Explainable Fashion Recommendation:
A Semantic Attribute Region Guided Approach

Min Hou[1,2] , **Le Wu**[3] , **Enhong Chen**[1,2*] , **Zhi Li**[1,2] , **Vincent W. Zheng**[4] and **Qi Liu**[1,2]

[1]Anhui Province Key Lab. of Big Data Analysis and Application, University of S&T of China
[2]School of Data Science, University of S&T of China
[3]Hefei University of Technology
[4]WeBank
{minho, zhili03}@mail.ustc.edu.cn, lewu@hfut.edu.cn, {qiliuql, cheneh}@ustc.edu.cn,
vincentz@webank.com

Abstract

In fashion recommender systems, each product usually consists of multiple semantic attributes (e.g., sleeves, collar, etc). When making cloth decisions, people usually show preferences for different semantic attributes (e.g., the clothes with v-neck collar). Nevertheless, most previous fashion recommendation models comprehend the clothing images with a global content representation and lack detailed understanding of users' semantic preferences, which usually leads to inferior recommendation performance. To bridge this gap, we propose a novel *Semantic Attribute Explainable Recommender System* (SAERS). Specifically, we first introduce a fine-grained interpretable semantic space. We then develop a Semantic Extraction Network (SEN) and Fine-grained Preferences Attention (FPA) module to project users and items into this space, respectively. With SAERS, we are capable of not only providing cloth recommendations for users, but also explaining the reason why we recommend the cloth through intuitive visual attribute semantic highlights in a personalized manner. Extensive experiments conducted on real-world datasets clearly demonstrate the effectiveness of our approach compared with the state-of-the-art methods.

1 Introduction

The ubiquity of online fashion shopping has led to information explosion in the fashion industry. That imposes an increasing challenge for users who have to choose from a large number of available fashion products for satisfying personalized demands. Moreover, to promote profits growth, the fashion retailers have to understand the preferences of different customers and provide more intelligent recommendation services. However, unlike generic objects, clothes usually present significant variations of visual appearance, which has a vital impact on consumer decision.

Figure 1: An example of user preferences for semantic attributes.

Along this line, some studies have enhanced the performance of fashion recommendation via various visual features. For example, ImageNet[1] pre-trained convolutional neural networks (CNNs) was adopted to extract visual representation [McAuley *et al.*, 2015; He and McAuley, 2016b; He *et al.*, 2016b]. As these visual features were trained for the image classification task, the pre-trained CNNs could only capture category information. Yu *et al.* [2018b] and Liu *et al.* [2017] attempted to get a better understanding of clothing by utilizing aesthetics and style features, respectively. These methods have hammered at comprehending clothing from a holistic perspective via global categories, aesthetics and style information, while failing to dive deep into the details to explore clothing fine-grained semantic aspects.

However, when purchasing clothing products, it is intuitive that we often have preferences for detailed semantic attributes (such as neckline, heel height, skirt length) in addition to global impressions. For instance, Figure 1 shows two users' purchase histories. The first user chooses three products in different categories with the same kind of neckline. Therefore, we can infer that she prefers V-neckline clothes. Meanwhile, the second user prefers the pointed toe incorporated with high-heeled shoes as evidenced by the frequent appearance of these two semantic attributes in her chosen products. In fact, clothing is not atomic and a piece of clothing usually consists of multiple attributes, such as neckline, sleeves length. In some cases, people choose a product because it contains the attributes they like. Therefore, the clothing semantic attribute can not only help us generate a comprehen-

---

*Corresponding Author

[1]http://image-net.org/

sive representation of items, but also make a deep understanding of user preferences. Unfortunately, there are still many unique challenges inherent in designing an effective solution to integrate semantic attribute information into fashion recommendation. On the one hand, it's difficult to obtain clothing semantic attributes features without the manual attribute annotation in the large-scale E-commerce data. On the other hand, the user preferences are sophisticated, while traditional methods usually transform the item image into a latent vector directly. These two aspects make it hard to generate explainable recommendations with current recommendation models.

To address the challenges mentioned above, in this paper, we propose a novel *Semantic Attribute Explainable Recommender System* (SAERS) for fashion recommendation. In SAERS, we introduce a fine-grained interpretable space named semantic attribute space, where each dimension corresponds to a semantic attribute. We project users and items into this space to capture the user's fine-grained preferences and generate explainable recommendations. Specifically, we firstly develop a Semantic Extraction Network (SEN), which is used to extract the region-specific attribute representations in a weakly-supervised manner. Through SEN, each item is projected to the semantic attribute space. Then, to capture the diversity of the semantic attribute preference among items, we design a Fine-grained Preferences Attention (FPA) module to automatically match the user preference for each attribute in semantic attribute space, and aggregate all attributes with different weights. Finally, we optimize the SAERS in Bayesian Personalized Rank (BPR) framework. Extensive experiments on the large-scale real-world Amazon dataset reveal that SAERS not only significantly outperforms several baselines on the visual recommendation task, but also provides interpretable insights by highlighting attribute semantic in a personalized manner.

2 Related Work

Generally, the related work can be grouped into the following three categories, i.e., visual-aware recommendation, attribute localization and explainable recommendation.

Visual-aware Recommendation. Visual-aware recommendation uses visual signals presented in the underlying data to model visual characteristics of items and user preferences. Some previous works [McAuley *et al.*, 2015; He and McAuley, 2016b; He and McAuley, 2016a; He *et al.*, 2016b; Le *et al.*, 2019] directly utilized the ImageNet pre-trained CNN to generate items' visual representation. To further explore the information contained in the product image, Liu *et al.* [2017] extracted style features to represent item via eliminating the corresponding categorical information from the visual feature vector generated by CNN. Yu *et al.* [2018b] leveraged an aesthetic network to obtain image aesthetic elements for fashion recommendation task. In [Kang *et al.*, 2017; Lei *et al.*, 2016], Siamese framework was utilized to obtain task-specific visual representation. This framework allowed CNN to be trained comparatively with two 'copies' and joined by a triplet loss function. Unfortunately, these methods focused on products' global characteristics and
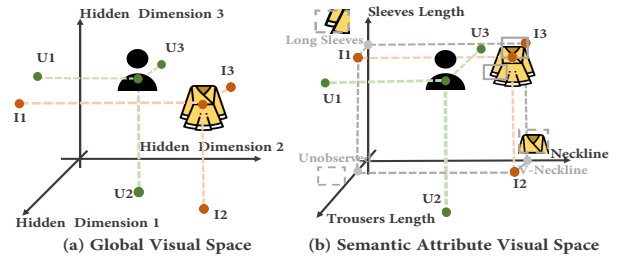

Figure 2: Difference between the conventional (a) Global Visual Space and our (b) Semantic Attribute Visual Space.

transferred item image into a fixed latent vector. However, the fine-grained attribute has been largely under-exploited.

Attribute Localization. In computer vision, the semantic attribute is a type of important auxiliary information and has been leveraged as a fine-grained representation for image understanding [Chen *et al.*, 2012; Kovashka *et al.*, 2012]. Researchers have demonstrated that utilizing localized representation to model attributes is more effective [Xiao and Jae Lee, 2015]. Singh and Lee [2016] presented a ranking framework that combines a spatial transformer network to localize attributes. Ak *et al.* [2018] proposed an efficient weakly-supervised attribute localization and representation method through Class Activation Maps (CAM) [Zhou *et al.*, 2016], which could highlight the most informative image regions relevant to the predicted class. Nevertheless, in CAM, the fully connected layer of CNN needs to be replaced to global average pooling (GAP) layer and re-trained. For the sake of convenience, in this paper, we use Grid-CAM [Selvaraju *et al.*, 2017] instead to extract attributes' relevant regions.

Explainable Recommendation. Although conventional recommender systems [Yu *et al.*, 2018a; Zhu *et al.*, 2018; Liu *et al.*, 2011; Li *et al.*, 2018] have made great strides, their recommendations are often unexplained and therefore difficult to convince users. To solve this urgent issue, explainable recommendation [Zhang and Chen, 2018; Abdollahi, 2017] has become a very important research direction in recent years. Bilgic *et al.* [2005] adopted statistical neighbors' ratings and the most impacted bought items to help users understand the recommendations. McAuley *et al.* [2013] and Zhang *et al.* [2014] improved the interpretability of collaborative filtering models by utilizing explicit factor based matrix factorization (MF) as explainable components. They aligned each latent dimension in MF with a particular explicit product aspects/topics. In deep learning based methods, researchers discovered that the attention mechanism was capable of improving model explainability by automatically learning the importance of explicit features and also refined user/item embedding [Chen *et al.*, 2018; Gao *et al.*, 2019; Chen *et al.*, 2017]. Chen *et al.* [2018] proposed an attentive architecture with the supervision of user implicit feedback as well as textual reviews to capture users' visual preferences. These works attempted to make intuitional explanations for the recommendations but were limited on the item level. In this paper, we take a further step to explain the user preferences on the visual attribute level.
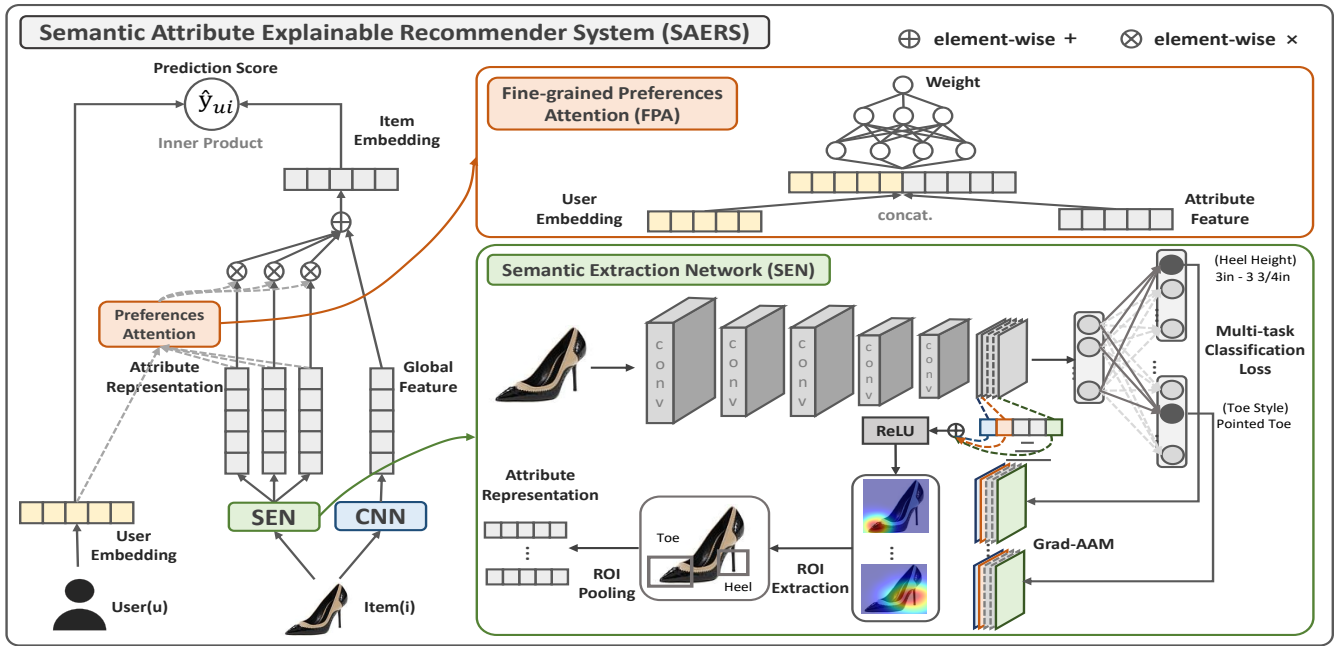
Figure 3: The architecture for Semantic Attribute Explainable Recommender System (SAERS)

3 SAERS: Semantic Attribute Explainable Recommender System

In this section, we introduce our proposed SAERS for addressing the personalized fashion recommendation problem.

Most existing works represent users and items in a global visual space (as shown in Figure 2(a)), but the meaning of each dimension is unknown, reducing the interpretability of recommendations. To bridge this gap, SAERS utilizes a new semantic attribute visual space. In this space, each dimension represents an attribute, which corresponds to different regions of the clothing. As shown in Figure 2(b), items are split into several semantic attributes via Semantic Extraction Network, then projected into the semantic attribute visual space. After that, users are projected according to fine-grained preferences for clothing attributes. As shown in Figure 3, SAERS contains two main components, i.e., Semantic Extraction Network (SEN) and Fine-grained Preferences Attention (FPA). Specifically, with SEN, we first obtain the fashion item projections in the semantic feature space. Next, we design FPA to project users into the same semantic feature space. Then, we jointly learn the item representation in both global visual space and semantic attribute visual space under a pair-wise learning framework. Finally, with attribute preference inference, we can generate the explainable recommendations.

3.1 Projecting Item into Semantic Attribute Space

In this subsection, we investigate how to project items into semantic attribute space. First, we need to divide the item into several parts (semantic attributes), where each part corresponds to one dimension in the space. And then, we extract attribute representations for them. Whereas, most real-world E-commerce datasets lack attribute annotations to learn se-

mantic attribute representation directly. To this end, we borrow a fine-grained labeled data and pre-train a Semantic Extraction Network (SEN), which is used to extract the region-specific attribute representations and simultaneously locate and classify attributes.

Due to the lack of training data with bounding box annotations, we cannot leverage traditional object detection methods [Girshick, 2015] to locate and extract attribute representations straightforwardly. Therefore, we utilize Gradient-weighted Attribute Activation Maps (Grad-AAM) [Selvaraju et al., 2017] to develop SEN in a weakly-supervised manner. Grad-AAM is capable of producing a coarse location highlighting according to the gradient information of target attribute in CNN. In SEN, we first train a classification CNN. Then, according to the results of CNN, we calculate the Grad-AAM to get attributes location. Finally, the region-specific attribute representations can be obtained by ROI pooling layer.

We combine the *UT-Zap50K shoes*[2] dataset and the *Tianchi Apparel*[3] dataset, which contains 50,025 shoe and over 180,000 apparel image-level attribute annotations respectively. As listed in Table 1, each row represents an attribute. We employ 12 attributes that cover the various locations of the clothing, with each attribute is classified into several classes as shown after the colon. Then, the combined dataset is used to train a multi-task classification network. The network architecture can be arbitrary. In this work, ResNet-50 [He et al., 2016a] is utilized. We use the following classifica-

---

[2]http://vision.cs.utexas.edu/projects/finegrained/utzap50k/
[3]https://tianchi.aliyun.com/competition/entrance/231671/ information

tion loss to train the network:

$$L_C = -\sum_{I=1}^{N}\sum_{a=1}^{A} \log\left(p\left(\hat{y}_{Ia}|y_{Ia}\right)\right),\qquad(1)$$

where $\hat{y}_{Ia}$ represents the ground truth class of the ath attribute of the Ith image. y_{Ia} is the corresponding predicted result. N is the number of training examples and A is the number of attributes. The posterior probability estimates the probability of y_{Ia} to be classified as $\hat{y}_{Ia}$.

The classification loss is used to discover the most relevant regions of attributes. After the classification network has converged, we start to calculate the Grad-AAM $M_{\text{Grad-AAM}}^{a_c}$ for class c of an attribute a, where c is determined by the class that maximizes the classification confidence. We first compute the gradient of the score for class a_c, y^{a_c} (before the softmax), with respect to feature maps F^t of the last convolutional layer, *i.e.,* $\frac{\partial y^{a_c}}{\partial F^t}$. These gradients flowing back are global-averaged-pooled to obtain the neuron importance weights $\alpha_t^{a_c}$:

$$\alpha_t^{a_c} = \overbrace{\frac{1}{Z}\sum_m\sum_n}^{\text{global average pooling}} \underbrace{\frac{\partial y^{a_c}}{\partial F_{mn}^t}}_{\text{gradients via backpop}}.\qquad(2)$$

Here, the (m,n) indicates the spatial location of the tth channel feature map F^t in the last convolutional layer. This weight $\alpha_t^{a_c}$ represents a *partial linearization* of the deep network downstream from F, and captures the 'importance' of feature map t for the target class a_c.

We perform a weighted combination of forwarding activation maps, and follow it by a ReLU to obtain,

$$M_{\text{Grad-AAM}}^{a_c} = ReLU\underbrace{\left(\sum_t \alpha_t^{a_c} F^t\right)}_{\text{linear combination}}.\qquad(3)$$

Actually, each element of $M_{\text{Grad-AAM}}^{a_c}$ indicates attribute class a_c's contribution of the activation. Meanwhile, these elements can be associated with the location of the input image. Consequently, by using $M_{\text{Grad-AAM}}^{a_c}$, attribute location can be added to the network. In order to do so, all attribute locations are estimated with a simple hard threshold technique. As the implementation in [Ak *et al.*, 2018], the pixel values that are above 20% of the maximum value in the generated map are segmented. This is followed by estimating a bounding box, that covers the largest connected region in the Grad-AAM. This step is repeated for each attribute. With attribute's bounding box and the last convolutional layer's feature map as input, then the ROI pooling layer [Girshick, 2015] is used to generate region-specific attribute representation.

Accordingly, given a clothing image, we can obtain 12 attribute representation through the pre-trained SEN as introduced above. We use f_a^k to denote the kth attribute feature generated by SEN.

3.2 Projecting User into Semantic Attribute Space

In this subsection, we describe how to project users into the semantic attribute space and capture fine-grained user preference on attributes.

Category	Attribute: Class
Top	*high neck*: ruffle semi-high, turtle,... *collar*: rib collar, puritan collar,... *lapel*: notched, shawl, collarless,... *neckline*: V, square, round,... *sleeves length*: sleeveless, cap, short,... *body length*: high waist, long, regular,...
Bottom	*skirt length*: short, knee, midi, ankle,... *trousers length*: short, mid, 3/4, cropped,...
Shoes	*heel height*: flat, 1 in-7/4 in, under 1 in,... *boots height*: ankle, knee-high, mid-calf,... *closure*: lace-up, slip-on, zipper,... *toe style*: round, pointed, peep, open,...

Table 1: List of semantic attributes used in our method

To get the final embedding of item i with respect to all the attribute features, our first intuition is to equally fuse all the features with an average pooling as follows:

$$f(i) = \frac{1}{A}\sum_{k=1}^{A} E^k f_a^k(i),\qquad(4)$$

where $f_a^k(i) \in \mathbb{R}^m$ denotes the kth attribute feature of item i, $E^k \in \mathbb{R}^{d\times m}$ is a matrix for transferring the kth attribute feature to lower dimension d. However, as discussed in the introduction, there may be a variety of clothing attributes that users are interested in. Different attributes may have different impacts on the candidate item i when considering whether user u will buy i or not. Therefore, we argue that the item embedding should be fused with adaptive weights to capture users' dynamic preference over attributes and enhance the interpretability of recommendations. Inspired by the recent success of attention mechanism [Chen *et al.*, 2018], which allows different parts to contribute differently when compressing them to a single representation, we propose an attribute-level attention module, namely *Fine-grained Preferences Attention* (FPA). Specifically, as illustrated in the right upper part of Figure 3, we use a DNN $\mathcal{D}$ (in this work is a two-layer multilayer perceptron) to compute attention weight α_{ui}^k with user embedding $f(u) \in \mathbb{R}^d$ and item's k^{th} transferred attribute feature $E^k f_a^k(i)$'s concatenated vector as input:

$$\begin{aligned}\alpha_{ui}^k &= \text{softmax}\left(\mathcal{D}\left(f(u), E^k f_a^k(i)\right)\right)\\ &= \frac{\exp\left(\mathcal{D}\left(f(u), E^k f_a^k(i)\right)\right)}{\sum_{k=1}^{A}\exp\left(\mathcal{D}\left(f(u), E^k f_a^k(i)\right)\right)}.\end{aligned}\qquad(5)$$

The embedding of item i can thus be calculated as the weighted sum of its attribute features:

$$f(i) = \sum_{k=1}^{A}\alpha_{ui}^k E^k f_a^k(i).\qquad(6)$$

Through this way, users are aligned in the semantic attribute space, and the preference towards each attribute can be captured.

Besides the semantic attributes, clothes usually have global characteristics information such as style, categories and aesthetics. Consequently, the embedding of item i can be calculated as the weighted sum of its attribute features plus the

global characteristics feature:

$$f(i) = \sum_{k=1}^{A} \alpha_{ui}^k E^k f_a^k(i) + f_g(i), \qquad (7)$$

where $f_g(i) \in \mathbb{R}^d$ denotes the global characteristics feature. In this work, we leverage a Siamese CNN architecture based on the AlexNet [Krizhevsky *et al.*, 2012] to obtain global characteristics feature.

3.3 Model Learning Stage

When making predictions, we feed the user embedding $f(u) \in \mathbb{R}^d$ and the final item embedding $f(i) \in \mathbb{R}^d$ into a function:

$$\hat{y}_{ui} = \mathcal{P}\left(f(u), f(i)\right), \qquad (8)$$

where $\mathcal{P}$ is an arbitrary prediction function, such as the inner product between the corresponding embeddings, or a prediction neural network [He *et al.*, 2017]. Here, we choose the inner product $\hat{y}_{ui} = f(u) \cdot f(i)$ as a specific implementation, as it gives us better training efficiency and avoids overfitting on our large-scale data. Since we address the fashion recommendation task from the ranking perspective, we employ a pairwise learning method to optimize model parameters. The assumption of pairwise learning is that the model could predict a higher score for an observed interaction that its unobserved counterparts. We adopt the Bayesian Personalized Ranking (BPR) [Rendle *et al.*, 2009] loss, which has been widely used in recommender systems:

$$L = \sum_{(u,i,j) \in \mathcal{D}_S} -\ln \sigma\left(\hat{y}_{ui} - \hat{y}_{uj}\right) + \lambda\|\Theta\|^2, \qquad (9)$$

where σ is the sigmoid function and λ is the regularization hyper-parameter. Θ includes all model parameters. $\mathcal{D}_S$ is the training set, which consists of triples in the form (u, i, j). u denotes the user together with an interacted item i and a non-observed item j. In each iteration, we sample a user u, a positive item $i \in \mathcal{I}_u^+$, and a negative item $j \in \mathcal{I}\backslash\mathcal{I}_u^+$. Thus, the convolutional neural network Φ, which is used to extract item's global characteristic feature, has two images: $\mathbf{X}_i$ and $\mathbf{X}_j$ to consider. Both CNNs $\Phi\left(\mathbf{X}_i\right)$ and $\Phi\left(\mathbf{X}_j\right)$ share the same weights. To optimize the above objective function, we employ the widely used Adam [Kingma and Ba, 2014] optimization algorithm.

3.4 Attribute Preference Inference

We project users and items into a new interpretable space. In this space, the user's preferences for attributes can be calculated, making it possible to generate personalized explainable recommendations. Specifically, when we recommend an item to a user, the user's preference towards each attribute of the item can be obtained according to the attention weight (Equation 5). The attribute with a large attention weight indicates the user's preference. Utilizing the SEN's location (Equation 3) and classification (Equation 1) ability, we can further go back to find out where the attribute is located and what the specific class is.

Accordingly, the explanation consists of three parts: (1) SAERS uses a bounding box to highlight which part of the

product image the user might like. (2) SAERS provides which semantic attribute the highlighted part belongs to. (3) SAERS provides the possibility that the user likes the semantic attribute. We will demonstrate the explainable recommendation results in the experiment section.

4 Experiments

In this section, we conduct experiments on a real-world dataset to verify the feasibility of our proposed framework. We first introduce the experimental setup, followed by the experiment results. We will also discuss the effectiveness of semantic attribute features and give a case study of SAERS's visual explainablility.

4.1 Experimental Setup

Dataset. We evaluate our methods on a real-world e-commerce dataset, i.e., Amazon Fashion. This dataset was introduced in [He and McAuley, 2016b; Kang *et al.*, 2017] and consist of reviews of clothing items crawled from *Amazon.com*. Amazon Fashion contains six representative fashion categories (men/women's tops, bottoms and shoes). We treat users' reviews as implicit feedback. There are 45,184 users, 166,270 items, and 358,003 records in total. The sparsity of the dataset is 99.9952%. In Amazon Fashion dataset, we discard inactive users with purchase records less than 5, similar as previous works [He and McAuley, 2016b; Kang *et al.*, 2017]. For each user, we randomly select one record for validation and another one for test.

Evaluation Protocol. We evaluate our method in two experiment scenarios: (1) In pair-wise recommendation, we use the *Area Under the Roc Curve* (AUC) [Rendle *et al.*, 2009] to measure the probability that a model will rank a randomly chosen positive instance higher than a randomly chosen negative one. (2) In list-wise top-*N* recommendation, the *Normalized Discounted Cumulative Gain* (NDCG@*N*) are calculated to evaluate the performance of the baselines and our model. In order to reduce the computational cost, we randomly select a user, and then randomly sample 500 unrated items at each time and combine them with the positive item the user likes in the ranking process. We repeated this procedure 10000 times and report the average ranking results [Wu *et al.*, 2017].

Implementation Details. For all the models (except Random and PopRank), we tune hyper-parameters via grid search on the validation set, with a regularizer selected from [0, 0.0001, 0.001, 0.01, 0.1, 1], learning rate selected from [0.0001, 0.001, 0.01], and the latent feature dimension of [10, 30, 50, 100]. We use mini-batch size of 256 to train all the models until they converge. All experiments are trained with NVIDIA K80 graphics card and implemented by *Tensorflow*.

4.2 Comparison Methods

To demonstrate the effectiveness of SAERS, we compare it with the following alternative methods:

- **Random:** This baseline ranks items randomly for all users. By definition, this method has AUC=0.5.
- **PopRank:** This is a nontrivial method that items are ranked in order of their popularity within the system.
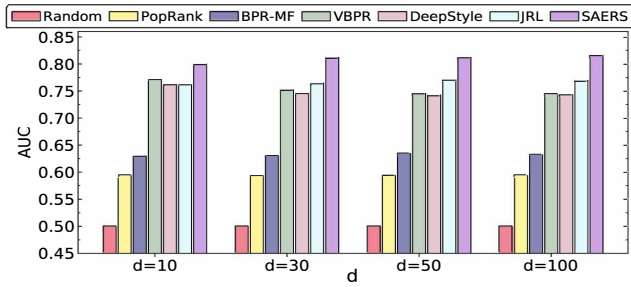
Figure 4: Overall performance of difference choices of embedding dimension d in *All Items* scenario.
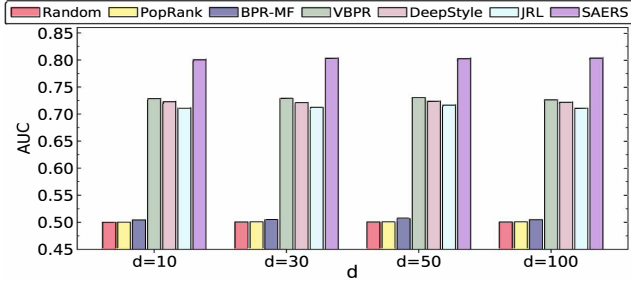

Figure 5: Overall performance of difference choices of embedding dimension d in *Cold Items* scenario.

- **BPR-MF** [Rendle *et al.*, 2009]: This is a classical matrix factorization method based on pairwise personalized ranking for implicit feedback datasets.
- **VBPR** [He and McAuley, 2016b]: One of the state-of-the-art methods for visually-aware personalized ranking from implicit feedback. VBPR is a form of hybrid content-aware recommendation that makes use of pre-trained CNN features of product images.
- **DeepStyle** [Liu *et al.*, 2017]: A state-of-the-art method for visual recommendation, based on the learned style features and the BPR framework. The style features are obtained by subtracts categories information from visual features of items generated by CNN.
- **JRL** [Zhang *et al.*, 2017]: It is a state-of-the-art neural recommender system, which can leverage multimodal side information for Top-N recommendation. We only use image visual information in this baseline.

4.3 Recommendation Performance

In this section, we evaluate our model for fashion recommendation. Figure 4 and Figure 5 present the recommendation performance on AUC in *All Items* and *Cold Items* scenarios, respectively. For the latter, we seek to estimate relative preference scores among items that have never been observed at training time. Figure 6 depicts the Top-*N* recommendation performance on two scenarios when embedding dimension d=100. Figure 7 shows the AUC during training. The main findings are summarized as follows:

- We can find that our proposed SAERS achieves the best performance on AUC, with the improvement by up to 5.8% compared to the second-best method across all datasets, and 10.2% in cold-start scenarios. In Figure 6, we can also
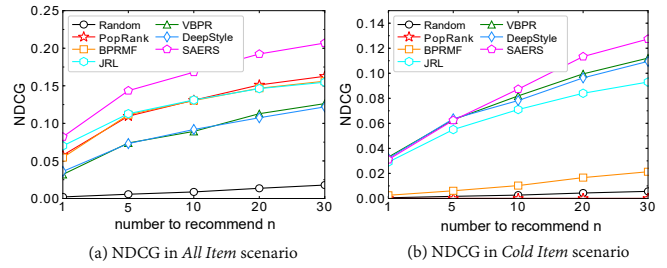

Figure 6: NDCG@*N* in top-*N* recommendation for *All Items* and *Cold Items* scenarios.
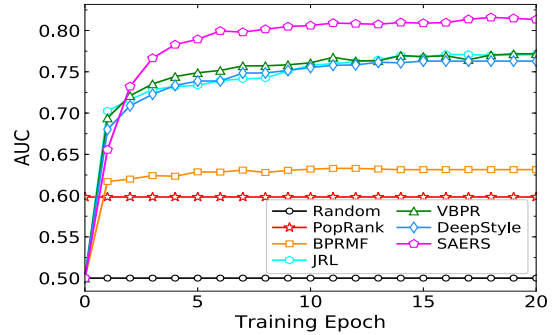

Figure 7: Performance of training epochs (test set)

find that our model outperforms the other baseline methods in Top-*N* recommendation at two scenarios. The result demonstrates the significant benefits of incorporating semantic attribute features in fashion recommendation.

- Our method surpasses the strongest collaborative filtering based method (BPR-MF) significantly, especially in *Cold Items* scenarios. This phenomenon is largely due to the fact that the dataset is particularly sparse. The sparsity of the dataset is 99.9952% and on average each item is only be purchased 2.15 times. This further illustrates the importance of using content-based recommendation methods and in-depth understanding of image information in fashion recommendation tasks.
- The performance of DeepStyle using style information does not work better than VBPR, by directly using CNN pre-trained features. DeepStyle assumes that style feature can be obtained by subtracting categorical information from visual features of items generated by CNN. However, in fashion recommendation tasks, there are few categories of clothing, but the style is indeed diverse. Therefore, the assumption could not work well. This reflects the difficulty of effectively incorporating content information into the recommender system.
- Figure 4 and 5 show the sensitivity of the models to the embedding dimension d. It shows that SAERS is robust to hyper-parameter tuning.

4.4 Necessity of the Semantic Attribute

In this subsection, we discuss the necessity of using the semantic attribute features extracted by SEN. We construct three variant implements based on VBPR [He and McAuley, 2016b] and SAERS to verify the effectiveness of incorporat-

Models	AUC	Improvement
SAFo	0.7569	-
VBPR	0.7718	-
VBPR + SAF	0.7826	1.4%
SAERS - SAF	0.8017	-
SAERS	**0.8161**	**1.8%**

Table 2: Improvement of using semantic attribute feature.

ing the semantic attribute information.

- **SAFo:** This is a model with **S**emantic **A**ttribute **F**eatures **o**nly. Items are represented as the fusion of all semantic attributes representations, as shown in Equation 4.
- **VBPR+SAF:** Add the **S**emantic **A**ttribute **F**eatures to the CNN features VBPR used as the item representation.
- **SAERS-SAF:** A variant of our model that uses only global characteristic features extract by Siamese CNN to represent items.

As shown in Table 2, the usage of SAF can increase the effect of VBPR and our model by 1.4% and 1.8%, respectively. Experimental results show our SAERS model, which captures both local attribute semantic information and global characteristic information, performs the best result on the *Amazon* dataset since these two kinds of information mutually enhance each other to a certain extent.

4.5 User Attribute Interest Visualization

In order to better demonstrate the interpretability of our model, we randomly selected four users. As shown in Figure 8, for each user, we present 3 logs from the purchase history as well as our recommendation result. The value in the red box indicates the attention weight. The greater weight demonstrates the higher personal preferences for this attribute. We show the user's favorite attribute calculated by the Fine-grained Preferences Attention (FPA) and locate it via Grad-AAM. (1) Attribute class, (2) attribute location and (3) attention weight are corresponding to three visual explanations we mentioned in Section 3.4. For example, SAERS recommends a T-shirt to user A and tells him that the reason for the recommendation is short sleeves and the position corresponding to the sleeve is highlighted. This is reasonable because the user has purchased three short sleeves clothing. For user C, SAERS recommends a shirt with V-neckline because it accurately captures the user's preference on this attribute. In the recommendation results, we can see that the neckline of this shirt has the largest attention weight, which accounts for 12.76%. This explains the reason we recommend this shirt: user C likes this kind of neckline.

5 Conclusions

In this paper, we proposed SAERS, aiming at comprehending user preference in fine-grained attribute level and providing visual interpretability in fashion recommendation tasks. To achieve this goal, we first introduced a fine-grained interpretable space, namely, semantic attribute space. Then we developed a Semantic Extraction Network (SEN) and Fine-grained Preferences Attention (FPA) module to project

Figure 8: Examples of the visual explanations (The IDs of User A, B, C and D are A2ZEFKXUBN2IIS, A2OLJU8M0L4YZD, AIUZT213W0WDE and AQ6BC76U77C7X)

users and items into this space, respectively. Finally, we jointly learned the item representation in both global visual space and semantic attribute space under a pair-wise learning framework. The experimental results on a large-scale real-world dataset clearly demonstrated the effectiveness and explanatory power of SAERS.

Acknowledgements

References

[Abdollahi, 2017] Behnoush Abdollahi. Accurate and justifiable: new algorithms for explainable recommendations. *ETD*, 2017.

[Ak *et al.*, 2018] Kenan E Ak, Ashraf A Kassim, Joo Hwee Lim, and Jo Yew Tham. Learning attribute representations with localization for flexible fashion search. In *CVPR*, pages 7708–7717, 2018.

[Bilgic and Mooney, 2005] Mustafa Bilgic and Raymond J Mooney. Explaining recommendations: Satisfaction vs. promotion. In *IUI*, volume 5, page 153, 2005.

[Chen *et al.*, 2012] Huizhong Chen, Andrew Gallagher, and Bernd Girod. Describing clothing by semantic attributes. In *ECCV*, pages 609–623. Springer, 2012.

[Chen *et al.*, 2017] Jingyuan Chen, Hanwang Zhang, Xiangnan He, Liqiang Nie, Wei Liu, and Tat-Seng Chua. Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In *SIGIR*, pages 335–344. ACM, 2017.

[Chen *et al.*, 2018] Xu Chen, Yongfeng Zhang, Hongteng Xu, Yixin Cao, Zheng Qin, and Hongyuan Zha. Visually explainable recommendation. *arXiv preprint arXiv:1801.10288*, 2018.

[Gao *et al.*, 2019] Jingyue Gao, Xiting Wang, Yasha Wang, and Xing Xie. Explainable recommendation through attentive multi-view learning. In *AAAI*, 2019.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *ICCV*, 2015.

[He and McAuley, 2016a] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, pages 507–517, 2016.

[He and McAuley, 2016b] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[He *et al.*, 2016a] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[He *et al.*, 2016b] Ruining He, Chunbin Lin, Jianguo Wang, and Julian McAuley. Sherlock: sparse hierarchical embeddings for visually-aware one-class collaborative filtering. In *IJCAI*, pages 3740–3746, 2016.

[He *et al.*, 2017] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *WWW*, pages 173–182, 2017.

[Kang *et al.*, 2017] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. Visually-aware fashion recommendation and design with generative image models. In *ICDM*, pages 207–216. IEEE, 2017.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Kovashka *et al.*, 2012] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, pages 2973–2980, 2012.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[Le *et al.*, 2019] Wu Le, Chen Lei, Hong Richang, Fu Yanjie, Xie Xing, and Wang Meng. A hierarchical attention model for social contextual image recommendation. *TKDE*, 2019.

[Lei *et al.*, 2016] Chenyi Lei, Dong Liu, Weiping Li, Zheng-Jun Zha, and Houqiang Li. Comparative deep learning of hybrid representations for image recommendations. In *CVPR*, pages 2545–2553, 2016.

[Li *et al.*, 2018] Zhi Li, Hongke Zhao, Qi Liu, Zhenya Huang, Tao Mei, and Enhong Chen. Learning from history and present: Next-item recommendation via discriminatively exploiting user behaviors. In *SIGKDD*, pages 1734–1743. ACM, 2018.

[Liu *et al.*, 2011] Qi Liu, Yong Ge, Zhongmou Li, Enhong Chen, and Hui Xiong. Personalized travel package recommendation. In *ICDM*, pages 407–416. IEEE, 2011.

[Liu *et al.*, 2017] Qiang Liu, Shu Wu, and Liang Wang. Deepstyle: Learning user preferences for visual recommendation. In *SIGIR*, pages 841–844. ACM, 2017.

[McAuley and Leskovec, 2013] Julian McAuley and Jure Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *RecSys*, pages 165–172. ACM, 2013.

[McAuley *et al.*, 2015] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52. ACM, 2015.

[Rendle *et al.*, 2009] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461. AUAI Press, 2009.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE, 2017.

[Singh and Lee, 2016] Krishna Kumar Singh and Yong Jae Lee. End-to-end localization and ranking for relative attributes. In *ECCV*, pages 753–769, 2016.

[Wu *et al.*, 2017] Le Wu, Yong Ge, Qi Liu, Enhong Chen, Richang Hong, Junping Du, and Meng Wang. Modeling the evolution of users' preferences and social links in social networking services. *TKDE*, 29(6):1240–1253, 2017.

[Xiao and Jae Lee, 2015] Fanyi Xiao and Yong Jae Lee. Discovering the spatial extent of relative attributes. In *ICCV*, pages 1458–1466, 2015.

[Yu *et al.*, 2018a] Runlong Yu, Yunzhou Zhang, Yuyang Ye, Le Wu, Chao Wang, Qi Liu, and Enhong Chen. Multiple pairwise ranking with implicit feedback. In *CIKM*, pages 1727–1730. ACM, 2018.

[Yu *et al.*, 2018b] Wenhui Yu, Huidi Zhang, Xiangnan He, Xu Chen, Li Xiong, and Zheng Qin. Aesthetic-based clothing recommendation. In *WWW*, 2018.

[Zhang and Chen, 2018] Yongfeng Zhang and Xu Chen. Explainable recommendation: A survey and new perspectives. *arXiv preprint arXiv:1804.11192*, 2018.

[Zhang *et al.*, 2014] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. Explicit factor models for explainable recommendation based on phrase-level sentiment analysis. In *SIGIR*, pages 83–92. ACM, 2014.

[Zhang *et al.*, 2017] Yongfeng Zhang, Qingyao Ai, Xu Chen, and W Bruce Croft. Joint representation learning for top-n recommendation with heterogeneous information sources. In *CIKM*, pages 1449–1458, 2017.

[Zhou *et al.*, 2016] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[Zhu *et al.*, 2018] Feng Zhu, Yan Wang, Chaochao Chen, Guanfeng Liu, Mehmet A Orgun, and Jia Wu. A deep framework for cross-domain and cross-system recommendations. In *IJCAI*, pages 3711–3717, 2018.