NeuroStylist: Neural Compatibility Modeling for Clothing Matching

Xuemeng Song†, Fuli Feng§, Jinhuan Liu†, Zekun Li†, Liqiang Nie†, Jun Ma†

† Shandong University, Jinan, China, § National University of Singapore, Singapore

{sxmustc, fulifeng93, liujinhuan.sdu, lizekunlee, nieliqiang}@gmail.com, majun@sdu.edu.cn

ABSTRACT
In modern society, clothing has gradually become an important beauty-enhancing product, playing an important role in human’s social life. In fact, the key to a proper outfit usually lies in the harmonious clothing matching. Nevertheless, not everyone is good at clothing matching. Fortunately, with the proliferation of fashion-oriented online communities, fashion experts can share their fashion tips to the public by showcasing their outfit compositions, where each fashion item (e.g., a top, a bottom) usually has an image and context metadata (e.g., title and category). Such rich fashion data offer us the new opportunity to investigate the code in clothing matching. However, challenges co-exist with opportunities. The first challenge lies in the complicated factors, such as color, material and shape, that affect the compatibility of complementary fashion items. Second, as each fashion item involves multiple modalities (i.e., image and text), how to cope with the heterogeneous multimodal data also poses a great challenge. Third, our pilot study shows that the relationship between fashion items is rather sparse, which makes the traditional matrix factorization methods not applicable. Towards this end, in this work, we propose a content-based neural scheme to model the compatibility between fashion items based on the Bayesian Personalized Ranking (BPR) framework, which is able to jointly model the coherent relationship between different modalities of fashion items and their implicit preference. Extensive experiments verify the effectiveness of our scheme, based on which we also provide deep insights that can benefit the future research.

1 INTRODUCTION
According to the Goldman Sachs, the 2016 online retail market of China for fashion products, including apparel, footwear, and accessories, has reached 187.5 billion US dollars, which demonstrates people’s great demand for clothing. In fact, apart from physiological needs, people also have esteem needs of clothes as dressing properly is of importance in daily life. As each outfit usually involves multiple complementary items (e.g., tops, bottoms, and shoes), the key to a proper outfit lies in the harmonious clothing matching to a great extent. However, not everyone is a natural-born fashion stylist, which makes choosing the matching clothes a tedious and even annoying daily routine. It thus deserves our attention to develop an effective clothing matching scheme to help people figure out the suitable match for a given item and make a harmonious outfit. Meanwhile, recent years have witnessed the proliferation of various online fashion-oriented communities, such as Polyvore and Chictopic, where fashion experts can share their fashion tips by showcasing their outfit compositions to the public, as shown in Figure 1. Currently, Polyvore embraces 20 million unique hits and has more than 3 million outfits created per month. Moreover, clothing items on Polyvore have not only the visual images with clean background but also rich contextual metadata, such as titles and categories. Such tremendous volume of outfit compositions with rich metadata naturally makes Polyvore a wonderful venue to investigate the code in clothing matching.

In this work, we aim to investigate the practical problem of clothing matching, without loss of generality, by particularly answering the question “which bottom matches the given top”. The problem we pose here primarily requires modeling human notion of the compatibility between fashion items. However, modeling such subtle notion regarding compatibility is non-trivial due to the following challenges. First, the compatibility between fashion items usually involves color, material, pattern, shape and other design factors. In addition, human notion of compatibility is not absolute but relative, as people can only tell that a pair of items...
As a matter of fact, similar to visual images, contextual descriptions of fashion items and the implicit preference among them model the coherent relationship between different modalities of network (BPR-DAE) for compatibility modeling, which jointly top and bottom). Ultimately, we propose a dual autoencoder scheme seamlessly integrates the multi-modal data (i.e., visual and contextual modalities) of fashion items to comprehensively model the compatibility among fashion items. Moreover, considering the factors affecting the compatibility among items can be highly sophisticated, we employ the autoencoder neural model to exploit the latent compatibility space and unify the complementary fashion items from heterogeneous spaces.

- We seamlessly exploit the knowledge from multiple modalities (visual and contextual modalities) of fashion items and model the modality relatedness to enhance the performance of compatibility modeling among fashion items.
- We construct a comprehensive fashion dataset FashionVC, which consists of both the images and contextual metadata of fashion items on Polyvore. We have released our compiled dataset, codes, and parameters \(^4\) to facilitate other researchers to repeat our experiments and verify their approaches.

The remainder of this paper is structured as follows. Section 2 briefly reviews the related work. The proposed BPR-DAE is introduced in Section 3. Section 4 details the dataset construction and the feature extraction. Section 5 presents the experimental results, followed by our concluding remarks in Section 6.

2 RELATED WORK

2.1 Fashion Analysis

Fashion domain recently has been attracting increasing attention from both the computer vision and multimedia research communities. Existing efforts mainly focus on clothing retrieval [26], clothing recommendation [13, 25], and fashionability prediction [24, 33]. For example, Liu et al. [25] proposed a latent Support Vector Machine (SVM) [5] model for occasion-oriented outfit and item recommendation, where the dataset of wild street photos was created by human annotation. Iwata et al. [15] proposed a topic model to recommend tops for bottoms with a small dataset collected from magazines. Due to the infeasibility of human annotated dataset, several pioneering works have resorted to other sources, where rich data can be harvested automatically. For example, Hu et al. [14] studied the problem of personalized whole outfit recommendation over a dataset collected from Polyvore. McAuley et al. [27] presented a general framework to model human visual preference for a pair of objects based on the Amazon co-purchase dataset. They extracted visual features with CNNs and introduced a similarity metric to model human notion of complement objects. Similarly, He et al. [10] introduced a scalable matrix factorization approach that incorporates visual signals of product images to fulfill the recommendation task. Although these works have achieved huge success, previous efforts on fashion analysis mainly focus on the visual signals but fail to take the contextual information into account.

4 http://neurostylis.farbox.com/
where we jointly model the coherent relationship between visual and contextual modalities and the implicit preference among items via the Bayesian Personalized Ranking. C: category, T: title. “\text{-}” indicates the category hierarchy.

2.2 Representation Learning

Representation learning has long been an active research topic for machine learning, which aims to learn more effective representations for data, as compared to hand-designed representations, and hence achieve better performance for machine learning tasks [23, 41, 42]. In particular, recently, the advances in neural networks also propelled a handful of models, such as autoencoders (AE) [28], deep belief networks (DBN) [12], deep Boltzmann machine (DBM) [8] and convolutional neural networks (CNN) [22] to tackle various problems. For example, Want et al. [37] utilized deep autoencoders to capture the highly non-linear network structure and thus learn accurate network embedding. Due to the increasingly complex data and tasks, multi-view representation learning has attracted several research attempts. One basic training criterion that has been applied to multi-view representation learning is to learn a latent compact representation that can reconstruct the input as much as possible [36], where autoencoders are naturally adopted [6]. For example, Ngiam et al. [28] first proposed a structure based on multimodal autoencoders to learn the shared representation for speech and visual inputs and solve the problem of speech recognition. In addition, Wang et al. [36] proposed a multimodal deep model to learn image-text unified representations to tackle the cross-modality retrieval problem. Although representation learning has been successfully applied to solve cross-modality retrieval [4, 6], phonetic recognition [36] and multilingual classification [30], limited efforts have been dedicated to the fashion domain, which is the research gap we aim to bridge in this work.

3 NEURAL COMPATIBILITY MODELING

3.1 Notation

Formally, we first declare some notations. In particular, we use bold capital letters (e.g., \(X\)) and bold lowercase letters (e.g., \(x\)) to denote matrices and vectors, respectively. We employ non-bold letters (e.g., \(x\)) to represent scalars and Greek letters (e.g., \(\beta\)) to stand for parameters. If not clarified, all vectors are in column forms. Let \(\|A\|_F\) and \(\|x\|_2\) denote the Frobenius norm of matrix \(A\) and the Euclidean norm of vector \(x\), respectively.

3.2 Problem Formulation

In a sense, people prefer to choose clothes with high compatibility, such as a silk pushy bow blouse plus a mini skirt or a wool pullover plus a tweed lap skirt, to make a harmonious outfit. Consequently, in this work, we focus on the compatibility modeling towards clothing matching. Suppose we have a set of tops \(T = \{t_1, t_2, \ldots, t_{N_t}\}\) and bottoms \(B = \{b_1, b_2, \ldots, b_{N_b}\}\), where \(N_t\) and \(N_b\) denote the total number of tops and bottoms, respectively. For each top \(t_i\) (bottom \(b_j\)), we use \(v^T_i\) (\(v^B_j\)) \(\in \mathbb{R}^{D_v}\) and \(c^T_i\) (\(c^B_j\)) \(\in \mathbb{R}^{D_c}\) to represent its visual and contextual input features, respectively. \(D_v\) and \(D_c\) denote the dimensions of the corresponding input features. In addition, we have a set of positive top-bottom pairs \(S = \{(t_1, b_j), (t_2, b_{j_2}), \ldots, (t_{N_t}, b_{j_{N_b}})\}\) extracted from the outfit compositions on Polyvore, where \(N\) denotes the number of positive pairs. Accordingly, each top \(t_i\) has a positive bottom set \(B_i^+ = \{b_j \in B | (t_i, b_j) \in S\}\). Let \(m_{ij}\) denote the compatibility between top \(t_i\) and bottom \(b_j\). In this work, we aim to propose an accurate model to measure \(m_{ij}\), based on which we can generate a ranking list of \(b_j\)’s for a given \(t_i\).
3.3 Non-linear Compatibility Space

Obviously, it is not advisable to directly measure the compatibility between fashion items from distinct spaces due to their heterogeneity. Therefore, we assume that there exists a latent compatibility space that is able to bridge the gap between heterogeneous fashion items, where highly compatible fashion items that share similar style, material or functionality should also show high similarity. In fact, the factors contributing to compatibility may diversely range from style and color, to material and shape. Moreover, the relationship among these factors can be highly sophisticated. For example, a white casual T-shirt goes well with a black casual jeans but not a black suit, while a pair of high boots prefers skinny leggings rather than flared pants. Towards this end, in this work, we further assume that the subtle compatibility factors lie in a highly non-linear space, which can be learned by the advanced neural network models. In particular, we employ the autoencoder networks to learn the latent space, which has been proven to be effective in the latent space learning [37].

Autoencoder which works in an unsupervised manner, consists of two parts: the encoder and decoder. The encoder maps the input data to the latent representation space, while the decoder works toward mapping the latent representation space to a reconstruction space. Both of the encoder and decoder work based on multiple non-linear functions. Suppose the encoder consists of $K$ layers of non-linear transformation. Given the input $x$, the hidden representation for each layer can be calculated as follows,

$$
\begin{align*}
    h_1 &= s(W_1 x + b_1), \\
    h_k &= s(W_k h_{k-1} + b_k), & k = 2, \cdots, K,
\end{align*}
$$

where $h_k$ is the hidden representation, $W_k$ and $b_k$, $k = 1, \cdots, K$ are the matrices of weights and biases, respectively. $s : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear function applied element wise. In practice, the biases $b_k$ can be horizontally merged into the weight matrix $W_k$, while the input $x/h_k$ can be vertically appended by an entry 1. Therefore, to simplify the notation, we only consider $W_k$ and ignore the bias terms in the following discussion. We treat the output of the $K$-th layer as the latent representation $\hat{x} = h_K \in \mathbb{R}^t$, where $t$ denotes the dimensionality of the latent representation. Then the decoder computes inversely from the latent representation $\hat{x}$ to the reconstructed representation $\hat{x}$. Overall, for the input $x$, the autoencoder aims to minimize the reconstruction error as follows,

$$
    l(x) = \frac{1}{2} \| \hat{x} - x \|_2^2.
$$

3.4 Compatibility Measure

Table 1 lists several examples of fashion items in our dataset. Each fashion item is associated with an image, a title and several categories in terms of different granularity. Apparently, visual signals play significant roles in the compatibility measure, as many visual factors such as color and shape are encoded by the visual information. Moreover, we also observed that the context of each fashion item also present important characteristics of fashion items, such as the functionality and shape. Therefore, to comprehensively measure the compatibility between fashion items, we seamlessly explore the knowledge from both visual and contextual modalities.

In particular, we first feed the visual and contextual input features of tops and bottoms to four autoencoder networks $A^i_v$, $A^i_c$, $A^b_v$, and $A^b_c$, respectively. The superscripts $v$ and $c$ refer to the top and bottom. We thus obtain the latent visual and contextual representation for $t_i$ and $b_j$ as $\hat{v}_i$, $\hat{c}_i$, $\hat{v}_j$, $\hat{c}_j$. Then the decoder computes inversely from the latent representation to the reconstructed representation $\hat{v}_i$, $\hat{c}_i$, $\hat{v}_j$, $\hat{c}_j$, respectively. Based on such latent visual and contextual representations of tops and bottoms, we can define the compatibility between top $t_i$ and bottom $b_j$ as follows,

$$
    m_{ij} = (1 - \beta)(\hat{v}_i^T \hat{c}_j + \beta(\hat{c}_i^T \hat{c}_j)).
$$

where $\beta$ is the non-negative trade-off parameter.

Inspired by [34, 35], considering the coherent relationship between items’ images and contextual metadata, we further introduce the regularization to encourage the consistency between visual and contextual latent representation of the same fashion item $x_i$,

$$
    L_{mod}(\hat{v}_i, \hat{c}_i) = -\ln(\sigma(\hat{v}_i^T \hat{c}_i)).
$$

3.5 BPR-DAE

In a sense, we can easily identify the positive top-bottom pairs as which have been paired within the same outfits by fashion experts. Regarding the non-paired items (e.g., top-bottom pairs), they may just indicate the incompatibility between pairs or the missing potential positive pairs (i.e., pairs may be created in the future). Therefore, to fully take advantage of these implicit relationship between tops and bottoms, we naturally adopt the BPR framework. We assume that bottoms from the positive set $B^+_i$ are more favorable to top $t_i$ than those unobserved neutral bottoms. According to BPR, we build a training set:

$$
    D_S := \{(i, j, k) | t_i \in T, b_j \in B^+_i \land b_k \in B \setminus B^+_i\},
$$

where the triple $(i, j, k)$ indicates that bottom $b_j$ is more compatible than bottom $b_k$, with top $t_i$.

Then according to [31], we have the following objective function,

$$
    L_{bpr} = \sum_{(i, j, k) \in D_S} -\ln(\sigma(m_{ijk})),
$$

where $m_{ijk} := m_{ij} - m_{ik}$, capturing the compatibility preference between top $t_i$, bottom $b_j$ compared to bottom $b_k$, and the $\sigma$ is the sigmoid function. In addition, according to Eqn.(4) and taking the
modality consistency into consideration, we have
\[
\mathcal{L}_{mod} = \sum_{(i, j, k) \in \mathcal{D}_S} \left( \mathcal{L}_{mod}(\tilde{v}_i^j, \tilde{z}_i^j) + \mathcal{L}_{mod}(\tilde{v}_j^i, \tilde{z}_j^i) + \mathcal{L}_{mod}(\tilde{v}_k^i, \tilde{z}_k^i) \right).
\]
Finally, we have the following objective function,
\[
\mathcal{L} = \mathcal{L}_{bpr} + \gamma \mathcal{L}_{mod} + \mu \mathcal{L}_{rec} + \frac{\lambda}{2} \|\Theta\|^2_F,
\]
where \(\mathcal{L}_{rec} = \mathcal{L}_{rec} - \mathcal{L}_{rec} = \sum_{(i, j, k) \in \mathcal{D}_S} \left( l(v_i^j) + l(v_j^i) + l(v_k^i) \right), \) and \(\mathcal{L}_{rec} = \sum_{(i, j, k) \in \mathcal{D}_S} \left( l(v_i^j) + l(v_j^i) + l(v_k^i) \right). \) \(\mu, \gamma, \lambda\) are the nonnegative trade-off hyperparameters, and \(\Theta\) refers to the set of parameters (i.e., \(W_k\) and \(W_k\)). The last regularizer term is designed to avoid overfitting.

### 3.6 Optimization

Towards the optimization, the core step is to calculate the partial derivative with respect to parameters \(\partial \mathcal{L} / \partial W_k^{(bpr)}\) and \(\partial \mathcal{L} / \partial W_k^{(rec)}\), \(x \in \{t, b\}, y \in \{t, c\}\). Due to the space limitation, here we only introduce the detailed calculation for \(\partial \mathcal{L} / \partial W_k^{(bpr)}\) and \(\partial \mathcal{L} / \partial W_k^{(rec)}\), while the other partial derivative can be solved in similar fashion.

Taking advantage of the back-propagation strategy, we first calculate the \(\partial \mathcal{L}_{bpr} / \partial W_k^{(bpr)}\) and \(\partial \mathcal{L}_{mod} / \partial W_k^{(mod)}\) as follows,
\[
\begin{align*}
\frac{\partial \mathcal{L}_{bpr}}{\partial W_k^{(bpr)}} &= -\sigma(-m_{ijk}) \frac{\partial \hat{v}_i^j}{\partial W_k^{(bpr)}} (v_i^j - v_k^i), \\
\frac{\partial \mathcal{L}_{mod}}{\partial W_k^{(mod)}} &= -\gamma (z_i^j) \frac{\partial \hat{v}_i^j}{\partial W_k^{(mod)}} z_i^j, \\
\frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}} &= \mu (\hat{v}_i^j - v_i^j) \frac{\partial \hat{v}_i^j}{\partial W_k^{(rec)}}.
\end{align*}
\]

As \(\frac{\partial \hat{v}_i^j}{\partial W_k^{(bpr)}}\) and \(\frac{\partial \hat{v}_i^j}{\partial W_k^{(mod)}}\) can be derived from \(\hat{v}_i^j = \sigma(W_t^{(bpr)} h_{K-1} + b_t^{(bpr)})\) and \(\hat{v}_i^j = \sigma(W_t^{(mod)} h_{K-1} + b_t^{(mod)})\), we can easily access \(\frac{\partial \mathcal{L}_{bpr}}{\partial W_k^{(bpr)}}\) and \(\frac{\partial \mathcal{L}_{mod}}{\partial W_k^{(mod)}}\) and \(\frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}}\).

We then can iteratively obtain \(\frac{\partial \mathcal{L}_{bpr}}{\partial W_k^{(bpr)}}\) and \(\frac{\partial \mathcal{L}_{mod}}{\partial W_k^{(mod)}}\) and \(\frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}}\), \(k = K, \ldots, 1\). Meanwhile, we obtain \(\frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}}\) and \(\frac{\partial \mathcal{L}_{mod}}{\partial W_k^{(mod)}}\), \(k = K, \ldots, 1\) in the similar manner. We then employ the stochastic gradient descent to optimise the proposed model, where the network parameters can be updated as follows,
\[
\begin{align*}
W_k^{(bpr)} &\leftarrow W_k^{(bpr)} - \eta \frac{\partial \mathcal{L}_{bpr}}{\partial W_k^{(bpr)}} + \gamma \frac{\partial \mathcal{L}_{mod}}{\partial W_k^{(mod)}} + \mu \frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}} + \lambda W_k^{(rec)}, \\
W_k^{(rec)} &\leftarrow W_k^{(rec)} - \eta \frac{\partial \mathcal{L}_{rec}}{\partial W_k^{(rec)}} + \lambda W_k^{(rec)}.
\end{align*}
\]

where \(\eta\) is the learning rate.

### 4 DATASET AND FEATURES

#### 4.1 Dataset

In fact, several fashion datasets have been collected for different research purposes, for instance, the \(WwW\) [25], \(Exact\ Street2Shop\) [9], and \(Fashion-136K\) [16] datasets. However, most of existing released datasets are collected from wild street photos and thus inevitably involves clothing parsing technique, which still remains a great challenge in computer vision domain [39, 40]. In addition, these datasets lack the rich contextual metadata of each fashion item, which makes it intractable to fully model the fashion items. Therefore, to guarantee the evaluation quality and facilitate the experiment conduction, we constructed our own dataset \(Fashion\VC\) by crawling outfits created by fashion experts on Polyvore. In particular, we first collected a seed set of popular outfits on Polyvore, based on which we tracked 248 fashion experts. We then crawled the historical outfits published by them, based on which we construct the ground truth for positive item pairs. Considering that certain improper outfits can be accidentally created by users on Polyvore, we also set a threshold \(z = 50\) with respect to the number of “likes” for each outfit to ensure the quality of the positive fashion pairs. Finally, we obtained 20,726 outfits with 14,871 tops and 13,663 bottoms. For each fashion item, we particularly collected its visual image, categories and title description.

#### 4.2 Insights

Due to the limited space, we only show the most popular matching pairs of top and bottom categories\(^4\) in our dataset in Figure 5. Each circle denotes a fashion category, where the light blue refers to the top categories and the dark blue denotes the bottom ones. The areas of the circles and the widths of the links are proportional to the number of fashion items with the given category and the co-occurrence frequency between categories, respectively. It can be seen that knee length skirts, sweaters and T-shirts are the most compatible items, as they are all matched with various other category items. In addition, we found that coats go more with day dresses while sweaters match more knee length skirts. This also implies that the contextual information regarding each fashion item can be helpful in cloth matching.

\(^4\)Here we only consider the category at the finest granularity for each item.
We ultimately obtained a vocabulary of 3 where the evaluation pairs per top we randomly sample with less than 3 characters, which are more likely to be noise. 

**Contextual Modality.** Considering the short length of such contextual information, we utilize the bag-of-words scheme [17], which has been proven to be effective to encode contextual metadata [7]. We first constructed a style vocabulary based on the categories and the words in all the titles in our dataset. As such user-generated metadata can be inevitably noisy, we thus filtered out the categories and the words in all the titles in our dataset. As such user-generated contextual information, we utilize the bag-of-words scheme [17, 44], which is defined as,

$$AUC = \frac{1}{\left| T \right|} \sum \frac{1}{E(i)} \sum_{(j,k) \in E(i)} \delta(m_{ij} > m_{jk}).$$

where the evaluation pairs per top i are defined as,

$$E(i) := \{(j,k) | (i,j) \in S_{test} \land (i,k) \not\in S\}. \quad (12)$$

$$\delta(b)$$ is the indicator function that returns one if the argument b is true and zero otherwise.

For optimization, we employ the stochastic gradient descent (SGD) [1] with the momentum factor as 0.9. We adopt the grid search strategy to determine the optimal values for the regularization parameters (i.e., λ, µ, γ) among the values \$10^r | r \in \{-5, \ldots, -1\}\$. In addition, the mini-batch size, the number of hidden units and learning rate for all methods were searched in [32, 64, 128, 256, 512, 1024], [128, 256, 512, 1024], and [0.001, 0.01, 0.1], respectively. The proposed model was fine-tuned based on training set and validation set for 30 epochs, and the performance on testing set was reported. We experimentally found that the proposed model achieves the optimal performance with K = 1 hidden layer of 512 hidden units. All the experiments were conducted over a server equipped with four NVIDIA Titan X GPUs.

**5.2 On Model Comparison**

Due to the sparsity of our dataset, where matrix factorization based methods [29, 38] are not applicable, we only consider the following content-based baselines regarding compatibility modeling to evaluate the proposed model BPR-DAE. 

POD: We utilize the “popularity” of bottom b.j to measure its compatibility with top t.i. The “popularity” is defined as the number of tops that has been paired with b.j, and we thus have,

$$m_{ij} = |(i',j)|(i',j) \in S_{train}. \quad (13)$$

**RAND:** We randomly assign the scores of m_{ij} and m_{ik} to evaluate the compatibility between items.

**RAW:** We measure the compatibility score between top t.i and bottom b.j based on the similarity between their raw features directly as,

$$m_{ij} = \langle v_i^r | v_j^b + \beta(c_i^r) c_j^b \rangle. \quad (14)$$

**IBR:** We choose the image-based recommendation method proposed by [27], which aims to model the relationships between objects based on their visual appearance. This work also learns a visual style space, in which the retrieval of related objects is performed by nearest-neighbor search. Different from our model, this baseline learns the latent space by linear transformation and consider positive samples and negative samples independently. Moreover, this method only focuses on the visual information.

**ExIBR:** We extend IBR to handle both the visual and contextual data of fashion items, where we modify the distance function between top t.i and bottom b.j [27] as follows,

$$d_{ij} = \| v_i^r - v_j^b \|_2 + \beta \| (e_i^r - e_j^b) \|_2, \quad (15)$$

where Y, ∈ R^{Dx×K'} and Y, ∈ R^[Dx×K'] are the projection matrices for visual and contextual modality input, respectively. K’ refers to the dimension of the style space.

Table 3 shows the performance comparison among different approaches. From this table, we have the following observations: 1)
Table 3: Performance comparison of different approaches in terms of AUC.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>POP</td>
<td>0.4206</td>
</tr>
<tr>
<td>RAND</td>
<td>0.5094</td>
</tr>
<tr>
<td>RAW</td>
<td>0.5494</td>
</tr>
<tr>
<td>IBR</td>
<td>0.6075</td>
</tr>
<tr>
<td>ExIBR</td>
<td>0.7033</td>
</tr>
<tr>
<td>BPR-DAE</td>
<td>0.7616</td>
</tr>
</tbody>
</table>

POP achieves the worst performance, which propels us to further check the popular items in our dataset. Table 2 shows the five most popular tops and bottoms, respectively. We noticed that the popular fashion items are all in the basic style, such as plain T-shirts and jeans, which maybe due to the fact that they can go with many other items. Therefore, we can easily find the limitations of POP method. For example, most of the popular bottoms are jeans, which maybe not suitable for professional tops and sport outfits. Therefore, it is not advisable to adopt recommendation strategy based on popularity. 2) ExIBR and BPR-DAE both outperform the visual-based baseline IBR, which confirms the necessity of considering the contextual modality in compatibility modeling. 3) BPR-DAE shows superiority over ExIBR. One possible explanation is that the highly sophisticated compatibility space would be better characterized by the autoencoder neural networks rather than the linear transformation.

5.3 On Component Comparison

To verify the effectiveness of each component of our model, we also compared BPR-DAE with the following methods.

**BPR-DAE-Norec**: To check the component that regularizes the reconstruction error, we removed the $L_{rec}$ by setting $\mu = 0$.

**BPR-DAE-Nomod**: To check the modality regularizer component that controls the consistency between latent representations of different modalities, we removed the $L_{mod}$ by setting $\gamma = 0$.

**BPR-DAE-No**: We removed both the reconstruction and modality regularizers by setting $\mu = 0$ and $\gamma = 0$.

Table 4 shows the performance of our model with different component configurations. It can be seen that BPR-DAE outperforms all the other derivative models, which verifies the impact of each component in our model. For example, we noticed that BPR-DAE shows superiority over BPR-DAE-Nomod, which implies that the visual and contextual information of the same fashion items does share certain consistency in terms of characterizing the fashion items. Besides, the worse performance achieved by BPR-DAE-Norec as compared to BPR-DAE suggests that the latent compatibility space can be helpful to reconstruct the fashion items.

Table 4: Performance comparison of our model with different component configurations with respect to AUC.

<table>
<thead>
<tr>
<th>Approaches</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BPR-DAE</td>
<td>0.7616</td>
</tr>
<tr>
<td>BPR-DAE-Norec</td>
<td>0.7533</td>
</tr>
<tr>
<td>BPR-DAE-Nomod</td>
<td>0.7539</td>
</tr>
<tr>
<td>BPR-DAE-No</td>
<td>0.7421</td>
</tr>
</tbody>
</table>

5.4 On Modality Comparison

To verify the effectiveness of multi-modal integration, we also conducted experiments over different modality combinations. In particular, we adapt our model to BPR-DAE-V and BPR-DAE-C to cope with the visual and contextual modality of fashion items, respectively, by removing the other unnecessary autoencoder networks as well as the $L_{mod}$ regularizer. Figure 7 shows the comparative performance of different approaches with respect to AUC. We observed that BPR-DAE outperforms both BPR-DAE-V and BPR-DAE-C, which suggests that the visual and contextual information does complement each other and both contributes to the compatibility measurement between fashion items. It is surprising that BPR-DAE-C is more effective that BPR-DAE-V. One plausible explanation is that the contextual information is more concise to present the key features of fashion items.

To intuitively illustrate the impact of contextual information, we show the comparison between BPR-DAE and BPR-DAE-V on testing triples in Figure 8. As can be seen, contextual metadata works better in cases when the given two bottom candidates $b_j$ and $b_k$ share similar visual signals, such as color or shape, where visual signals could be insufficient to distinguish the compatibility between them with the given top $t_i$. Nevertheless, such contextual information may also lead to certain failed triples due to the category matching bias, especially when visual signals of bottom candidates differ significantly. For example, it is popular to match blouses with knee length skirts according to our dataset, which may thus lead to the first failed testing triple in the right most column.

To gain more detailed insights, we further check the performance of the proposed models on the seven most popular top categories. As can be seen from Figure 7, BPR-DAE still consistently shows superiority over both BPR-DAE-V and BPR-DAE-C on each of the seven top categories. Meanwhile, we found that contextual information significantly improves the performance on top categories such as “Jacket” and “Coat”, compared to “T-shirt” and “Cardigan” categories. One possible explanation is that the matching for coats and jackets would be more complicated [3] due to the fact that they serve people in more seasons and thus apart from the common color and pattern factors, we also need further consider other factors such as various material (e.g., silk and leather) and length (e.g., long and short). These factors may not be easy-learned from visual signals but can be effectively captured by the contextual information. On the contrary, regarding tops in basic styles, such as T-shirts and
Due to the limited space, we only list the key phrases of items’ contextual metadata.

Figure 8: Illustration of the comparison between BPR-DAE and BPR-DAE-V on testing triples. All the triples satisfy $m_{ij} > m_{ik}$.

Figure 9: Performance of different models with respect to MRR at different numbers of the bottom candidates $T$.

cardigans, where color and shape factors play more important roles in matching, the visual signal is more powerful than the context.

5.5 On Complementary Fashion Item Retrieval

To efficiently evaluate the proposed BPR-DAE towards the complementary fashion item retrieval, we adopted the common strategy [11, 21] that feeds each top $t_i$ appeared in $S_{test}$ as a query, and randomly selects $T$ bottoms as the candidates, where there is only one positive candidate. Then by passing them to the trained neural networks, getting their latent representations and calculating the compatibility score $m_{ij}$ according to Eqn.(3), we can generate a ranking list of these bottoms for the given top. In our setting, we care about the average position of the positive bottom in the ranking list and thus adopt the mean reciprocal rank (MRR) metric [19]. In total, we have 1, 954 unique tops in $S_{test}$, among which 1, 262 tops have never appeared in $S_{train}$ or $S_{valid}$.

Figure 9 shows the performance of different models in terms of MRR at different numbers of the bottom candidates $T$. It is worth mentioning that we dropped the POP baseline here due to the fact that the majority of tops share the same popularity of 1, which makes it intractable to generate the ranking. As can be seen, our model shows superiority over all the other baselines consistently at different numbers of bottom candidates, which verifies the effectiveness of our model in complementary fashion item retrieval and coping with the cold start problem. Certain intuitive ranking results for testing tops can be found in Figure 10. We noticed that although BPR-DAE sometimes failed to accurately rank the positive bottom at the first place, the neutral bottoms ranked before the positive one are also compatible with the given top, which is reasonable in the real application.

6 CONCLUSION AND FUTURE WORK

In this work, we present a content-based neural scheme (BPR-DAE) for compatibility modeling towards clothing matching (i.e., matching the tops and bottoms), which is able to jointly model the coherent relationship between different modalities of fashion items and the implicit preference among items via a dual autoencoder network. In addition, we constructed a comprehensive fashion dataset FashionVC, consisting of both the images and contextual metadata of fashion items on Polyvore. Experimental results demonstrated the effectiveness of our proposed scheme and verified the advantages of taking the contextual modality into consideration in terms of compatibility modeling. Surprisingly, we found that contextual modality even shows superiority over the visual modality, especially towards complicated tops (e.g., coats) rather than the basic ones (e.g., T-shirts). Currently, we fail to explore the category hierarchy to further enhance the compatibility modelling, which can be the future work direction.

ACKNOWLEDGMENTS

The work is supported by the National Natural Science Foundation of China under Grant No.: 61672322