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 multi-modal 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.

CCS CONCEPTS

•**Information systems → Retrieval tasks and goals;** *World Wide Web;*

KEYWORDS

Fashion Analysis, Compatibility Modeling, Multi-modal

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[1], 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[2] and Chictopic[3], 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.

(a) Composition 1. (b) Composition 2. (c) Composition 3.

Figure 1: Example outfit compositions on Polyvore.

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

* Corresponding author: Liqiang Nie (nieliqiang@gmail.com).

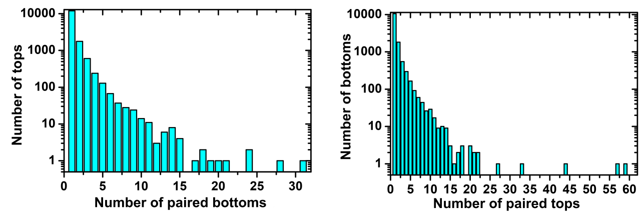
MM'17, Mountain View, CA USA

DOI: xx.xxx/xxx_x

[1]http://www.chinainternetwatch.com/19945/online-retail-2020.
[2]http://www.polyvore.com/.
[3]http://www.chictopia.com/.

(a) Hybrid tank top shirt. (b) Fairy Tulle Black Skirt.

Figure 2: Illustration of the importance of contextual modality in compatibility measurement.



(a) Top distribution (b) Bottom distribution

Figure 3: Distribution of tops and bottoms in our dataset. The Y-axis is in the logarithmic scale.

is more suitable to each other than other items. Therefore, how to accurately measure the compatibility between items constitutes a tough challenge. Second, existing works mainly focused on measuring the compatibility based on images of items but failed to take the contextual information of fashion items into consideration. As a matter of fact, similar to visual images, contextual descriptions also present the key features of fashion items and thus can be helpful in distinguishing compatible fashion items. For example, as shown in Figure 2, it maybe hard to predict whether the "Hybrid tank top shirt" is compatible with the "Fairy Tulle Black Skirt" with the current computer vision techniques due to their compatible color. However, if we further take the contextual descriptions into account, we can safely draw the conclusion that the lady skirt is not much suitable for the neutral top shirt. Therefore, how to model the intrinsic relatedness between the visual and contextual modalities of the same fashion item and further boost the performance is another crucial challenge. Last but not least, according to our pilot study on Ployvore, only 1,181 (7.94%) of 14,871 fashion tops have been paired with more than 2 bottoms, as can be seen from Figure 3. Such sparse relationship between fashion items makes the matrix factorization-based methods [29, 38] not applicable and hence poses another challenge for us.

To address these challenges, we present a content-based neural scheme for clothing matching (i.e., matching tops with bottoms), as shown in Figure 4. To deal with the sparse relationship between tops and bottoms, the proposed scheme learns a latent compatibility space to unify the complementary fashion items that come from the heterogenous spaces. In particular, the proposed 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 that the factors affecting the compatibility among items can be highly sophisticated, we employ the autoencoder neural model to exploit the latent compatibility space. Meanwhile, to take full advantage of the rich implicit semantics regarding the compatibility among fashion items on Polyvore, we further employ the advanced Bayesian Personalized Ranking (BPR) framework [31] to exploit the pairwise preference between complementary fashion items (i.e., tops and bottoms). Ultimately, we propose a dual autoencoder network (BPR-DAE) for compatibility modeling, which jointly models the coherent relationship between different modalities of fashion items and the implicit preference among them.

Our main contributions can be summarized in threefold:

- We propose a content-based neural scheme to model the compatibility between fashion items based on the BPR framework, which is able to learn the highly non-linear 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 fulfil 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

[4] http://neurostylist.farbox.com/.

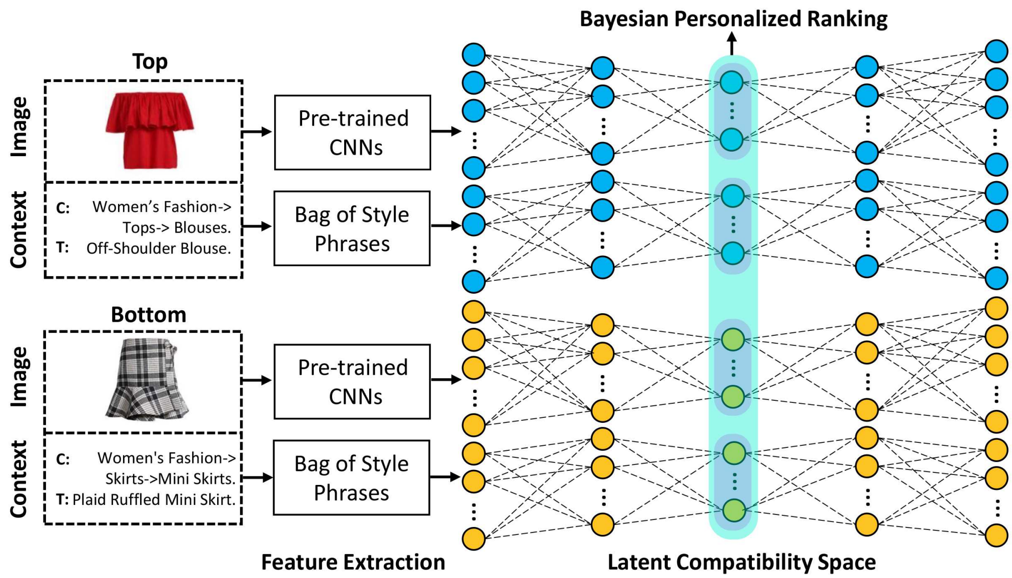


Figure 4: Illustration of the proposed scheme. We employ a dual autoencoder network to learn the latent compatibility space, where we jointly model the coherent relationship between visual and contextual modalities and the implicit preference among items via the Bayesian Personalized Ranking. C: category, T: title. "->" indicates the category hierarchy.

consideration. To bridge this gap, Li et al. [24] proposed a multi-modal multi-instance deep learning system to classify an given outfit as a popular or nonpopular one. Distinguished from the above works, we particularly focus on modeling the sophisticated compatibility between fashion items by seeking the non-linear latent compatibility space with neural networks. Moreover, we seamlessly aggregate the multi-modal data of fashion items and exploit the inherent relationship between different modalities to comprehensively model the compatibility between fashion items.

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., $\mathbf{X}$) and bold lowercase letters (e.g., $\mathbf{x}$) to denote matrices and vectors, respectively. We employ non-bold letters (e.g., x) to represent scalars and Greek letters (e.g., β) to stand for parameters. If not clarified, all vectors are in column forms. Let $\|\mathbf{A}\|_F$ and $\|\mathbf{x}\|_2$ denote the Frobenius norm of matrix $\mathbf{A}$ and the Euclidean norm of vector $\mathbf{x}$, 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 flap 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 $\mathcal{T} = \{t_1, t_2, \cdots, t_{N_t}\}$ and bottoms $\mathcal{B} = \{b_1, b_2, \cdots, b_{N_b}\}$, where N_t and N_b denote the total number of tops and bottoms, respectively. For each t_i (b_i), we use $\mathbf{v}_i^t$ ($\mathbf{v}_i^b$) $\in \mathbb{R}^{D_v}$ and $\mathbf{c}_i^t$ ($\mathbf{c}_i^b$) $\in \mathbb{R}^{D_c}$ 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 $\mathcal{S} = \{(t_{i_1}, b_{j_1}), (t_{i_2}, b_{j_2}), \cdots, (t_{i_N}, b_{j_N})\}$ 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 $\mathcal{B}_i^+ = \{b_j \in \mathcal{B} | (t_i, b_j) \in \mathcal{S}\}$. 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.

Table 1: Fashion item examples. "->": category hierarchy.

Id	Image	Category	Title
1		Women's Fashion -> Clothing ->Tops	River Island resort light blue denim halter neck top
2		Women's Fashion -> Clothing ->Skirts -> Mini Skirts	Plaid Ruffled Mini Skirt
3		Women's Fashion -> Jeans -> Flared Jeans	MiH Jeans mid-rise stretch-velvet flares jeans

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 heterogenous 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 nonlinear transformation. Given the input $\mathbf{x}$, the hidden representation for 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 $\mathbf{h}_k$ is the hidden representation, $\mathbf{W}_k$ and $\mathbf{b}_k$, $k = 1, \cdots, K$ are the matrices of weights and biases, respectively. $s : \mathbb{R} \mapsto \mathbb{R}$ is a non-linear function applied element wise[5]. In practice, the biases $\mathbf{b}_k$ can be horizontally merged into the weight matrix $\mathbf{W}_k$, while the input $\mathbf{x}/\mathbf{h}_k$ can be vertically appended by an entry 1. Therefore, to simplify the notation, we only consider $\mathbf{W}_k$ and ignore the bias terms in the following discussion. We treat the output of the K-th layer as the latent representation $\tilde{\mathbf{x}} = \mathbf{h}_K \in \mathcal{R}^L$, where L denotes the dimensionality of the latent representation. Then the decoder computes inversely from the latent representation $\tilde{\mathbf{x}}$ to the reconstructed representation $\widehat{\mathbf{x}}$. Overall, for the input $\mathbf{x}$, the autoencoder aims to minimize the reconstruction error as follows,

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

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_v^t, A_c^t, A_v^b and A_c^b, respectively. The superscripts t and b refer to the top and bottom. We thus obtain the latent visual and contextual representation for t_i and b_j as $\tilde{\mathbf{v}}_i^t$, $\tilde{\mathbf{c}}_i^t$, $\tilde{\mathbf{v}}_j^b$, $\tilde{\mathbf{c}}_j^b$. Then the decoder computes inversely from the latent representation to the reconstructed representation $\widehat{\mathbf{v}}_i^t$, $\widehat{\mathbf{c}}_i^t$, $\widehat{\mathbf{v}}_j^b$, $\widehat{\mathbf{c}}_j^b$, 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,

$$m_{ij} = (1 - \beta)(\tilde{\mathbf{v}}_i^t)^T \tilde{\mathbf{v}}_j^b + \beta(\tilde{\mathbf{c}}_i^t)^T \tilde{\mathbf{c}}_j^b, \tag{3}$$

where β 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,

$$\mathcal{L}_{mod}(\tilde{\mathbf{v}}_i, \tilde{\mathbf{c}}_i) = -ln(\sigma(\tilde{\mathbf{v}}_i{}^T \tilde{\mathbf{c}}_i)). \tag{4}$$

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 $\mathcal{B}_i^+$ are more favorable to top t_i than those unobserved neutral bottoms. According to BPR, we build a training set:

$$\mathcal{D}_S := \{(i, j, k) | t_i \in \mathcal{T}, b_j \in \mathcal{B}_i^+ \wedge b_k \in \mathcal{B} \backslash \mathcal{B}_i^+\}, \tag{5}$$

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,

$$\mathcal{L}_{bpr} = \sum_{(i,j,k) \in \mathcal{D}_S} -ln(\sigma(m_{ijk})), \tag{6}$$

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 σ is the sigmoid function. In addition, according to Eqn.(4) and taking the

[5] In this work, we use the sigmoid function $s(x) = \frac{1}{1+e^{-x}}$.

modality consistency into consideration, we have $\mathcal{L}_{mod}$ =

$$\sum_{(i,j,k)\in\mathcal{D}_S}\left(\mathcal{L}_{mod}(\tilde{\mathbf{v}}_i^t,\tilde{\mathbf{c}}_i^t)+\mathcal{L}_{mod}(\tilde{\mathbf{v}}_j^b,\tilde{\mathbf{c}}_j^b)+\mathcal{L}_{mod}(\tilde{\mathbf{v}}_k^b,\tilde{\mathbf{c}}_k^b)\right). \quad (7)$$

Finally, we have the following objective function,

$$\mathcal{L}=\mathcal{L}_{bpr}+\gamma\mathcal{L}_{mod}+\mu\mathcal{L}_{rec}+\frac{\lambda}{2}\left\|\Theta\right\|_F^2, \quad (8)$$

where $\mathcal{L}_{rec}=\mathcal{L}_{rec}^v+\mathcal{L}_{rec}^c$, $\mathcal{L}_{rec}^v=\sum_{(i,j,k)\in\mathcal{D}_S}\left(l(\mathbf{v}_i^t)+l(\mathbf{v}_j^b)+l(\mathbf{v}_k^b)\right)$, and $\mathcal{L}_{rec}^c=\sum_{(i,j,k)\in\mathcal{D}_S}\left(l(\mathbf{c}_i^t)+l(\mathbf{c}_j^b)+l(\mathbf{c}_k^b)\right)$. μ,γ,λ are the nonnegative trade-off hyperparameters, and Θ refers to the set of parameters (i.e., $\mathbf{W}_k$ and $\hat{\mathbf{W}}_k$). 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 $\partial\mathcal{L}/\partial\mathbf{W}_k^{xy}$ and $\partial\mathcal{L}/\partial\hat{\mathbf{W}}_k^{xy}$, $x\in\{t,b\}$, $y\in\{v,c\}$. Due to the space limitation, we here only introduce the detailed calculation for $\partial\mathcal{L}/\partial\mathbf{W}_k^{tv}$, $\partial\mathcal{L}/\partial\hat{\mathbf{W}}_k^{tv}$, while the other partial derivative can be solved in similar fashion.

Taking advantage of the back-propagation strategy, we first calculate the $\frac{\partial\mathcal{L}_{bpr}}{\partial\mathbf{W}_K^{tv}}$, $\frac{\partial\mathcal{L}_{mod}}{\partial\mathbf{W}_K^{tv}}$, and $\frac{\partial\mathcal{L}_{rec}}{\partial\mathbf{W}_K^{tv}}$ as follows,

$$\begin{cases}\dfrac{\partial\mathcal{L}_{bpr}}{\partial\mathbf{W}_K^{tv}}=-\sigma(-m_{ijk})\dfrac{\partial(\tilde{\mathbf{v}}_i^t)}{\partial\mathbf{W}_K^{tv}}(\mathbf{v}_j^b-\mathbf{v}_k^b)\\ \dfrac{\partial\mathcal{L}_{mod}}{\partial\mathbf{W}_K^{tv}}=-\gamma\sigma(-z_i^t)\dfrac{\partial(\tilde{\mathbf{v}}_i^t)}{\partial\mathbf{W}_K^{tv}}\tilde{\mathbf{c}}_i^t\\ \dfrac{\partial\mathcal{L}_{rec}}{\partial\hat{\mathbf{W}}_K^{tv}}=\mu(\hat{\mathbf{v}}_i^t-\mathbf{v}_i^t)\dfrac{\partial(\hat{\mathbf{v}}_i^t)}{\partial\hat{\mathbf{W}}_K^{tv}}.\end{cases} \quad (9)$$

As $\frac{\partial(\tilde{\mathbf{v}}_i^t)}{\partial\mathbf{W}_K^{tv}}$ and $\frac{\partial(\hat{\mathbf{v}}_i^t)}{\partial\hat{\mathbf{W}}_K^{tv}}$ can be derived from $\hat{\mathbf{v}}_i^t=\sigma(\hat{\mathbf{W}}_K^{tv}\hat{\mathbf{h}}_{K-1}^{tv}+\hat{\mathbf{b}}_K^{tv})$ and $\tilde{\mathbf{v}}_i^t=\sigma(\mathbf{W}_K^{tv}\mathbf{h}_{K-1}+\mathbf{b}_K^{tv})$, we can easily access $\partial\mathcal{L}_{bpr}/\partial\mathbf{W}_K^{tv}$, $\partial\mathcal{L}_{mod}/\partial\mathbf{W}_K^{tv}$ and $\partial\mathcal{L}_{rec}/\partial\hat{\mathbf{W}}_K^{tv}$. Then we can iteratively obtain $\partial\mathcal{L}_{bpr}/\partial\mathbf{W}_k^{tv}$ and $\partial\mathcal{L}_{mod}/\partial\mathbf{W}_k^{tv}, k=K,\cdots,1$. Meanwhile, we obtain the $\partial\mathcal{L}_{rec}/\partial\hat{\mathbf{W}}_k^{tv}$ and $\partial\mathcal{L}_{rec}/\partial\mathbf{W}_k^{tv}$, $k=K,\cdots,1$. 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,

$$\begin{cases}\mathbf{W}_k^{tv}\leftarrow\mathbf{W}_k^{tv}-\eta(\dfrac{\partial\mathcal{L}_{bpr}}{\partial\mathbf{W}_k^{tv}}+\gamma\dfrac{\partial\mathcal{L}_{mod}}{\partial\mathbf{W}_k^{tv}}+\mu\dfrac{\partial\mathcal{L}_{rec}}{\partial\mathbf{W}_k^{tv}}+\lambda\mathbf{W}_k^{tv})\\ \hat{\mathbf{W}}_k^{tv}\leftarrow\hat{\mathbf{W}}_k^{tv}-\eta(\mu\dfrac{\partial\mathcal{L}_{rec}}{\partial\mathbf{W}_k^{kv}}+\lambda\hat{\mathbf{W}}_k^{kv}),\end{cases} \quad (10)$$

where η 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 *WoW* [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 **FashionVC** 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.

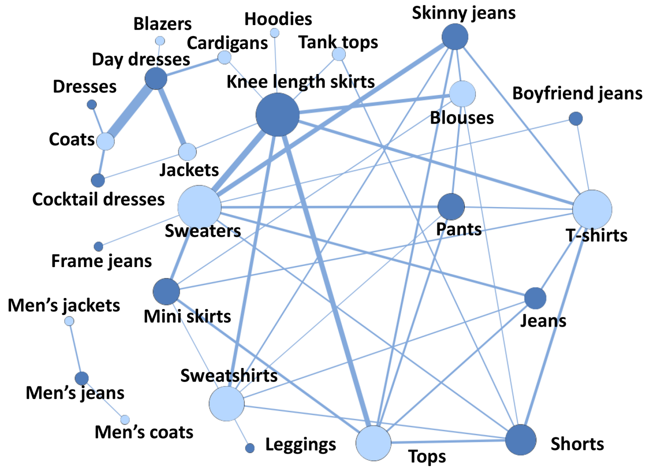


Figure 5: Illustration of the most popular matchings pairs between top and bottom categories.

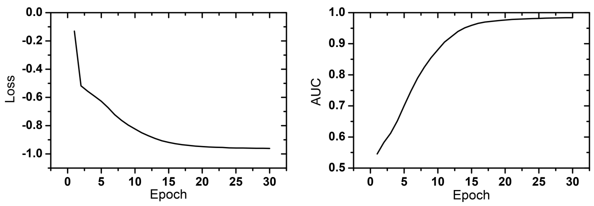
4.2 Insights

Due to the limited space, we only show the most popular matching pairs of top and bottom categories[6] 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.

4.3 Feature Extraction

Visual Modality. In this work, we take advantage of the advanced deep convolutional neural networks, which has been

[6]Here we only consider the category at the finest granularity for each item



(a) Training loss (b) AUC

Figure 6: Training loss and AUC with respect to each epoch.

proven to be the state-of-the-art model for image representation learning [2, 20, 43]. In particular, we chose the pre-trained ImageNet deep neural network provided by the Caffe software package [18], which consists of 5 convolutional layers followed by 3 fully-connected layers. We fed the image of each fashion item to the CNNs, and adopted the fc7 layer output as the visual feature. Therefore, for each item, its visual modality is represented by a 4096-D vector.

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 words that appeared in less than 5 items as well as the words with less than 3 characters, which are more likely to be noise. We ultimately obtained a vocabulary of 3,529 phrases, and hence compiled the contextual modality of each fashion item with a 3,529-D boolean vector.

5 EXPERIMENT

5.1 Experiment Settings

We separate the positive pair set $\mathcal{S}$ into three chunks: 80% of tripes for training, 10% for validation, and 10% for testing, which are denoted as $\mathcal{S}_{train}$, $\mathcal{S}_{valid}$ and $\mathcal{S}_{test}$, respectively. Then we generate the triple set $\mathcal{D}_{\mathcal{S}_{train}}$, $\mathcal{D}_{\mathcal{S}_{valid}}$ and $\mathcal{D}_{\mathcal{S}_{test}}$ according to Eqn.(5). In particular, for each positive top-bottom pair t_i and b_j, we randomly sample M bottoms b_k's to construct M triples (i, j, k), where $b_k \notin \mathcal{B}_i^+$ and M is set as 3. We adopt the widely used metric AUC (Area Under the ROC curve) [32, 44], which is defined as,

$$AUC = \frac{1}{|\mathcal{T}|}\sum_i \frac{1}{E(i)} \sum_{(j,k)\in E(i)} \delta(m_{ij} > m_{ik}), \tag{11}$$

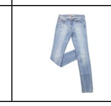
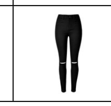
where the evaluation pairs per top i are defined as,

$$E(i) := \{(j,k)|(i,j) \in \mathcal{S}_{test} \wedge (i,k) \notin \mathcal{S}\}. \tag{12}$$

$\delta(b)$ 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 $\{10^r | r \in \{-5, \cdots, -1\}\}$. 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.

Table 2: Illustration of the most popular tops and bottoms.

Rank	1	2	3	4	5
Top					
Bottom					

We first experimentally verify the convergence of the proposed learning algorithm. We show the changes of the objective function in Eqn.(8) and the training AUC with one run of the training algorithm in Figure 6. As we can see, both values first change rapidly within a few epochs and then tend to go steady finally, which well demonstrates the convergence of our model.

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**.

POP: 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,

$$m_{ij} = |(i', j)|(i', j) \in \mathrm{S}_{train}|. \tag{13}$$

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,

$$m_{ij} = (\mathbf{v}_i^t)^T \mathbf{v}_j^b + \beta (\mathbf{c}_i^t)^T \mathbf{c}_j^b. \tag{14}$$

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 in [27] as follows,

$$d_{ij} = \left\|(\mathbf{v}_i^t - \mathbf{v}_j^b)\mathbf{Y}_v\right\|_2^2 + \beta \left\|(\mathbf{c}_i^t - \mathbf{c}_j^b)\mathbf{Y}_c\right\|_2^2, \tag{15}$$

where $\mathbf{Y}_v \in \mathcal{R}^{D_v \times K'}$ and $\mathbf{Y}_c \in \mathcal{R}^{D_c \times K'}$ 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.

Approaches	AUC
POP	0.4206
RAND	0.5094
RAW	0.5494
IBR	0.6075
ExIBR	0.7033
BPR-DAE	**0.7616**

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 $\mathcal{L}_{rec}$ 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 $\mathcal{L}_{mod}$ 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.

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

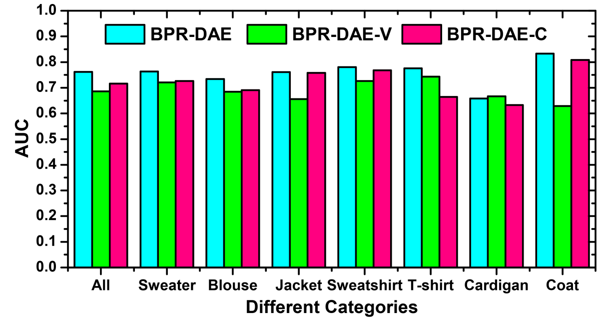


Figure 7: Performance of the proposed models on tops of different categories. "All" refers to the whole testing set.

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 $\mathcal{L}_{mod}$ 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

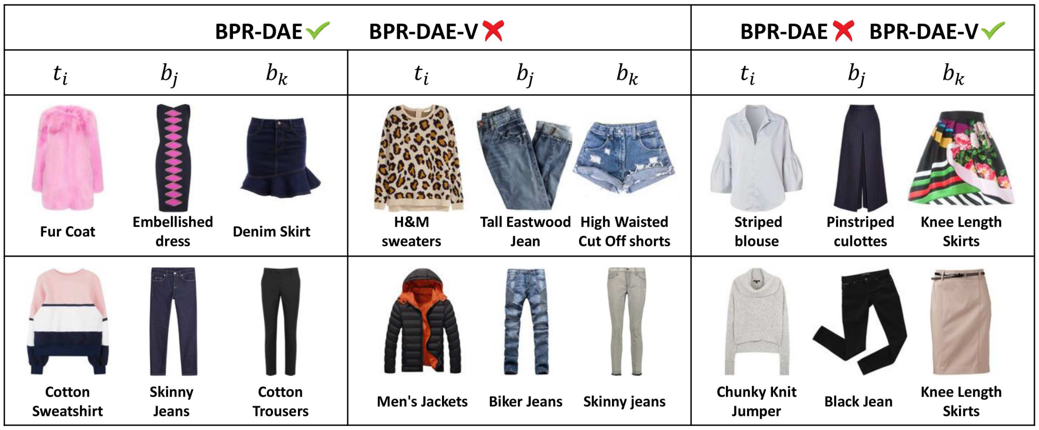


Figure 8: Illustration of the comparison between BPR-DAE and BPR-DAE-V on testing triples. All the triples satisfy $m_{ij} > m_{ik}$. Due to the limited space, we only list the key phrases of items' contextual metadata.

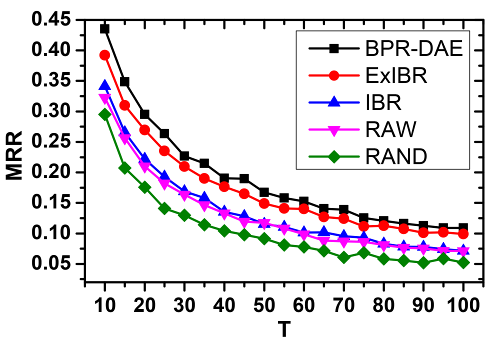


Figure 9: Performance of different models with respect to MRR at different numbers of the bottom candidates T.

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 $\mathcal{S}_{test}$ 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 $\mathcal{S}_{test}$, among which 1,262 tops have never appeared in $\mathcal{S}_{train}$ or $\mathcal{S}_{valid}$.

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

Figure 10: Illustration of the ranking results for given testing tops. The bottoms highlighted in the red boxes are the positive ones.

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

REFERENCES

[1] Léon Bottou. 1991. Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes* 91, 8 (1991).
[2] Jingyuan Chen, Xuemeng Song, Liqiang Nie, Xiang Wang, Hanwang Zhang, and Tat-Seng Chua. 2016. Micro tells macro: predicting the popularity of micro-videos via a transductive model. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 898–907.
[3] Wei Di, Catherine Wah, Anurag Bhardwaj, Robinson Piramuthu, and Neel Sundaresan. 2013. Style finder: Fine-grained clothing style detection and retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 8–13.
[4] Aviv Eisenschtat and Lior Wolf. 2016. Linking Image and Text with 2-Way Nets. *arXiv preprint arXiv:1608.07973* (2016).
[5] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. 2008. A discriminatively trained, multiscale, deformable part model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.
[6] Fangxiang Feng, Xiaojie Wang, and Ruifan Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 7–16.
[7] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. 2013. Visual-textual joint relevance learning for tag-based social image search. *IEEE Transactions on Image Processing* 22, 1 (2013), 363–376.
[8] Kostadin Georgiev and Preslav Nakov. 2013. A non-IID Framework for Collaborative Filtering with Restricted Boltzmann Machines.. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 1148–1156.
[9] M Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C Berg, and Tamara L Berg. 2015. Where to buy it: Matching street clothing photos in online shops. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 3343–3351.
[10] Ruining He and Julian McAuley. 2016. VBPR: Visual Bayesian Personalized Ranking from Implicit Feedback. In *Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, 144–150.
[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural Collaborative Filtering. In *Proceedings of the ACM International Conference on World Wide Web*. ACM, 173–182.
[12] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh. 2006. A fast learning algorithm for deep belief nets. *Neural computation* 18, 7 (2006), 1527–1554.
[13] Diane J Hu, Rob Hall, and Josh Attenberg. 2014. Style in the long tail: Discovering unique interests with latent variable models in large scale social e-commerce. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1640–1649.
[14] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: a functional tensor factorization approach. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 129–138.
[15] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the International Joint Conference on Artificial Intelligence*, Vol. 22. AAAI Press, 2262.
[16] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1925–1934.
[17] Rongrong Ji, Xing Xie, Hongxun Yao, and Wei-Ying Ma. 2009. Mining city landmarks from blogs by graph modeling. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 105–114.
[18] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 675–678.
[19] Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, and Alexander G Hauptmann. 2015. Fast and accurate content-based semantic search in 100m internet videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 49–58.
[20] Aditya Khosla, Atish Das Sarma, and Raffay Hamid. 2014. What makes an image popular?. In *Proceedings of the ACM International Conference on World Wide Web*. ACM, 867–876.
[21] Yehuda Koren. 2008. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 426–434.
[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*. 1097–1105.
[23] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.
[24] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining Fashion Outfit Composition Using An End-to-End Deep Learning Approach on Set Data. *IEEE Transactions on Multimedia* (2017).
[25] Si Liu, Jiashi Feng, Zheng Song, Tianzhu Zhang, Hanqing Lu, Changsheng Xu, and Shuicheng Yan. 2012. Hi, magic closet, tell me what to wear!. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 619–628.
[26] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. 2012. Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3330–3337.
[27] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. 2015. Image-based recommendations on styles and substitutes. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.
[28] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. 2011. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*. JMLR.org, 689–696.
[29] Xueming Qian, He Feng, Guoshuai Zhao, and Tao Mei. 2014. Personalized recommendation combining user interest and social circle. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2014), 1763–1777.
[30] Janarthanan Rajendran, Mitesh M Khapra, Sarath Chandar, and Balaraman Ravindran. 2015. Bridge correlational neural networks for multilingual multimodal representation learning. *arXiv preprint arXiv:1510.03519* (2015).
[31] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 452–461.
[32] Steffen Rendle and Lars Schmidt-Thieme. 2010. Pairwise interaction tensor factorization for personalized tag recommendation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*. ACM, 81–90.
[33] Edgar Simo-Serra, Sanja Fidler, Francesc Moreno-Noguer, and Raquel Urtasun. 2015. Neuroaesthetics in fashion: Modeling the perception of fashionability. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 869–877.
[34] Xuemeng Song, Liqiang Nie, Luming Zhang, Mohammad Akbari, and Tat-Seng Chua. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 213–222.
[35] Xuemeng Song, Liqiang Nie, Luming Zhang, Maofu Liu, and Tat-Seng Chua. 2015. Interest Inference via Structure-Constrained Multi-Source Multi-Task Learning.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2371–2377.
[36] Daixin Wang, Peng Cui, Mingdong Ou, and Wenwu Zhu. 2015. Deep Multimodal Hashing with Orthogonal Regularization.. In *Proceedings of the International Joint Conference on Artificial Intelligence*. AAAI Press, 2291–2297.
[37] Daixin Wang, Peng Cui, and Wenwu Zhu. 2016. Structural deep network embedding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1225–1234.
[38] Xinxi Wang and Ye Wang. 2014. Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 627–636.
[39] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. 2012. Parsing clothing in fashion photographs. In *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3570–3577.
[40] Kota Yamaguchi, M Hadi Kiapour, Luis E Ortiz, and Tamara L Berg. 2015. Retrieving similar styles to parse clothing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 5 (2015), 1028–1040.
[41] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. 2017. Visual Translation Embedding Network for Visual Relation Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[42] Hanwang Zhang, Xindi Shang, Huanbo Luan, Meng Wang, and Tat-Seng Chua. 2016. Learning from collective intelligence: Feature learning using social images and tags. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13 (2016).
[43] Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, and Tat-Seng Chua. 2016. Online collaborative learning for open-vocabulary visual classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2809–2817.
[44] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 33–42.