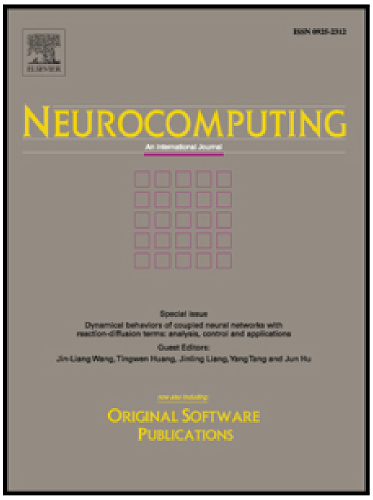

ISSN 0925-2312
NEUROCOMPUTING
An International Journal
Special issue
Dynamical behaviors of coupled neural networks with reaction-diffusion terms: analysis, control and applications
Guest Editors:
Jin-Liang Wang, Tingwen Huang, Jinling Liang, Yang Tang and Jun Hu
now also including:
ORIGINAL SOFTWARE PUBLICATIONS

Neural Fashion Experts: I Know How to Make the Complementary Clothing Matching

Jinhuan Liu, Xuemeng Song, Zhumin Chen, Jun Ma

PII: S0925-2312(19)30805-7
DOI: https://doi.org/10.1016/j.neucom.2019.05.081
Reference: NEUCOM 20873

To appear in: *Neurocomputing*

Received date: 4 November 2018
Revised date: 31 March 2019
Accepted date: 27 May 2019

Please cite this article as: Jinhuan Liu, Xuemeng Song, Zhumin Chen, Jun Ma, Neural Fashion Experts: I Know How to Make the Complementary Clothing Matching, *Neurocomputing* (2019), doi: https://doi.org/10.1016/j.neucom.2019.05.081

Neural Fashion Experts: I Know How to Make the Complementary Clothing Matching

Jinhuan Liu, Xuemeng Song*, Zhumin Chen, Jun Ma

School of Computer Science and Technology, Shandong University, Qingdao 266237, China

Abstract

Clothing has gradually become the beauty enhancing product, while the harmonious clothing matching is critical for a suitable outfit. The existing clothing matching techniques mainly rely on the visual features but overlook the textual metadata, which may be insufficient to comprehensively encode the fashion items. Nowadays, fashion experts are enabled to share their fashion tips by demonstrating their outfit compositions on the fashion-oriented online communities. Each outfit usually consists of several complementary fashion items (e.g., a top, a bottom and a pair of shoes), which involves an image along with the textual metadata (e.g., the categories and titles). The rich fashion data provide us an opportunity for the clothing matching, especially the complementary fashion item matching. In this work, we propose a multiple autoencoder neural network based on the Bayesian Personalized Ranking, dubbed BPR-MAE. Seamlessly exploring the multi-modalities (i.e., the visual and textual modalities) of fashion items, this framework is able to not only comprehensively model the compatibility between fashion items (e.g., tops and bottoms, bottoms and shoes) but also fulfill the complementary fashion item matching among multiple fashion items. Experimental results on the real-world dataset **FashionVC+** demonstrate the effectiveness of BPR-MAE, based on which we provide certain deep insights that can benefit the future research.

*Corresponding author

Email addresses: liujinhuan.sdu@gmail.com (Jinhuan Liu), sxmustc@gmail.com (Xuemeng Song), chenzhumin@sdu.edu.cn (Zhumin Chen), majun@sdu.edu.cn (Jun Ma)

Keywords: Multi-modal, compatibility modeling, complementary fashion item matching, fashion analysis.

1. Introduction

According to the Kantar China Insights, 71 percent of consumers purchased the apparel, while 51 percent of consumers purchased the shoes in the 2017 Double 11 shopping carnival[1], demonstrating the great demand of people for clothes. As a matter of fact, apart from the needs of daily necessities, more and more people are pursuing to dress up properly and in fashion. Due to that an outfit usually consists of multiple complementary items (e.g., a top, a bottom, and a pair of shoes), the key to a suitable outfit is the harmonious clothing matching to a great extent. Nevertheless, not everyone has a good taste in the clothing matching. Many people find it difficult to match the complementary fashion items and make proper outfits, facing the huge amount of fashion items. Consequently, it deserves our attention to develop an effective clothing matching scheme to help people figure out the suitable match for a given set of fashion items and make a harmonious outfit.

In recent years, various fashion-oriented online communities (e.g., Polyvore[2] and Chictopic[3]) have came into vogue. In comparison with the common shopping websites (e.g., eBay[4], Amazon[5] and Tmall[6]), these fashion sharing websites encourage fashion experts to create outfits, as shown in Fig. 1. Take the Polyvore as an example. Polyvore possesses more than two million products, and there are three million outfits created per month. Moreover, the fashion items on Polyvore not only have the visual images with clean background but also the rich textual metadata, such as the categories and titles. The tremen-

[1] `https://cn-en.kantar.com/consumer/shoppers/2017/2017-pre-singles-day/`.

[2] `http://www.polyvore.com/`.

[3] `http://www.chictopia.com/`.

[4] `https://www.ebay.com/`.

[5] `https://www.amazon.cn/`.

[6] `https://www.tmall.com/`.

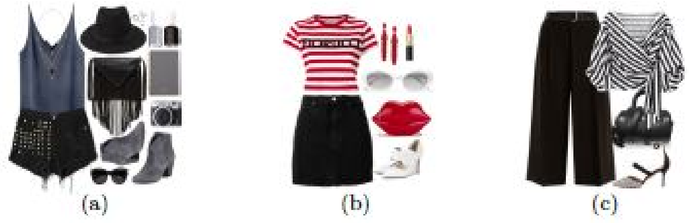


Figure 1: (a) Outfit 1. (b) Outfit 2. (c) Outfit 3. Examples of outfit compositions.

dous volume of fashion outfits naturally make Polyvore a good venue for us to investigate the fashion code of the clothing matching.

In this work, we investigate a practical problem of the complementary clothing matching. In fact, many research efforts have been dedicated to the clothing matching, which can be roughly organized into two groups: collaborative methods [1–3] and content-based methods [4–6]. The former one recommends fashion items that people with similar tastes and preferences liked based on their historical behaviors [7], such as the users' purchase behaviors [1], users' textual descriptions [2] and the behaviors of other users [3]. Apparently, such methods inevitably suffer from the data sparsity problem [8, 9]. The latter one tackles the problem by modeling the human preferences between fashion items based on their visual compatibility [4–6]. Despite the great success achieved by these efforts, most of them only leverage the visual modality of fashion items. Nevertheless, it may be insufficient to comprehensively model the compatibility between fashion items, as the textual modality can also contribute to the compatibility modeling. As can be seen from Fig. 2, the "Black Short with Lace Hem" seems to be visually compatible with the "Biker Jacket". Nevertheless, if we further consider the textual metadata, we can conclude that the lady short cannot go well with the jacket. As such, the compatibility measurement of fashion items can be benefited from exploiting the multi-modal data.

As a matter of fact, there are few multi-modal studies have been committed to the research on clothing matching [10–12]. For example, Song et al. [11] studies the compatibility modeling between two fashion items (e.g., a top and a

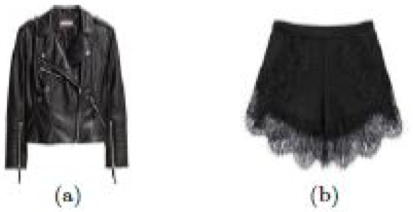


Figure 2: (a) Biker Jacket. (b) Black Short with Lace Hem. Illustration of the importance of the multi-modal data in clothing matching.

bottom) with both the visual and textual modalities, in the context of recommending bottoms for a given top. However, in most cases, one outfit contains more than two items (e.g., a top, a bottom and a pair of shoes), which makes their work less useful in practice. This paper builds upon our previous work presented in [11]. We take one step forward and propose a complementary fashion item matching scheme, which is able to recommend the complementary fashion items given a set of fashion items rather than a single item. As the top, bottom and shoes are the most essential fashion categories, we focus on answering the practical question of how to recommend a pair of shoes for the given top and bottom to make a suitable outfit. However, the complementary clothing matching with multiple fashion items is extremely challenging due to the following reasons.

a) In fact, different modalities can represent the same fashion item from different perspectives, as the color and pattern features of an item can be easily encoded by the visual information, while the functionality and material characters may be better conveyed by the textual information. Hence, how to mine the intrinsic relatedness between the textual and visual modalities is a big challenge.

b) The compatibility among different fashion items (e.g., tops, bottoms and shoes) can be rather complex, which is usually affected by many factors such as the color, shape and functionality. Therefore, how to accurately measure the compatibility among complementary fashion items is a difficult challenge.

c) According to our preliminary study based on the dataset crawled from

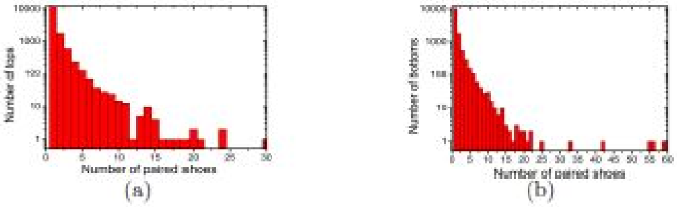


Figure 3: Distribution of tops and bottoms with respect to the number of paired shoes in our dataset. The Y-axis adopts the logarithmic scale.

the polyvore, only 1, 183 (7.96%) of the 14, 870 tops and 1, 328 (9.72%) of the 13, 662 bottoms have been matched by fashion experts with more than two pairs of shoes, respectively, as shown in Fig. 3. Accordingly, the last challenge lies in the sparse relationship among fashion items.

To address the above challenges, we propose a multiple autoencoder neural network based on the Bayesian Personalized Ranking, dubbed BPR-MAE. Fig. 4 shows the the proposed BPR-MAE scheme in the context of the complementary clothing matching (i.e., matching a pair of shoes for the given top and bottom). To comprehensively model the compatibility among fashion items, the proposed framework exploits both the visual and textual modalities of fashion items. We employ the pretrained deep convolutional neural networks and the bag-of-words scheme to encode the visual images and the textual metadata, respectively. Furthermore, taking into account the factors affecting the compatibility among items can be rather complicated, we use the multiple autoencoder neural network to learn the latent compatibility space. Meanwhile, in order to make the utmost of the implicit feedback pertaining to the compatibility among fashion items, we further adopt the BPR framework to explore the matching preferences among the complementary fashion items (i.e., the tops, bottoms and shoes). In a sense, the proposed scheme is devised to unify the complementary fashion items by learning the latent compatibility space. Ultimately, we present a content-based BPR-MAE framework, which is able to jointly model the implicit preferences among fashion items and the relationship between different

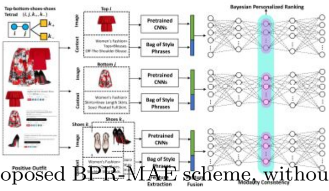


Figure 4: Illustration of the proposed BPR-MAE scheme, without loss of generality, devised to match the suitable shoes for a given top and bottom. “>” represents the category hierarchy.

modalities of them.

Our contributions can be summarized in three-folds.

a) We present a multiple autoencoder neural network, which is able to not only model the compatibility between fashion items, but also accomplish the complementary clothing matching by seamlessly exploring the multi-modalities (i.e., the visual and textual modalities) of fashion items.

b) We propose a content-based neural scheme (BPR-MAE) based on the BPR framework, which aims to learn the latent compatibility space and bridge the semantic gap among fashion items from heterogeneous spaces.

c) We construct the fashion dataset **FashionVC+**, consisting of both the visual and textual metadata of fashion items (i.e., the tops, bottoms and shoes) on Polyvore. We have released our code and parameters[7] to facilitate other researchers to repeat our work and verify their ideas.

The rest of the paper is organized as follows. Section II briefly reviews the related work. In Section III, we present the BPR-MAE framework. We then introduce the dataset construction, observation and feature extraction in Section IV. The experimental results and analyses are given in Section V, followed by the conclusion and future work in Section VI.

2. Related Work

2.1. Fashion Analyses

Fashion analysis has received extensive research interests in the computer vision and multimedia communities. Previous research works mainly focus on the

[7] `http://bpr_mae.bitcron.com/`.

clothing recognition [13, 14], clothing retrieval [15–17], clothing recommendation [5, 6, 18], compatibility modeling [19] and fashionability prediction [12, 20]. For example, Shen et al. [2] introduced a scenario-oriented fashion recommendation system which is devised to recommend outfits simply based on the attributes of the fashion items (e.g., brand, type and material). Obviously, here the high-level visual features of fashion items are not considered. Later, Jagadeesh et al. [21] harnessed both the visual contents of street fashion items and their annotations to recommend tops for a given bottom. Similarly, Liu et al. [4] integrated visual features and attributes to realize the occasion-oriented outfit and item recommendation by employing a latent Support Vector Machine (SVM) [22] framework, where the dataset is annotated manually. In fact, the annotation for the dataset can be rather intractable. Therefore, some efforts began to seek other sources to harvest rich data more effectively. For example, Zhang et al. [18] studied the correlation between clothing and locations by employing the fine-grained fashion items which are collected from several travel websites. Moreover, Hu et al. [6] presented a functional tensor factorization strategy for the personalized outfit recommendation with a dataset of fashion items with clean background crawled from Polyvore. McAuley et al. [5] introduced a scalable matrix factorization strategy for clothes recommendation by uncovering human notions of the visual relationships based on the Amazon clean background images. Despite the great success achieved by these works, existing efforts mainly focus on the visual features but overlook the textual modality. Towards this end, Li et al. [12] presented a multi-modal end-to-end deep learning framework to predict the set popularity. Different from previous works, we attempt to seamlessly exploit the multiple modalities of fashion items for the compatibility modeling between the fashion items and the complementary fashion item matching.

2.2. Representation Learning

As an important research point of machine learning, representation learning aims to learn more effective representations for data. Compared to the hand-designed representations, representation learning effectively improves the performance of machine learning tasks [23–25]. Particularly, with the advanced development of deep neural networks, a number of deep models are proposed to tackle the representation learning problems, such as deep boltzmann machine (DBM) [26, 27], deep belief networks (DBN) [28, 29], autoencoders (AE) [30, 31] and convolutional neural networks (CNN) [32, 33]. Want et al. [34] utilized the deep AE to obtain the highly non-linear network, where the accurate network embedding is learned. Due to the increasingly complex data and tasks, multi-model representation learning has received extensive research interests. For example, Ngiam et al. [30] presented the multi-modal deep AE to learn the shared representation for speech and visual inputs for speech recognition. Furthermore, Wang et al. [35] proposed a multi-modal deep model to learn an image-text representation for cross-modality retrieval. Despite the representation learning has been successfully applied for the multilingual classification [36], cross modality retrieval [37] and object recognition [38], there are limited efforts for the fashion analysis community. Accordingly, the major motivation of our work is to integrate the representation learning effectively.

3. Complementary Clothing Matching

3.1. Notation

Formally, we declare some notations used in this paper. Let bold capital letters (e.g., $\mathbf{X}$) represent the matrices, bold lowercase letters (e.g., $\mathbf{x}$) stand for the vectors and non-bold letters (e.g., x) denote scalars. In addition, we employ Greek letters (e.g., β) to denote parameters and all vectors are in column forms without clarification. The Frobenius norm of matrix $\mathbf{A}$ and the Euclidean norm of vector $\mathbf{x}$ are represented by $\left\|\mathbf{A}\right\|_F$ and $\left\|\mathbf{x}\right\|_2$, respectively.

3.2. Problem Formulation

We aim to propose a complementary fashion item matching scheme, which is able to recommend the complementary fashion items (e.g., a pair of shoes) for a given set of fashion items (e.g., a top and a bottom). In fact, the task posed here primarily requires us to model the compatibility between a pair of complementary fashion items. For simplicity, we first describe the compatibility modeling between bottoms and shoes, based on which we then present the complementary fashion item matching scheme in the context of recommending the compatible shoes for the given top and bottom.

Let $\mathcal{T} = \{t_1, t_2, \cdots, t_{N_t}\}$, $\mathcal{B} = \{b_1, b_2, \cdots, b_{N_b}\}$ and $\mathcal{S} = \{s_1, s_2, \cdots, s_{N_s}\}$ denote the set of tops, bottoms and shoes, where N_t, N_b and N_s refers to the number of the corresponding fashion items, respectively. For each fashion item x_i, which can be a top, a bottom and a pair of shoes, we denote its visual feature by $\mathbf{v}_i \in \mathbb{R}^{D_v}$ and textual feature as $\mathbf{c}_i \in \mathbb{R}^{D_c}$, where D_v and D_c stand for the feature dimensions of the respective modality. Let $\mathcal{P}^o = \{(t_{i_1}, b_{j_1}, s_{k_1}), (t_{i_2}, b_{j_2}, s_{k_2}), \cdots, (t_{i_M}, b_{j_M}, s_{k_M})\}$ stands for the positive top-bottom-shoes sets obtained from the online outfit composition community—Polyvore, where M is the total number of the positive sets. Accordingly, we can derive a positive shoes set $\mathcal{S}_{ij}^{+} := \{s_k \in \mathcal{S} | (t_i, b_j, s_k) \in \mathcal{P}^o\}$ for the given top t_i and bottom b_j.

Given the top t_i and bottom b_j, our goal is to recommend the appropriate shoes s_k to make proper outfits. In particular, we generate a ranking list of s_k for t_i and b_j by measuring the compatibility m_{ijk} among t_i, b_j and s_k. Towards this end, we use m_{ik} and m_{jk} to represent the compatibility between top t_i and shoes s_k and that between bottom b_j and shoes s_k, respectively.

3.3. Compatibility Modeling Between Fashion Items

Formally, we first present the compatibility modeling between the bottom and shoes. Obviously, measuring the compatibility between the bottom and shoes directly from their heterogenous spaces is not advisable as the semantic gap. Therefore, it is natural to assume that there exists a latent compatibility space, which can bridge the semantic gap between the heterogenous fashion items. In a sense, highly compatible fashion items should possess large similarity in the latent compatible space, as they may share similar color, style or material.

Considering that the compatibility factors can be rather complicated, we employ the neural networks, especially the deep AE, to learn the latent representation of the visual and textual modalities. Deep AE has been proven to be effective in the latent space learning [34].

As an unsupervised manner, AE consists of two components: the encoder and the decoder. The encoder maps the input data to the latent representation space, while the decoder maps the latent representation space to the reconstruction space. Both the encoder and decoder are implemented based on the multiple non-linear functions. Given the input $\mathbf{x}$, the hidden representation of each layer can be calculated as follows:

$$\begin{aligned} &\mathbf{h}_1 = s(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1), \\ &\mathbf{h}_k = s(\mathbf{W}_k\mathbf{h}_{k-1} + \mathbf{b}_k),\ k = 2, \cdots, K, \end{aligned} \tag{1}$$

where $\mathbf{h}_k$ denotes the hidden representation of the k-th layer, $\mathbf{W}_k$ and $\mathbf{b}_k$ are the matrices of weights and biases, respectively. $s : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear activation function and we use the sigmoid function $s(x) = \frac{1}{1+e^{-x}}$ in this paper. Suppose the encoder has K layers, and the output of the K-th layer hence can be regarded as the latent representation $\tilde{\mathbf{x}} = \mathbf{h}_K \in \mathbb{R}^L$, where L represents the dimensionality of the latent representation.

Then the decoder computes inversely from the latent representation $\tilde{\mathbf{x}}$ to the reconstructed representation $\hat{\mathbf{x}}$. Analogously, we use $\hat{\mathbf{v}}_{\mathbf{x}}$ and $\hat{\mathbf{c}}_{\mathbf{x}}$ to represent the reconstructed visual and textual representation of $\mathbf{v}_{\mathbf{x}}$ and $\mathbf{c}_{\mathbf{x}}$, respectively. The reconstruction error is defined as follows:

$$l(\mathbf{x}) = l(\mathbf{v}_{\mathbf{x}}) + l(\mathbf{c}_{\mathbf{x}}) = \frac{1}{2}\left\|\hat{\mathbf{v}}_{\mathbf{x}} - \mathbf{v}_{\mathbf{x}}\right\|_2^2 + \frac{1}{2}\left\|\hat{\mathbf{c}}_{\mathbf{x}} - \mathbf{c}_{\mathbf{x}}\right\|_2^2. \tag{2}$$

Moreover, to fully measure the compatibility between fashion items, we seamlessly explore both the visual and textual modalities. We thus use the visual and textual features of bottoms and shoes as input to deep AE A_v^b, A_c^b, A_v^s and A_c^s, respectively. The superscripts b and s stand for the bottom and shoes, while the subscripts v and c refer to the visual and textual features of each fashion item, respectively. Then the decoder computes inversely from the latent representation $\tilde{\mathbf{v}}_j^b$, $\tilde{\mathbf{c}}_j^b$, $\tilde{\mathbf{v}}_k^s$, $\tilde{\mathbf{c}}_k^s$ of bottom b_j and shoes s_k to the reconstructed representation $\hat{\mathbf{v}}_j^b$, $\hat{\mathbf{c}}_j^b$, $\hat{\mathbf{v}}_k^s$, $\hat{\mathbf{c}}_k^s$, respectively. Based on such latent visual and textual representations, we calculate the compatibility between bottom b_j and shoes s_k as follows:

$$m_{jk} = (1-\beta)(\tilde{\mathbf{v}}_j^b)^T \tilde{\mathbf{v}}_k^s + \beta(\tilde{\mathbf{c}}_j^b)^T \tilde{\mathbf{c}}_k^s, \tag{3}$$

where β is a non-negative trade-off parameter.

In reality, different modalities (e.g., the visual and textual modalities) can coherently characterize the same fashion item. Therefore, we define the $l_{mod}(x_i)$ to express the consistency between the visual latent representation $\tilde{\mathbf{v}}_i$ and textual latent representation $\tilde{\mathbf{c}}_i$ of the same fashion item x_i as follows:

$$l_{mod}(x_i) = -ln(\sigma(\tilde{\mathbf{v}}_i^T \tilde{\mathbf{c}}_i)), \tag{4}$$

where σ denotes the sigmoid function.

In a sense, it is easy to derive the positive bottom-shoes pairs from the outfits composed by fashion experts. Namely, we can derive $\mathcal{P} = \{(b_{j_1}, s_{k_1}), (b_{j_2}, s_{k_2}), \cdots, (b_{j_N}, s_{k_N})\}$ from the $\mathcal{P}^o$ as the positive bottom-shoes pairs, where N represents the number of positive matching pairs. Hence, for each bottom b_j, there is a positive shoes set $\mathcal{S}_j^+ := \{s_k \in \mathcal{S} | (b_j, s_k) \in \mathcal{P}\}$. In terms of the other unobserved pairs, they may imply the incompatibility or the potential positive pairs (i.e., bottom-shoes pairs that may be matched by experts in the future). Therefore, to fully tap the compatibility between bottoms and shoes, we adopt the BPR framework and we assume that the shoes from the positive set $\mathcal{S}_j^+$ are better matched to bottom b_j than those of unobserved neutral shoes. According to the BPR, we build the training set as follows:

$$\mathcal{D}_S := \{(j, k_+, k_-) | b_j \in \mathcal{B}, s_{k_+} \in \mathcal{S}_j^+ \wedge s_{k_-} \in \mathcal{S} \backslash \mathcal{S}_j^+\}, \tag{5}$$

where the triplet (j, k_+, k_-) indicates that shoes s_{k_+} is more compatible than shoes s_{k_-} with bottom b_j. Then according to [39], the BPR model can be

Figure 5: (a) Fringe Blanket Cardigan. (b) Scalloped Maxi Skirt. (c) Knee High Boots. Illustration of compatibility among the top, bottom and shoes.

defined as follows:

$$\mathcal{L}_{bpr} = \sum_{(j,k_+,k_-)\in\mathcal{D}_S} -ln(\sigma(m_{jk_+} - m_{jk_-})). \tag{6}$$

Moreover, according to Eqn.(4), we have:

$$\mathcal{L}_{mod} = \sum_{(j,k_+,k_-)\in\mathcal{D}_S} \Big(l_{mod}(b_j) + l_{mod}(s_{k_+}) + l_{mod}(s_{k_-})\Big). \tag{7}$$

On the basis of Eqn.(2), we define the reconstruction error as follows:

$$\mathcal{L}_{rec} = \sum_{(j,k_+,k_-)\in\mathcal{D}_S} \Big(l(b_j) + l(s_{k_+}) + l(s_{k_-})\Big). \tag{8}$$

With these notations, the following loss function is minimized to train the BPR-MAE network,

$$\mathcal{L} = \mathcal{L}_{bpr} + \gamma\mathcal{L}_{mod} + \mu\mathcal{L}_{rec} + \frac{\lambda}{2}\left\|\Theta\right\|^2, \tag{9}$$

where Θ refers to the weights $\mathbf{W}$ and bias $\mathbf{b}$ of BPR-MAE network. The hyper-parameters γ, μ, λ control the strength of reconstruction error $\mathcal{L}_{rec}$, modal loss $\mathcal{L}_{mod}$ and regularization term, respectively. Moreover, the regularization term is L_2 norm to prevent overfitting.

3.4. Complementary Fashion Item Matching

Based on the above compatibility formula modeling, by now, we have described the compatibility modeling between fashion items, and we thus can proceed to introduce the formulation towards the complementary fashion item

Table 1: Examples of fashion items in our dataset.

Id	1	2	3
Image			
Title	Off The Shoulder Ruffled Blouse	Plaid Ruffled Mini Skirt	Brown Pointed Toe Single Strap Over Ladies Heels
Category	Women's Fashion > Clothing > Skirts > Mini Skirts	Women's Fashion > Clothing > Tops > Blouses	Women's Fashion > Shoes > Pumps

matching. In particular, we aim to tackle the practical problem of recommending suitable shoes for the given top and bottom. Similarly, we use $\tilde{\mathbf{V}}^t = [\tilde{\mathbf{v}}_1^t, \tilde{\mathbf{v}}_2^t, \cdots, \tilde{\mathbf{v}}_{N_t}^t]$, $\tilde{\mathbf{C}}^t = [\tilde{\mathbf{c}}_1^t, \tilde{\mathbf{c}}_2^t, \cdots, \tilde{\mathbf{c}}_{N_t}^t]$ to represent the latent visual and textual representation of all tops. Thus, we use $\hat{\mathbf{v}}_i^t$, $\hat{\mathbf{c}}_i^t$ to denote the reconstructed representation of the latent representation $\mathbf{v}_i^t$, $\mathbf{c}_i^t$ for top t_i, respectively. Given the top t_i and bottom b_j, our goal is to recommend appropriate shoes s_k to make a proper outfit. In fact, the compatibility among multiple fashion items is more complicated than that between two items. Fig. 5 illustrates the compatibility among the top, bottom and shoes. As can be seen, the top "Fringe Blanket Cardigan" seems to be compatible with the pair of shoes "Knee High Boots". However, if we further consider the bottom "Scalloped Maxi Skirt", we can find that the pair of shoes "Knee High Boots" cannot go well with both the given top and bottom. Accordingly, we calculate the compatibility m_{ijk} among top t_i, bottom b_j and shoes s_k as follows:

$$m_{ijk} = m_{ik} + m_{jk}, \tag{10}$$

where m_{ik} and m_{jk} can be derived from Eqn.(3). Here, we assume that the given top and bottom would contribute equally regarding the compatibility measurement. Furthermore, we assume that the pairs of shoes from the positive set $\mathcal{S}_{ij}^+$ are more favorable to the given top t_i and bottom b_j than those of unobserved neutral shoes. We then have the following objective function:

$$\mathcal{L}_{bpr}^{o} = \sum_{(i,j,k_+,k_-)\in\hat{\mathcal{D}}_S} -ln(\sigma(m_{ijk_+} - m_{ijk_-})), \tag{11}$$

where $\hat{\mathcal{D}}_S$ is the training set constructed as follows:

$$\hat{\mathcal{D}}_S := \{(i,j,k_+,k_-)|(t_i,b_j) \in \mathcal{P}^o, s_{k_+} \in \mathcal{S}_{ij}^{+} \wedge s_{k_-} \in \mathcal{S}\backslash\mathcal{S}_{ij}^{+}\}, \tag{12}$$

where the quadruple (i,j,k_+,k_-) indicates that shoes s_{k_+} is more compatible than shoes s_{k_-} with the top and bottom pair (t_i, b_j). In addition, we have:

$$\mathcal{L}_{mod}^{o} = \sum_{(i,j,k_+,k_-)\in\hat{\mathcal{D}}_S} \Big(l_{mod}(t_i) + l_{mod}(b_j) + l_{mod}(s_{k_+}) + l_{mod}(s_{k_-})\Big). \tag{13}$$

According to Eqn.(2), we have:

$$\mathcal{L}_{rec}^{o} = \sum_{(i,j,k_+,k_-)\in\hat{\mathcal{D}}_S} \Big(l(t_i) + l(b_j) + l(s_{k_+}) + l(s_{k_-})\Big). \tag{14}$$

The training of the BPR-MAE network is optimized to minimized the following loss function,

$$\mathcal{L}^{o} = \mathcal{L}_{bpr}^{o} + \gamma\mathcal{L}_{mod}^{o} + \mu\mathcal{L}_{rec}^{o} + \frac{\lambda}{2}\left\|\Theta^{o}\right\|^{2}, \tag{15}$$

where Θ^o refers to the weights $\mathbf{W}$ and bias $\mathbf{b}$ between the layers of BPR-MAE framework. The hyperparameters γ, μ, λ control the strength of reconstruction error $\mathcal{L}_{rec}^{o}$, modal loss $\mathcal{L}_{mod}^{o}$ and regularization term, respectively. The regularization term is L_2 norm to prevent overfitting.

4. Dataset and Features

4.1. Dataset

We employ the rich fashion outfit data on Polyvore and construct our own fashion dataset **FashionVC+** to evaluate the compatibility fashion item matching. Firstly, a set of popular fashion outfits are collected from Polyvore, based on which we derived 248 fashion experts. Then, the historical outfits are further collected. Since this work aims to recommend the shoes for the given top and

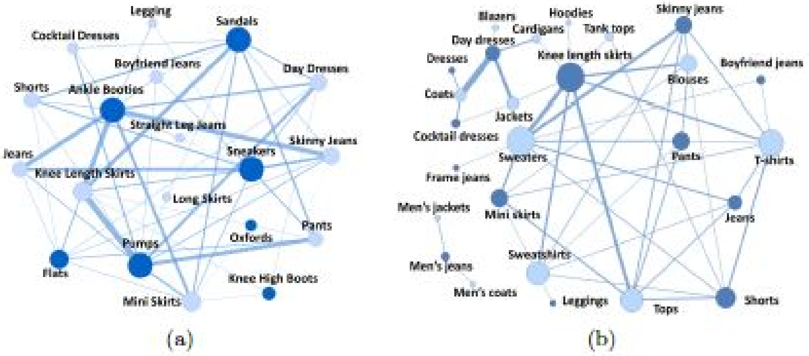


Figure 6: (a) The bottom and shoes categories. (b) The top and bottom categories. Illustration of the most popular matching pairs between different categories.

bottom, we only retain the outfits that contain at least a top, a bottom and a pair of shoes and take them as a positive example. Moreover, we set a threshold $z = 50$ regarding the number of "likes" for each outfit to ensure the quality of the positive fashion sets, due to the consideration that certain improper outfits can be accidentally created by users on Polyvore. Finally, we obtain 20,726 outfits with 14,870 tops, 13,662 bottoms and 14,093 pairs of shoes, respectively. In particular, the visual image, categories and title descriptions are collected for each fashion item. **ExpFashion** [40] dataset is crawled from the Polyvore website, which consists of 893991 outfits with 168682 tops and 117668 bottoms. Due to each outfit in ExpFashion dataset contains only a top and a bottom, we only carried out experiments on the compatibility modeling between tops and bottoms.

4.2. Data Observation

Table 1 shows several examples of fashion items in our dataset. As can be seen, the images with clean background depict the fashion items intuitively, while the textual metadata in short-length concisely summarize the key features of the fashion items. Accordingly, the coherent relationship between the textual metadata and the images propels us to explore the visual and textual modalities.

We further explore the validity of the textual information and illustrate the most popular matching pairs between the bottom and shoes categories[8] in our dataset. As shown in Fig. 6(a), we use the light blue and the dark blue circles to represent the bottom and shoes categories, respectively. The area of the circles is proportional to the number of fashion items with the corresponding categories. Furthermore, the width of the lines reflects the co-occurrence frequency between the different categories. As can be seen, ankle booties, sneakers, sandals, pumps, knee length skirts and skinny jeans seem to be the most compatible items, as they are all matched with various other categories of items. For example, it seems that knee length skirts go better with pumps while ankle booties match more with skinny jeans.

We also show the most popular matching pairs between the top and bottom categories in Fig. 6(b). The light blue represents the top categories, while the dark blue stands for the bottoms. We can see that the sweaters, T-shirts and knee length skirts are the most compatible items, as they can be matched with complementary items of various categories. Interestingly, we find that the coats match better with the day dresses while the sweaters match more with the knee-length skirts.

All of the above observations imply the textual metadata of each fashion item can contribute to the clothing matching.

4.3. Feature Extraction

Visual Modality: The advanced deep CNN feature which has been proven to be more discriminative than the traditional feature extraction methods (e.g., Speeded-Up Robust Features (SURF) [41], Scale Invariant Feature Transform (SIFT) [42] and Pyramid Histogram of Oriented Gradients (PHOG) [43]) for image representation learning [44–47] is utilized. Particularly, the pre-trained Alexnet model provided by the Caffe package [48] is employed. It consists of 5 convolutional layers followed by 3 fully-connected layers. The images of

[8] Here, we only consider the category at the finest granularity for each item.

the fashion items are fed into the CNN model, and the output of the 'fc7' layer is regarded as the visual feature. In this way, the visual modality can be represented by a $4,096$-D vector for each fashion item.

Textual Modality: In our dataset, each fashion item is associated with a textual title and several categories. Due to the short length of such textual information, we employ the bag-of-words (BOW) scheme [49] that has been proven to be capable of effectively encoding the textual metadata [50]. Firstly, a style vocabulary is constructed, which consists of representative categories and the words appeared in the titles of fashion items. Due to that user generated titles and categories can be inevitably noisy, the categories and words that appeared in less than 5 fashion items are filtered out. Moreover, the words with less than 3 characters are removed. Finally, a vocabulary with $3,345$ phrases is obtained. In this way, the textual modality is compiled with a $3,345$-D boolean vector for each fashion item.

5. Experiments

In this section, we introduce the experimental settings and provide the experimental results regarding the compatibility modeling between fashion items (i.e., bottoms and shoes, tops and bottoms), the complementary fashion item matching (i.e., matching a pair of shoes for the given top and bottom) and the complementary fashion item retrieval.

5.1. Experimental Settings

Regarding the compatibility modeling between the fashion items, for simplicity, we only introduce the experimental setting of that between the bottoms and shoes, where the settings of that between the tops and bottoms can be derived analogously. We randomly split the set of the positive bottom-shoes pairs $\mathcal{P}$ into three subsets: 80% for the training, 10% for the validation and 10% for the testing, which are denoted as $\mathcal{P}_{train}$, $\mathcal{P}_{valid}$ and $\mathcal{P}_{test}$, respectively. Then according to Eqn.(5), we can generate the triplet set $\mathcal{T}_{S_{train}}$, $\mathcal{T}_{S_{valid}}$ and

$\mathcal{T}_{S_{test}}$. Furthermore, for each positive bottom-shoes pair (b_j,s_{k+}), we randomly sample M shoes s_{k-} to construct M triplets (j, k_+, k_-), where M is set to 3 and $s_{k-} \notin \mathcal{S}_j^+$. Pertaining to the selection of hyperparameters and the performance comparison, we adopt the Area Under the ROC curve (AUC) [51, 52], which is defined as flows:

$$AUC = \frac{1}{|\mathcal{B}|} \sum_{(j)} \frac{1}{|E(j)|} \sum_{(k_+, k_-) \in E(j)} \delta(m_{jk_+} > m_{jk_-}), \tag{16}$$

where $\delta(\cdot)$ is the indicator function and the set of evaluation pairs for each bottom b_j is defined as follows:

$$E(j) := \{(k_+, k_-)|(j, k_+) \in \mathcal{P}_{test} \wedge (j, k_-) \notin \mathcal{P}\}. \tag{17}$$

In fact, the experimental settings of the complementary fashion item matching shares the same manner with that of the compatibility modeling between fashion items. In particular, we adaptively generate the quadruple sets $\mathcal{P}^o_{S_{train}}$, $\mathcal{P}^o_{S_{valid}}$ and $\mathcal{P}^o_{S_{test}}$ with a given set of fashion items (i.e., a top and a bottom). Ultimately, we define:

$$AUC^o = \frac{1}{|\mathcal{P}^o|} \sum_{(i,j)} \frac{1}{|E(i,j)|} \sum_{(k_+, k_-) \in E(i,j)} \delta(m_{ijk_+} > m_{ijk_-}), \tag{18}$$

where the set of evaluation pairs for each top and bottom pair (t_i,b_j) is constructed as:

$$E(i, j) := \{(k_+, k_-)|(i, j, k_+) \in \mathcal{P}^o_{test} \wedge (i, j, k_-) \notin \mathcal{P}^o\}. \tag{19}$$

Taking the complementary fashion item retrieval, we employ Mean Reciprocal Rank (MRR) to evaluate the retrieval performance, which is widely used in retrieval systems. The MRR can be calculated as follows:

$$MRR = \frac{1}{|\mathcal{P}^o|} \sum_{i=1}^{|\mathcal{P}^o|} \frac{1}{R_i}, \tag{20}$$

where $\mathcal{P}^o$ refers to the positive top-bottom-shoes sets obtained from the online outfit composition communityPolyvore. $|R_i|$ is the ranking position of the positive pair of shoes for the i-th top-bottom pair.

To optimize the objective function, we adopt the stochastic gradient descent (SGD) [53], which has proven to be effective on optimizing neural network models [54, 55]. To improve the performance, we set the momentum factor as 0.9. In addition, we employ the grid search strategy to determine the optimal values for the hyper parameters (i.e., μ, γ, λ) with the values of $\{10^r | r \in \{-5, \cdots, -1\}\}$. Moreover, we search the mini-batch size in $[32, 64, 128, 256, 512, 1024]$, the number of hidden units in $[128, 256, 512, 1024]$, the learning rate η in $[0.001, 0.01, 0.1]$ and the trade-off parameter β from 0 to 1 with the increment of 0.1. We fine-tuned the model based on the training set and validation set for 30 epochs, and reported the experimental results on the testing set.

5.2. Compatibility Modeling Between Bottoms And Shoes

We first evaluate the BPR-MAE framework in the context of the compatibility modeling between bottoms and shoes. In particular, we comprehensively evaluate the BPR-MAE framework on different models, different modalities and different components.

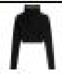
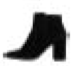
5.2.1. On Different Models

The sparse relationship among the fashion items in our dataset makes the methods based on matrix factorization [8, 9] not applicable. In this paper, we employ the following content-based methods to validate the **BPR-MAE** model.

Table 2: Performance comparison of the compatibility modeling between bottoms and shoes on different approaches.

Approaches	AUC
RAND	0.5073
POP	0.5793
RAW	0.5899
IBR	0.7184
ExIBR	0.7617
BPR-MAE	**0.8377**

Table 3: Examples of the five most popular fashion items.

Rank	1	2	3	4	5
Top					
Bottom					
Shoes					

RAND: We randomly assign the value of m_{jk_+} and m_{jk_-} to measure the compatibility between bottoms and shoes.

POP: We use the "popularity" of the shoes s_k to denote its compatibility with the bottom b_j. The "popularity" refers to the number of bottoms that have been paired with the shoes.

RAW: We directly measure the compatibility score m_{jk} based on the similarity between their raw features. The raw features are defined as $\mathbf{v}_j^b$, $\mathbf{c}_j^b$, $\mathbf{v}_k^s$, $\mathbf{c}_k^s$, where $\mathbf{v}$ and $\mathbf{c}$ denotes the raw visual and textual features of bottom b_j and shoes s_k. We thus have:

$$m_{jk} = (\mathbf{v}_j^b)^T \mathbf{v}_k^s + \beta(\mathbf{c}_j^b)^T \mathbf{c}_k^s. \quad (21)$$

IBR: We utilize the image-based recommendation method presented in [5], which models the relationships of their visual appearance between objects. This baseline only considers the visual information and learns the visual style space, in which the related objects are retrieved by the nearest-neighbor search. Nevertheless, this work learns the latent space via the simple linear transformation and independently explores the positive and negative samples.

ExIBR: We extend the **IBR** to deal with both the visual and textual metadata of fashion items, where we calculate the distance between the bottom b_j and the shoes s_k in [5] as follows:

$$d_{jk} = \left\|(\mathbf{v}_j^b - \mathbf{v}_s^k)\mathbf{Y}_v\right\|_2^2 + \eta \left\|(\mathbf{c}_j^b - \mathbf{c}_k^s)\mathbf{Y}_c\right\|_2^2, \quad (22)$$

where $\mathbf{Y}_v \in \mathbb{R}^{D_v \times K'}$ and $\mathbf{Y}_c \in \mathbb{R}^{D_c \times K'}$ refer to the projection matrices of the

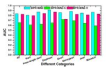


Figure 7: Performance of the proposed models on different bottom categories. “All” refers to the whole testing set.

visual and textual modalities, respectively. K' represents the dimension of the style space and η denotes the trade-off parameter from 0 to 1 with the increment of 0.1.

Table 2 shows the performance comparison in terms of the AUC among different approaches. From this table, we have the following observations:

a) As expected, **RAND** is the worst, achieving an AUC score around 0.5. Hence, it is not desirable to adopt the random matching strategy.

b) Although **POP** outperforms **RAND**, it still performs worse than the other methods. We thus further check the popular fashion items in our dataset. Table 3 shows the five most popular tops, bottoms and shoes in our dataset. As can be seen from the last two rows that the popular bottoms and shoes are all in basic styles and can match well with many other fashion items, such as the jeans, pumps and ankle booties. Meanwhile, it is easy to find the bias of **POP**. For example, most of the popular shoes are pumps and ankle booties, while they are not suitable for the formal and sport tops. Accordingly, it is not reasonable to adopt the popularity-based matching strategy.

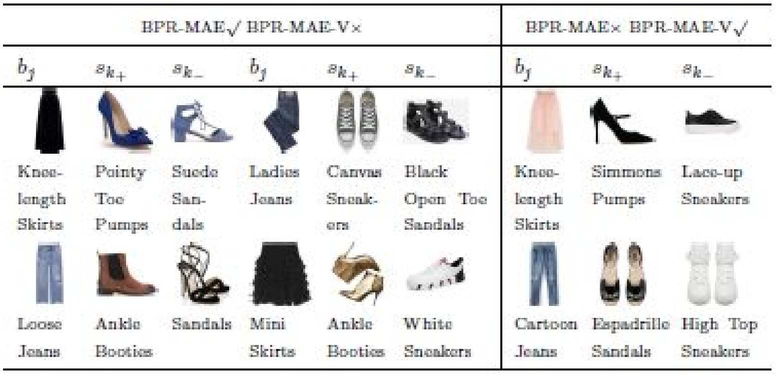
c) **ExIBR** and **BPR-MAE** are superior to the visual-based baseline **IBR**, which implies that the compatibility modeling does can be benefited from exploiting the textual modality.

d) **BPR-MAE** performs better than **ExIBR**. This maybe due to the fact that the highly complex compatibility space can be better characterized by the AE neural networks than the simple linear transformation.

5.2.2. On Different Modalities

To verify the importance of the multi-modal information, we conduct comparative experiments on different modality combinations. In particular, we adapt our model as **BPR-MAE-V** and **BPR-MAE-C** to handle the visual

Table 4: Illustration of the comparison between BPR-MAE and BPR-MAE-V on testing triples of the compatibility modeling between bottoms and shoes. Due to the limited space, we only list the key phrases of items' textual metadata.



BPR-MAE√ BPR-MAE-V×						BPR-MAE× BPR-MAE-V√		
b_j	s_{k_+}	s_{k_-}	b_j	s_{k_+}	s_{k_-}	b_j	s_{k_+}	s_{k_-}
Knee-length Skirts	Pointy Toe Pumps	Suede Sandals	Ladies Jeans	Canvas Sneakers	Black Open Toe Sandals	Knee-length Skirts	Simmons Pumps	Lace-up Sneakers
Loose Jeans	Ankle Booties	Sandals	Mini Skirts	Ankle Booties	White Sneakers	Cartoon Jeans	Espadrille Sandals	High Top Sneakers

and textual modalities of fashion items respectively, by removing the other unnecessary AE network and the $\mathcal{L}_{mod}$ regularizer. Fig. 7 shows the performance of different modality combinations with respect to the AUC. To obtain more insights in detail, in addition to the overall comparative assessment, we further test the performance of our model on the seven most popular bottom categories. From Fig. 7, we can observe that:

a) **BPR-MAE** shows superiority to the **BPR-MAE-V** and **BPR-MAE-C**, which indicates that both the visual and textual modalities contribute to the compatibility measurement between bottoms and shoes.

b) **BPR-MAE-V** is more effective than **BPR-MAE-C**. This may be on account of the fact that the visual signals capture more comprehensive and intuitive views of fashion items as compared to the textual signals.

c) Textual information significantly improves the performance of the bottom category "Short". This maybe due to that the common color and pattern factors of the shorts are easily obtained from the visual images, while the other factors such as the material (e.g., lace, chiffon, satin drill) and type (e.g., A-line, falbala, Ragged) factors can be easy-learned from the textual information.

To intuitively illustrate the impact of the textual information, we show the result comparison between **BPR-MAE** and **BPR-MAE-V** in Table 4. We can see that the textual metadata works better when the given positive shoes s_{k_+} and the negative shoes s_{k_-} share similar visual signals (e.g., the color and pattern), where visual signals may not be sufficient to distinguish the compatibility between fashion items. Meanwhile, the textual information may also cause some failures due to the category matching bias, especially when the visual signals of the bottom candidates are significantly different. For example, it is fashionable to match the knee-length skirts with pumps and the skinny jeans with ankle booties according to the textual statistics of our dataset. This may be the reason that lead to the failed testing triplet in the right most column.

5.2.3. On Different Components

To verify the importance of each component for the compatibility modeling between the bottoms and shoes, we compared our proposed model with the following methods.

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

BPR-MAE-Norec: To check the component that regularizes the reconstruction error, we removed $\mathcal{L}_{rec}$ by setting $\mu = 0$.

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

Table 5 shows the performance comparison among different components of our model, in which we can observe as follows:

Table 5: Performance comparison of the compatibility modeling between bottoms and shoes on different components.

Approaches	AUC
BPR-MAE	**0.8377**
BPR-MAE-Nomod	0.8132
BPR-MAE-Norec	0.7651
BPR-MAE-No	0.7607

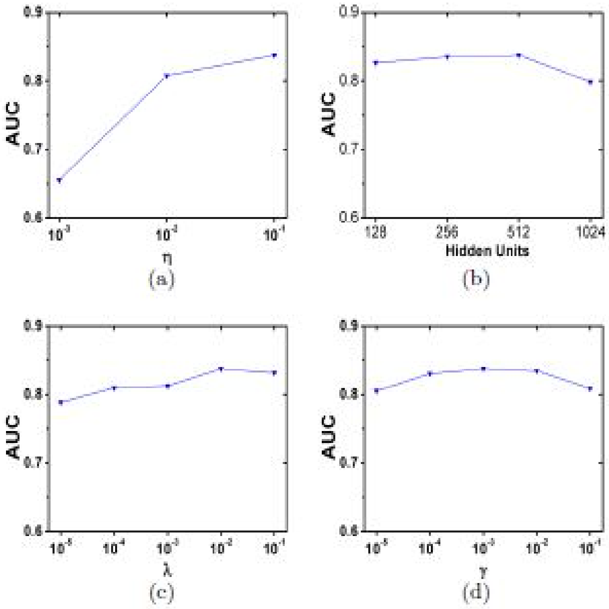


Figure 8: (a) Learning rate. (b) Hidden units. (c) Hyper parameter λ. (d) Hyper parameters γ. Illustration of the AUC values of our model respect to the different parameters.

a) **BPR-MAE** is superior to all the other derivative models, which verifies the effectiveness of each component in our model. For instance, we noticed that **BPR-MAE** performs better than **BPR-MAE-Nomod**, which means that the visual and textual information of the same fashion items is indeed consistent in describing the fashion items.

b) **BPR-MAE-No** performs worse than **BPR-MAE**, which indicates that both the compatibility space and the multi-modal information both contribute to the compatibility modeling between the bottoms and shoes.

5.2.4. On Different Parameters

We verify the performance of different parameters on the compatibility modeling between bottoms and shoes. Fig. 8 show the performance with respect to the learning rate η, the hidden units, the hyper parameters λ and γ. From this figure, we can see that overall, the performance of our model keeps steady

Table 6: Performance comparison of the compatibility modeling between tops and bottoms on FashionVC+ and ExpFashion datasets with different approaches.

Approaches	FashionVC+	ExpFashion
POP	0.4206	0.4975
RAND	0.5094	0.5392
RAW	0.5494	0.5916
IBR	0.6075	0.6845
ExIBR	0.7033	0.7394
BPR-MAE	**0.7616**	**0.7968**

nearby the optimal settings, which enables us to draw the conclusion that our model is non-sensitive to hyper-parameters. Experimental results show that the BPR-MAE model achieves the optimal performance with learning rate η of 0.1, hidden units of 512, λ of 0.01 and γ of 0.001.

5.3. Compatibility Modeling Between Tops And Bottoms

To fully evaluate the proposed model in terms of the compatibility modeling, we further conduct experiments on the compatibility modeling between the tops and bottoms. In particular, we evaluate the **BPR-MAE** in a similar manner, with different models, different modalities and different components.

5.3.1. On Different Models

Table 6 shows the performance on **FashionVC+** and **ExpFashion** datasets with different approaches. From this table, we can observe as follows:

a) **POP** performs the worst. Similarly, we checked the popular fashion items in our dataset. Table 2 shows the five most fashionable tops and bottoms. As can be seen that the popular tops and bottoms are also in the basic styles, such as the T-shirts and jeans, which also implies the limitation of **POP**. For instance, most of the fashionable bottoms are jeans, which maybe not suitable for the outdoor and professional tops. Therefore, it is also not appropriate

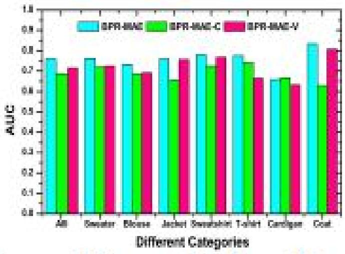


Figure 9: Performance of the proposed models on different top categories.

to realize the compatibility modeling between tops and bottoms based on the popularity of items.

b) **ExIBR** and **BPR-MAE** are more effective than **IBR**, which suggests the importance of the textual modality in the compatibility modeling between the tops and bottoms.

c) **BPR-MAE** is superior to the other methods on two dataset, which verifies the validity of our model. In addition, it is interesting that **BPR-MAE** performs better in the compatibility modeling between bottoms and shoes than that between tops and bottoms. We thus further analyze the visual and textual information of fashion items. We notice that the tops have more diverse attributes than the bottoms and shoes such as the various collar (e.g., round, v-neck and off-the-shoulder) and sleeve types (e.g., batwing, puff and mandarin), which may thus make the compatibility modeling between the tops and bottoms more difficult.

5.3.2. On Different Modalities

To verify the validity of the multi-modal information, we conducted the same experiments over different modality combinations as the compatibility modeling between bottoms and shoes. Fig. 9 shows the performance of different modalities combinations with respect to the AUC. From Fig. 9, we have the following observations:

a) **BPR-MAE** is superior to **BPR-MAE-V** and **BPR-MAE-C**, which indicates that both the visual and textual information can be conducive to the compatibility modeling between tops and bottoms.

Table 7: Illustration of the comparison between BPR-MAE and BPR-MAE-V on testing triples of the compatibility modeling between tops and bottoms. Due to the limited space, we only list the key phrases of items' textual metadata.

BPR-MAE√ BPR-MAE-V×						BPR-MAE× BPR-MAE-V√		
t_i	b_{j+}	b_{j-}	t_i	b_{j+}	b_{j-}	t_i	b_{j+}	b_{j-}
Fur Coat	Embellished Dress	Denim Skirt	HM Sweater	Eastwood Jeans	Waisted Shorts	Striped Blouse	Pinstriped Culottes	Knee Length Skirts
Cutton Sweatshirt	Skinny Jeans	Cotton Trousers	Men's Jackets	Biker Jeans	Skinny Jeans	Chubky Knit Jumper	Black Jeans	Knee Length Skirts

b) It is surprising that **BPR-MAE-C** is more effective than **BPR-MAE-V**. One reasonable explanation is that the textual information is more succinctly summarize the main features of tops and bottoms.

c) Compared with the "T-shirts" and "Cardigans" categories, the textual information significantly improves the performance of the "Jackets" and "Coats" categories. This maybe due to the fact that the coats and jackets serve people in more seasons [56]. In addition to the common color and pattern factors, we also consider other factors such as various materials (e.g., wool, silk, leather and fur) and length (e.g., short, medium and long). These factors may not be easy to capture from the visual signals but can easy-learned from the textual information. Neverthless, the tops in basic style categories, such as cardigans and T-shirts, where color, pattern and shape factors plays significant roles in the clothing matching between tops and bottoms, the visual information are more effective than the textual information.

To visually illustrate the impact of the textual modality, we illustrate the comparison between **BPR-MAE** and **BPR-MAE-V** in Table 7. We can see that the textual metadata works better when the given two bottom candidates b_{j+} and b_{j-} share similar visual signals, such as the color and shape, where visual

Table 8: Performance comparison of the compatibility modeling between tops and bottoms on different components.

Approaches	AUC
BPR-MAE	**0.7616**
BPR-MAE-Nomod	0.7539
BPR-MAE-Norec	0.7533
BPR-MAE-No	0.7421

signals may not be enough to distinguish the compatibility between them with the given top t_i. Furthermore, we get a consistent conclusion that such textual information may also lead to failures in the compatibility modeling due to the category matching bias, especially when there is a significant difference between the visual signals of the bottom candidates. For example, it is fashionable to match the blouses with knee length skirts in our dataset, which may lead to the failed matching in the right most column.

5.3.3. On Different Components

To verify the validity of each component of the compatibility modeling between tops and bottoms, we also compared our model with the **BPR-MAE-Norec**, **BPR-MAE-Nomod**, **BPR-MAE-No**.

Table 8 shows the performance comparison among different components of our model. We get the consistent conclusion that **BPR-MAE** is superior to all the other derivative models proposed in the compatibility modeling between bottoms and shoes. In addition, **BPR-MAE-No** performs the worst, which shows the necessity of both the latent compatibility space and the multi-modality information in terms of the compatibility modeling between tops and bottoms.

5.4. Complementary Fashion Item Matching

As we have evaluated our model regarding the compatibility modeling between fashion items, now we can proceed to evaluate the **BPR-MAE** in the context of the complementary fashion item matching. In particular, we aim

Table 9: Performance comparison of the complementary fashion item matching on different approaches.

Approaches	AUC
RAND	0.5025
POP	0.5928
RAW	0.5778
IBR	0.7012
ExIBR	0.7470
BPR-MAE	**0.8061**

to match the shoes for the given top and bottom and make a suitable outfit. Similarly, we also evaluate the proposed **BPR-MAE** framework in this context with different models, different modalities and different components.

5.4.1. On Different Models

Table 9 shows the performance comparison among different approaches. From this table, we have the following observations:

a) As expected, **RAND** is worst, achieving an AUC score around 0.5.

b) **RAW** performs better than **RAND**. This may be attributed to the fact that the negative shoes are randomly sampled to construct the quadruple sets, which is easy to choose the incompatibility examples. We further notice that **RAW** performs worse than other methods. The reasonable explanation is that the latent representation space can effectively model the complementary between the heterogeneous fashion items.

c) **ExIBR** and **BPR-MAE** are superior to the visual-based baseline **IBR**, which confirms the necessity of considering the textual modality in the complementary fashion item matching.

d) **BPR-MAE** shows superiority over **ExIBR**, which is consistent with the conclusion that in the compatibility modeling between fashion items.

Table 10: Performance comparison of the complementary fashion item matching on different components.

Approaches	AUC
BPR-MAE	**0.8061**
BPR-MAE-Nomod	0.7401
BPR-MAE-Norec	0.7627
BPR-MAE-No	0.7349

5.4.2. On Different Components

To verify the effectiveness of each component of the complementary fashion item matching, we compared the performance of **BPR-MAE** with the **BPR-MAE-Nomod**, **BPR-MAE-Norec**, **BPR-MAE-No**. Different from the compatibility modeling between fashion items, all these methods are adaptively generated by removing part of the $\mathcal{L}^o$.

Table 10 shows the performance compared with different components of our model, we can observe that:

a) **BPR-MAE** outperforms all the other derivative models, which verifies the impact of each component in our model.

b) **BPR-MAE-Norec** performs better than **BPR-MAE-Nomod**. This maybe due to the multi-modalities are more conducive to the complementary fashion item matching than the latent compatibility space.

c) The same as the aforementioned conclusion is that **BPR-MAE-No** performs the worst, which implies that both the compatibility space and the multi-modalities can be helpful to the complementary fashion item matching.

5.5. Complementary Fashion Item Retrieval

To comprehensively evaluate the proposed **BPR-MAE**, we further evaluate the performance of our model with three complementary item retrieval tasks: the retrieval of shoes for given bottoms, the retrieval of bottoms for given tops and the retrieval of shoes for given tops and bottoms.

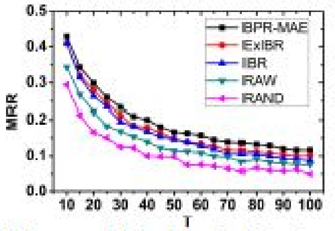


Figure 10: Performance of different models in the task of the shoes retrieval for given bottoms with various number of shoes candidates T.

5.5.1. The Retrieval Of Shoes For Given Bottoms

To evaluate the performance of **BPR-MAE** in the task of retrieving shoes for a given bottom, we adopt the common strategy [57, 58] that feeds each bottom b_j appeared in $\mathcal{P}_{test}$ as a query, and randomly selects T shoes as the candidates, among which there is only one positive candidate. Then by passing them to the neural networks trained on $\mathcal{P}_{train}$ and $\mathcal{P}_{valid}$, getting their latent representations and calculating the compatibility score m_{jk} according to Eqn.(3), we can generate a ranking list of these shoes for the given bottom. Since the ranking position of the one and only one positive pair of shoes is of great importance, we adopt the mean reciprocal rank (MRR) metric [59]. In total, we have $1,884$ unique bottoms in $\mathcal{P}_{test}$, among which $1,025$ bottoms have never appeared in $\mathcal{P}_{train}$ or $\mathcal{P}_{valid}$.

Fig. 10 shows the performance of different models in terms of MRR with various number of shoes candidates T. Here, we dropped the **POP** baseline due to the majority of bottoms are of the same popularity, which makes it intractable to generate the ranking. As can be seen, our model shows superiority over all the other baselines at different numbers of shoes candidates, which verifies the effectiveness of our model in the retrieval of shoes for given bottoms and coping with the cold start problem. Fig. 11 shows some instances illustrating

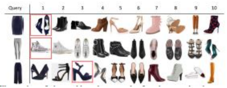


Figure 11: Illustration of the ranking shoes results for given testing bottoms. The shoes highlighted in the red boxes is the only positive ones.

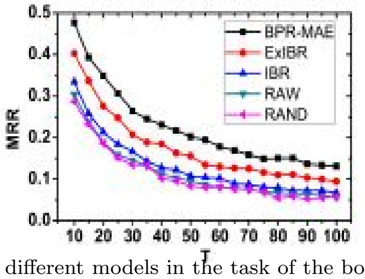


Figure 12: Performance of different models in the task of the bottom retrieval for given tops with various number of bottom candidates T.

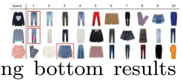


Figure 13: Illustration of the ranking bottom results for given testing tops. The bottoms highlighted in the red box is the positive one.

the ranking results for a given bottom. We notice that although **BPR-MAE** sometimes failed to accurately rank the positive pair of shoes at the first place, the other shoes ranked before the positive one are also compatible with the given bottom, which is reasonable in the practical application.

5.5.2. The Retrieval Of Bottoms For Given Tops

To evaluate the performance of **BPR-MAE** in the task of the retrieval of bottoms for given tops, we adopt the same strategy as above. In the testing set, there are $1,954$ unique tops, among which $1,262$ tops have never appeared in the training or validation set.

Fig. 12 shows the performance of different models in terms of MRR with various number of bottom candidates T. We can draw a similar conclusion that our model shows superiority over all the other baselines consistently at different numbers of bottom candidates. Fig. 13 shows some examples of the ranking results. We observe that **BPR-MAE** sometimes cannot accurately rank the positive bottom at the first place. The reason is that the other bottoms ranked before the positive one are potential positive bottoms.

5.5.3. The Retrieval Of Shoes For Given Tops And Bottoms

For the task of the retrieval of shoes for given tops and bottoms, we calculate the compatibility score m_{ijk} for top t_i, bottom b_j and shoes s_k according to Eqn.(10), and then generate a ranking list of shoes for the given top and bottom. In total, there are $2,076$ unique top and bottom pairs in the testing set, among which $1,235$ top and bottom pairs have never appeared in the training or validation set.

Fig. 14 shows the performance of different models in terms of MRR with various number of shoes candidates T. As can be seen, our model consistently outperforms the other baselines. Interestingly, we notice that our model performs better in the retrieval of the shoes given only the bottom than that given both the top and the bottom. Fig. 15 illustrates some examples of the ranking of shoes for a given top and bottom. We notice that **BPR-MAE** sometimes cannot accurately rank the positive shoes, the other shoes ranked before the positive one are also compatible with the given top and bottom.

6. Conclusion

In this work, we present a content-based neural scheme (BPR-MAE) regard to the compatibility modeling between fashion items (e.g., bottoms and shoes, tops and bottoms) and the complementary fashion item matching (e.g., matching a pair of shoes for the given top and bottom). This scheme is able to jointly model the coherent relation between different modalities (i.e., the visual and textual modalities) of fashion items and the implicit preferences among items via a multiple AE network. In addition, we construct a comprehensive fashion dataset **FashionVC+**, consisting of both images and textual metadata of fashion items on Polyvore. Experimental results demonstrate the effectiveness of our

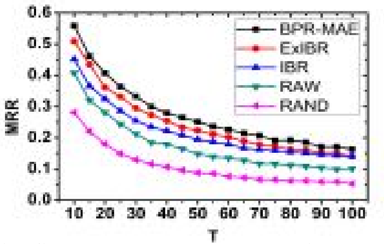


Figure 14: Performance of different models in the task of the shoes retrieval for given tops and bottoms with various number of shoes candidates T.

Figure 15: Illustration of the ranking shoes results for given testing tops and bottoms. The shoes highlighted in the red box is the only positive one.

proposed scheme and verify the advantages of taking the textual modality into consideration in terms of the compatibility modeling and the complementary clothing matching. Interestingly, we find that matching the shoes for a given bottom is easier than matching the bottom for a given top. In future work, we intend to explore the attention-based method to extract the important features of fashion items to further enhance the compatibility modeling.

7. Acknowledgements

This work is supported by the National Natural Science Foundation of China, No.: 61702300, and 61672322; the Fundamental Research Funds of Shandong University No.: 2018HW010.

References

[1] Y. Yue, C. Wang, K. El-Arini, C. Guestrin, Personalized collaborative clustering, in: ACM WWW, 2014, pp. 75–84.

[2] E. Shen, H. Lieberman, F. Lam, What am i gonna wear?: scenario-oriented recommendation, in: ACM IUI, 2007, pp. 365–368.

[3] J. Weston, C. Wang, R. Weiss, A. Berenzweig, Latent collaborative retrieval, in: Available: http://tensorflow.org/1206.4603, 2012.

[4] S. Liu, J. Feng, Z. Song, T. Zhang, H. Lu, C. Xu, S. Yan, Hi, magic closet, tell me what to wear!, in: ACM MM, 2012, pp. 619–628.

[5] J. McAuley, C. Targett, Q. Shi, A. Van Den Hengel, Image-based recommendations on styles and substitutes, in: ACM SIGIR, 2015, pp. 43–52.

[6] Y. Hu, X. Yi, L. S. Davis, Collaborative fashion recommendation: A functional tensor factorization approach, in: ACM MM, 2015, pp. 129–138.

[7] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions, IEEE TKDE 17 (6) (2005) 734–749.

[8] X. Wang, Y. Wang, Improving content-based and hybrid music recommendation using deep learning, in: ACM MM, 2014, pp. 627–636.

[9] X. Qian, H. Feng, G. Zhao, T. Mei, Personalized recommendation combining user interest and social circle, IEEE TKDE 26 (7) (2014) 1763–1777.

[10] C. Lynch, K. Aryafar, J. Attenberg, Images don't lie: Transferring deep visual semantic features to large-scale multimodal learning to rank, Available:http://arxiv.org/abs/1511.06746.

[11] X. Song, F. Feng, J. Liu, Z. Li, L. Nie, J. Ma, Neurostylist: Neural compatibility modeling for clothing matching, in: ACM MM, 2017, pp. 753–761.

[12] Y. Li, L. Cao, J. Zhu, J. Luo, Mining fashion outfit composition using an end-to-end deep learning approach on set data, IEEE TMM.

[13] M. H. Kiapour, K. Yamaguchi, A. C. Berg, T. L. Berg, Hipster wars: Discovering elements of fashion styles, in: CVPR, 2014, pp. 472–488.

[14] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: CVPR, 2016, pp. 1096–1104.

[15] S. Liu, Z. Song, G. Liu, C. Xu, H. Lu, S. Yan, Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set, in: CVPR, 2012, pp. 3330–3337.

[16] B. Siddiquie, R. S. Feris, L. S. Davis, Image ranking and retrieval based on multi-attribute queries, in: CVPR, 2011, pp. 801–808.

[17] P. Wu, S. C. Hoi, H. Xia, P. Zhao, D. Wang, C. Miao, Online multimodal deep similarity learning with application to image retrieval, in: ACM MM, 2013, pp. 153–162.

[18] X. Zhang, J. Jia, K. Gao, Y. Zhang, D. Zhang, J. Li, Q. Tian, Trip outfits advisor: Location-oriented clothing recommendation, IEEE TMM (2017) 2533–2544.

[19] X. Han, Z. Wu, Y.-G. Jiang, L. S. Davis, Learning fashion compatibility with bidirectional lstms, in: ACM MM, 2017, pp. 1078–1086.

[20] E. Simo-Serra, S. Fidler, F. Moreno-Noguer, R. Urtasun, Neuroaesthetics in fashion: Modeling the perception of fashionability, in: CVPR, 2015, pp. 869–877.

[21] V. Jagadeesh, R. Piramuthu, A. Bhardwaj, W. Di, N. Sundaresan, Large scale visual recommendations from street fashion images, in: ACM SIGKDD, 2014, pp. 1925–1934.

[22] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: CVPR, 2008, pp. 1–8.

[23] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436.

[24] S. Zhan, Q.-Q. Tao, X.-H. Li, Face detection using representation learning, Neurocomputing 187 (2016) 19–26.

[25] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, F. E. Alsaadi, A survey of deep neural network architectures and their applications, Neurocomputing 234 (2017) 11–26.

[26] R. Salakhutdinov, H. Larochelle, Efficient learning of deep boltzmann machines, in: AISTATS, 2010, pp. 693–700.

[27] T. Kuremoto, S. Kimura, K. Kobayashi, M. Obayashi, Time series forecasting using a deep belief network with restricted boltzmann machines, Neurocomputing 137 (2014) 47–56.

[28] G. E. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, Neural Computation 18 (7) (2006) 1527–1554.

[29] S. Zhou, Q. Chen, X. Wang, Fuzzy deep belief networks for semi-supervised sentiment classification, Neurocomputing 131 (2014) 312–322.

[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal deep learning, in: ACM ICML, 2011, pp. 689–696.

[31] Y. Zuo, J. Zeng, M. Gong, L. Jiao, Tag-aware recommender systems based on deep neural networks, Neurocomputing 204 (2016) 51–60.

[32] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: NIPS, 2012, pp. 1097–1105.

[33] S. Yu, S. Jia, C. Xu, Convolutional neural networks for hyperspectral image classification, Neurocomputing 219 (2017) 88–98.

[34] D. Wang, P. Cui, W. Zhu, Structural deep network embedding, in: DMKD, 2016, pp. 1225–1234.

[35] D. Wang, P. Cui, M. Ou, W. Zhu, Deep multimodal hashing with orthogonal regularization., in: IJCAI, Vol. 367, 2015, pp. 2291–2297.

[36] J. Rajendran, M. M. Khapra, S. Chandar, B. Ravindran, Bridge correlational neural networks for multilingual multimodal representation learning, Available:http://arxiv.org/abs/1510.03519.

[37] F. Feng, X. Wang, R. Li, Cross-modal retrieval with correspondence autoencoder, in: ACM MM, 2014, pp. 7–16.

[38] D. Ciregan, U. Meier, J. Schmidhuber, Multi-column deep neural networks for image classification, in: CVPR, 2012, pp. 3642–3649.

[39] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: UAI, 2009, pp. 452–461.

[40] Y. Lin, P. Ren, Z. Chen, Z. Ren, J. Ma, M. de Rijke, Explainable fashion recommendation with joint outfit matching and comment generation, arXiv preprint arXiv:1806.08977.

[41] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), CVIU 110 (3) (2008) 346–359.

[42] L. Nie, M. Wang, Z.-J. Zha, T.-S. Chua, Oracle in image search: a content-based approach to performance prediction, ACM TOIS 30 (2) (2012) 13.

[43] J. W. Ellison, Z. Wardak, M. F. Young, P. Gehron Robey, M. Laig-Webster, W. Chiong, Phog, a candidate gene for involvement in the short stature of turner syndrome, Human Molecular Genetics 6 (8) (1997) 1341–1347.

[44] J. Chen, X. Song, L. Nie, X. Wang, H. Zhang, T.-S. Chua, Micro tells macro: predicting the popularity of micro-videos via a transductive model, in: ACM MM, 2016, pp. 898–907.

[45] A. Khosla, A. Das Sarma, R. Hamid, What makes an image popular?, in: ACM WWW, 2014, pp. 867–876.

[46] H. Zhang, X. Shang, W. Yang, H. Xu, H. Luan, T.-S. Chua, Online collaborative learning for open-vocabulary visual classifiers, in: CVPR, 2016, pp. 2809–2817.

[47] Y.-G. Jiang, M. Li, X. Wang, W. Liu, X.-S. Hua, Deepproduct: Mobile product search with portable deep features, TOMM 14 (2) (2018) 50.

[48] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: ACM MM, 2014, pp. 675–678.

[49] R. Ji, X. Xie, H. Yao, W.-Y. Ma, Mining city landmarks from blogs by graph modeling, in: ACM MM, 2009, pp. 105–114.

[50] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, IEEE TIP 22 (1) (2013) 363–376.

[51] S. Rendle, L. Schmidt-Thieme, Pairwise interaction tensor factorization for personalized tag recommendation, in: ACM WSDM, 2010, pp. 81–90.

[52] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, T.-S. Chua, Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval, in: ACM MM, 2013, pp. 33–42.

[53] L. Bottou, Stochastic gradient learning in neural networks, Proc. Neuro-Nımes 91 (8).

[54] I. Sutskever, J. Martens, G. Dahl, G. Hinton, On the importance of initialization and momentum in deep learning, in: ACM ICML, 2013, pp. 1139–1147.

[55] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: ICML, 2004, p. 116.

[56] W. Di, C. Wah, A. Bhardwaj, R. Piramuthu, N. Sundaresan, Style finder: Fine-grained clothing style detection and retrieval, in: CVPR, 2013, pp. 8–13.

[57] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, T.-S. Chua, Neural collaborative filtering, in: ACM WWW, 2017, pp. 173–182.

[58] Y. Koren, Factorization meets the neighborhood: a multifaceted collaborative filtering model, in: ACM SIGKDD, 2008, pp. 426–434.

[59] L. Jiang, S.-I. Yu, D. Meng, Y. Yang, T. Mitamura, A. G. Hauptmann, Fast and accurate content-based semantic search in 100m internet videos, in: ACM MM, 2015, pp. 49–58.

Jinhuan Liu is currently a Ph.D. student from the school of Computer Science and Technology at Shandong University, Qingdao, China. Her research interests focus on information retrieval, machine learning and fashion analysis.

Xuemeng Song received the B.E. degree from University of Science and Technology of China in 2012, and the Ph.D. degree from the School of Computing, National University of Singapore in 2016. She is currently an assistant professor of Shandong University, Jinan, China. Her research interests include the information retrieval and social network analysis. She has published several papers in the top venues, such as ACM SIGIR, MM and TOIS. In addition, she has served as reviewers for many top conferences and journals.

Zhumin Chen is an associate professor in School of Computer Science and Technology of Shandong University. He is a member of the Chinese Information Technology Committee, Social Media Processing Committee, China Computer Federation Technical Committee (CCF) and ACM. He received his Ph.D. from Shandong University. His research interests mainly include information retrieval, big data mining and processing, as well as social media processing.

Jun Ma received the B.E., M.S., and Ph.D. degrees from Shandong University in China, Ibaraki University, and Kyushu University in Japan, respectively. He is currently a professor at Shandong University. He was a senior researcher in Ibaraki Univsity in 1994 and German GMD and Fraunhofer from 1999 to 2003. His research interests include information retrieval, Web data mining, recommendation systems and machine learning. He has published more than 150 International Journal and conference papers, including SIGIR, MM, TOIS and TKDE. He is a member of the ACM and IEEE.