be largely divided into two classes. On one hand, many methods [5, 9, 31, 36, 40] provide explanations in an "extractive" manner.

The basic idea is representing a target item by merging all its review information into a document, and leveraging user embedding as a query to search the most valuable parts in the item review documents to form the final recommendation explanations. In particular, D-Attn [31] and NARRE [5] leveraged attention mechanism to automatically identify important review information under the supervision of user-item ratings. Motivated from the intuition that a user's attentions on her previous reviews should be dependent on the item she is going to buy, CARL [40] and MPCN [36] utilized "Dynamic fusion" and "Co-attention" techniques to "extract" tailored explanations for the target item. DER [9] further took user dynamic preference into consideration. On the other hand, there are also many models [21, 22, 28, 35] exploring to provide explanation in a "generative (or abstractive)" manner. Instead of using some existing review information, these methods proposed to automatically generate some neural language sentences to explain the recommendation. In specific, NRT [21], gC2S [35] and NOR [22] leveraged recurrent neural network (RNN) or its variations to generate explanations word by word. ExpansionNet [28] further incorporated product "aspects" to provide more diverse explanation sentences.

In addition to user review information, recent years have also witnessed the emergence of explainable recommendation with knowledge graph [1, 4, 16, 38, 41]. The key idea of these methods is to extend item relationship with external knowledge, and make inference along knowledge paths to understand various user behavior patterns. In specific, KSR [16] utilized Key-Value Memory Network (KV-MN) to involve KB information in the context of sequential recommendation. ECFKG [1] borrowed the idea from translation-based objective function to build a personalization model based on multiple item relationships. RippleNet [38] proposed an end-to-end knowledge-aware recommender by taking the advantages of both embedding- and path-based methods. KPRN [41] considered dependencies among different entities, and leveraged attention mechanism to discover suitable knowledge paths for recommendation explanations. KTUP [4] studied how to mutually enhance the tasks of Top-N recommendation and knowledge graph completion by learning a joint model.

Although both the above methods and our model aim to construct explainable recommender models, we explore to capture users' visual preference, and correspondingly provide explanations from a novel visual perspective.

6 CONCLUSIONS

In this paper, we propose to jointly leverage image region-level features and user review information for enhancing fashion recommendation. To this end, we build a *visually explainable collaborative filtering* model based on a multimodal attention network to seamlessly couple different feature modalities. Extensive experiments verified that our model is not only able to provide accurate recommendations, but also can provide visual explanations for the recommended items.

This paper actually made a first step towards personalized fashion recommendation with visual explanations, and there is still much room to improve it. To begin with, even though visual explanations are intuitive and vivid in the fashion domain, not all features are appropriate to be explained visually, for example, the

quality of a clothing and the warmth of a pair of shoes. In the future, we will study the relationships between textual and visual explanations (*e.g.* their complementarity or substitutability), based on which we can explain different item aspects in their best-suited modalities. Then, as mentioned before, user review information, as a weak supervise signal, usually contains much noise, which may bias the model learning process. In the next step, we will pay more attention to extract more effective review information for better user profiling and recommendation explanations. In addition, we can also extend our framework to other domains, where visual features are important factors that influence user behaviors.

7 ACKNOWLEDGMENTS

REFERENCES

[1] Qingyao Ai, Vahid Azizi, Xu Chen, and Yongfeng Zhang. 2018. Learning heterogeneous knowledge base embeddings for explainable recommendation. *Algorithms* (2018).
[2] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. 2018. Latent Cross: Making Use of Context in Recurrent Recommender Systems. In *WSDM*.
[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *JMLR* (2003).
[4] Yixin Cao, Xiang Wang, Xiangnan He, Zikun Hu, and Chua Tat-seng. 2019. Unifying Knowledge Graph Learning and Recommendation: Towards a Better Understanding of User Preference. In *WWW*.
[5] Chong Chen, Min Zhang, Yiqun Liu, and Shaoping Ma. 2018. Neural attentional rating regression with review-level explanations. In *WWW*.
[6] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. 2017. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*.
[7] Xu Chen, Zheng Qin, Yongfeng Zhang, and Tao Xu. 2016. Learning to Rank Features for Recommendation over Multiple Categories. In *SIGIR*.
[8] Xu Chen, Hongteng Xu, Yongfeng Zhang, Jiaxi Tang, Yixin Cao, Zheng Qin, and Hongyuan Zha. 2018. Sequential recommendation with user memory networks. In *WSDM*.
[9] Xu Chen, Yongfeng Zhang, and Zheng Qin. 2019. Dynamic Explainable Recommendation based on Neural Attentive Models. In *AAAI*.
[10] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. 2017. Learning fashion compatibility with bidirectional lstms. In *MM*.
[11] Ruining He and Julian McAuley. 2016. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*.
[12] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback.. In *AAAI*.
[13] Xiangnan He, Tao Chen, Min-Yen Kan, and Xiao Chen. 2015. Trirank: Review-aware explainable recommendation by modeling aspects. In *CIKM*.
[14] Xiangnan He and Tat-Seng Chua. 2017. Neural factorization machines for sparse predictive analytics. In *SIGIR*.
[15] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *MM*.
[16] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji-Rong Wen, and Edward Y Chang. 2018. Improving Sequential Recommendation with Knowledge-Enhanced Memory Networks. In *SIGIR*.
[17] Kalervo Järvelin and Jaana Kekäläinen. 2000. IR evaluation methods for retrieving highly relevant documents. In *SIGIR*.
[18] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *ICDM*.
[19] George Karypis. 2001. Evaluation of item-based top-n recommendation algorithms. In *CIKM*.
[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
[21] Piji Li, Zihao Wang, Zhaochun Ren, Lidong Bing, and Wai Lam. 2017. Neural rating regression with abstractive tips generation for recommendation. In *SIGIR*.
[22] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten De Rijke. 2019. Explainable Outfit Recommendation with Joint Outfit Matching and Comment Generation. *TKDE* (2019).
[23] Yujie Lin, Pengjie Ren, Zhumin Chen, Zhaochun Ren, Jun Ma, and Maarten de Rijke. 2019. Improving Outfit Recommendation with Co-supervision of Fashion Generation. (2019).
[24] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning User Preferences for Visual Recommendation. In *SIGIR*.
[25] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML*.
[26] Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Recsys*.
[27] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015. Image-based recommendations on styles and substitutes. In *SIGIR*.
[28] Jianmo Ni and Julian McAuley. 2018. Personalized Review Generation by Expanding Phrases and Attending on Aspect-Aware Representations. In *ACL*.
[29] Zhaochun Ren, Shangsong Liang, Piji Li, Shuaiqiang Wang, and Maarten de Rijke. 2017. Social collaborative viewpoint regression with explainable recommendations. In *WSDM*.
[30] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *UAI*.
[31] Sungyong Seo, Jing Huang, Hao Yang, and Yan Liu. 2017. Interpretable Convolutional Neural Networks with Dual Local and Global Attention for Review Rating Prediction. In *Recsys*.
[32] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *ICLR* (2015).
[33] Xuemeng Song, Fuli Feng, Jinhuan Liu, Zekun Li, Liqiang Nie, and Jun Ma. 2017. Neurostylist: Neural compatibility modeling for clothing matching. In *MM*.
[34] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. 2016. Rating-Boosted Latent Topics: Understanding Users and Items with Ratings and Reviews.. In *IJCAI*.
[35] Jian Tang, Yifan Yang, Sam Carton, Ming Zhang, and Qiaozhu Mei. 2016. Context-aware natural language generation with recurrent neural networks. *arXiv preprint arXiv:1611.09900* (2016).
[36] Yi Tay, Luu Anh Tuan, and Siu Cheung Hui. 2018. Multi-Pointer Co-Attention Networks for Recommendation. (2018).
[37] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*.
[38] Hongwei Wang, Fuzheng Zhang, Jialin Wang, Miao Zhao, Wenjie Li, Xing Xie, and Minyi Guo. 2018. RippleNet: Propagating user preferences on the knowledge graph for recommender systems. In *CIKM*.
[39] Chao-Yuan Wu, Alex Beutel, Amr Ahmed, and Alexander J Smola. 2016. Explaining reviews and ratings with paco: Poisson additive co-clustering. In *WWW*.
[40] Libing Wu, Cong Quan, Chenliang Li, Qian Wang, Bolong Zheng, Tanmoy Chakraborty, Noseong Park, Márcia R Cappelle, Les Foulds, Humberto J Longo, et al. 2017. A Context-Aware User-Item Representation Learning for Item Recommendation. *arXiv preprint arXiv:1712.02342* (2017).
[41] Canran Xu Xiangnan He Yixin Cao Xiang Wang, Dingxian Wang and Tat-Seng Chua. 2019. Explainable Reasoning over Knowledge Graph Paths for Recommendation. In *AAAI*.
[42] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. 2014. Recurrent neural network regularization. *ICML* (2014).
[43] Yongfeng Zhang. 2017. Explainable Recommendation: Theory and Applications. *arXiv preprint arXiv:1708.06409* (2017).
[44] Y Zhang, Q Ai, X Chen, and W Croft. 2017. Joint representation learning for top-n recommendation with heterogeneous information sources. *CIKM* (2017).
[45] Yongfeng Zhang and Xu Chen. 2018. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192* (2018).
[46] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. 2014. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*.
[47] Yongfeng Zhang, Yi Zhang, and Min Zhang. 2018. SIGIR 2018 Workshop on ExplainAble Recommendation and Search (EARS 2018). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*.
[48] Lei Zheng, Vahid Noroozi, and Philip S Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *WSDM*.