Enhancing Fashion Recommendation with Visual Compatibility Relationship

Ruiping Yin
Beijing Institute of Technology
School of Computer Science and Technology
Beijing, China
University of Technology Sydney
Centre for Artificial Intelligence
Sydney, Australia
yrp@bit.edu.cn

Kan Li
Beijing Institute of Technology
School of Computer Science and Technology
Beijing, China
likan@bit.edu.cn

Jie Lu
University of Technology Sydney
Centre for Artificial Intelligence
Sydney, Australia
Jie.Lu@uts.edu.au

Guangquan Zhang
University of Technology Sydney
Centre for Artificial Intelligence
Sydney, Australia
Guangquan.Zhang@uts.edu.au

ABSTRACT

With the increasing of online shopping services, fashion recommendation plays an important role in daily online shopping scenes. A lot of recommender systems have been developed with visual information. However, few works take into account compatibility relationship when they are generating recommendations. The challenge is that fashion concept is often subtle and subjective for different customers. In this paper, we propose a fashion compatibility knowledge learning method that incorporates visual compatibility relationships as well as style information. We also propose a fashion recommendation method with domain adaptation strategy to alleviate the distribution gap between the items in target domain and the items of external compatible outfits. Our results indicate that the proposed method is capable of learning visual compatibility knowledge and outperforms all the baselines.

CCS CONCEPTS

• **Information systems** → **World Wide Web**; *Web interfaces*; Web log analysis.

KEYWORDS

Fashion Recommendation; Viusal Compatibility; Image Representation

ACM Reference Format:
Ruiping Yin, Kan Li, Jie Lu, and Guangquan Zhang. 2019. Enhancing Fashion Recommendation with Visual Compatibility Relationship. In *Proceedings of the 2019 World Wide Web Conference (WWW'19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3308558.3313739

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313739

1 INTRODUCTION

A large portion of sales in the e-commerce are affected by fashion and lifestyle, which constitute apparel, footwear, bags, accessories, etc. Intelligent fashion recommendation received a lot of attention in computer vision and machine learning community [3, 9, 22]. They have huge potential profits for the fashion industry. A lot of companies have established their own recommender systems to give users advice to enhance their shopping experience, such as Amazon, Alibaba and eBay [14, 23].

Many approaches have been proposed to analyze user preferences on fashion criterion and generate personalized recommendation. Most of fashion recommendation approaches take into account characteristics of image, as visual information is one of the most important factor in describing fashion items [5, 9, 13]. Such approaches can substantially improve recommendation accuracy than others without visual information.

However, a few of them considers the problem of compatibility of fashion items. We know that when we choose a piece of clothing, it is not just a matter of considering the style of the dress. We also need to consider its effect with other clothes we wear. Some examples of compatibility and incompatibility outfits have been shown in Figure 1. Normally one would not pair a red T-shirt with green pants. Moreover, a black robe is incompatible with a pair of pink running shoes. This is partially because it is difficult to model compatibility relationship between fashion items.

When designing this recommendation system, we mainly consider the problem of learning visual compatibility relationship of items on pixel level. The visual compatibility relationship that needs to be learnt is whether fashion matching between one item and another item conforms to the human aesthetic by understanding the picture. In the traditional fashion recommendation, authors often only consider the styles and categories of clothes and ignore the sense of harmony between items as an outfit. In this paper, our approach considers visual compatibility relationship to recommend fashion items, which is closer to the actual needs of people.

(a) compatible outfits.

(b) incompatible outfits.

Figure 1: Examples of compatible and incompatible outfits.

In order to learn the knowledge of the matching model between fashion items, we face two challenges: 1) How to learn the common domain knowledge about fashion compatibility relationship between items. In other words, there are a few of outfits that we observed on the online shopping website for a single person and fashion concept is often subtle and subjective for different customers. 2) How to incorporate the learnt domain knowledge into our recommender system. For the first challenge, we propose a novel method to incorporate the compatibility relationship knowledge into the image representation. Our method allows learning an embedding from the images of the fashion items to a latent space, so that two items that is a good match are close in this latent space and items that don't match are far apart. An external dataset which contains a number of outfits being given by experts is also been used to train our model. For the second challenge, we adjust the popular Bayesian personalized ranking (BPR) [17] model to include the compatibility relationship knowledge that we learnt. Moreover, because we use the external dataset to learn the domain knowledge between the items in order to solve the problem of the distribution gap between the source domain and target domain, we propose a domain adaptation method to alleviate this difference.

In this paper, our contributions are as follows:

- We propose a fashion compatibility relationship learning method that incorporates visual compatibility relationships as well as style information into a visual embedding.
- We propose a fashion recommendation method with domain adaptation strategy to alleviate the distribution gap between the items in target domain and the items of external compatible outfits.
- We conduct a case study to illustrate how our method understands images. Furthermore, through an extensive set of experiments on several datasets, we demonstrate our method significantly outperforms several alternative methods.

The rest of this paper is organized as follows. In Section 2, we review some of the relevant methods. Section 3 introduces some notations and presents our approach in detail. The experiments and results analysis are demonstrated in Section 4. Lastly, conclusions and future work are discussed in Section 5.

2 RELATED WORK

In this section, we first review current approaches about visual fashion compatibility learning. Then we have a survey about fashion recommendation.

2.1 Visual Fashion Compatibility Learning

Visual compatibility measures whether fashion items complement one another across visual categories [1, 12, 20, 24]. For example, in Figure 1, a brown jacket is more compatible with a black casual pants. Oramas and Tuytelaars [15] introduced a hierarchical method to model visual compatibility by discovering mid-level visual elements. Li et al. [10] incorporated the appearances and metadata into their automatic composition system using an end-to-end deep neural network. Veit et al. [21] proposed a learning framework to recover a style space for fashion items from co-occurrence information and category labels which is able to learns compatibility between items. Han et al. [4] jointly train a Bi-LSTM model and a visual-sematic embedding for fashion compatibility learning. All of these methods are designed to generate outfits in which each item belongs to different categories. However, on one hand, category information is not accessible in many situations. On the other hand, these methods require a prior knowledge about which two categories can be put together. In contrast to these approaches, our method learns the compatibility relationships without a strict constraint of categories.

2.2 Fashion Recommendation

As mentioned above, there are a few approaches for recommending fashion items. One of the most famous method is visual Bayesian personalized ranking (VBPR) which learns user visual preference from implicit feedback [5]. The preference predictor in VBPR can be formulated as follows:

$$\hat{r}_{u,i} = \alpha + \beta_u + \beta_i + \gamma_u^T \gamma_i + \theta_u^T \theta_i \tag{1}$$

where α is the global offset, β_u and β_i are user/item bias, and γ_u and γ_i are latent feature vectors describing user u and item i, respectively, θ_u and θ_i are visual feature vectors of user u and item i. The inner product $\gamma_u^T \gamma_i$ and $\theta_u^T \theta_i$ indicates the scores that user u assigned to item i in terms of latent aspect and visual aspect.

This work improves the performance of recommender system significantly. The reason is that it incorporates visual information into the preference prediction model. However, in this approach, the authors did not consider the compatibility relationship between fashion items.

There are also other approaches which concentrate on fashion recommendation [2, 7, 8, 18]. Hu et al. [6] proposed a functional tensor factorization method to recommend outfits to users by learning from the interactions between user and fashion items. Packer et al. [16] learnt an interpretable image representation and modelled the dynamics of personalized users' visual preference to generate clothing recommendation. Liu et al. [11] separated style and category information from the image representation using a multilayer

Table 1: Major Notations Used in This Paper

Notation	Description
$\mathcal{U}, \mathcal{I}$	user set, item set
$\mathcal{V}$	image set of items
R	implicit feedback matrix
X	the set of images in clothing collocation dataset
C	clothing collocation pair list
$\mathcal{P}_u, \mathcal{V}_u, T_u$	positive item set of user u in training/validation/test sets
f_i	image representation vector of image i
γ_u, γ_i	latent features of user u, item i
θ_u, θ_i	visual features of user u, item i
W_{enc}, W_{dec}	weights of encoders and decoders, respectively

perceptron. However, all these methods only considered visual information and co-occurrence information from the online shopping website. It cannot sufficient learn compatibility relationship knowledge. In this paper, we utilize an external expert fashion collection dataset which contains a lot of good matches and propose a domain adaptation method to transfer the knowledge from the external dataset to our recommendation scene.

3 NOTATIONS AND PROBLEM FORMULATION

We first explain the symbols used in the next paper. We will consider $\mathcal{U}$ be the set of all users and $\mathcal{I}$ the set of all items. For each item, we have a corresponding images $\mathcal{V}_i$ which can represent this product. $R \subseteq len(\mathcal{U}) \times len(\mathcal{V})$ is the implicit feedback matrix whose rows correspond to customers and whose columns correspond to products. This means that $R_{ui} = 1$ stands for user u has bought item i, and 0 otherwise. We also have a hand-crafted clothing collocation dataset $X = \{x_1, x_2, ..., x_n\}$ labelled by experts which contains a set of clothing images and a clothing collocation pair list $C = \{(x_i, x_j)|x_i, x_j \in X\}$. Note that despite in the item set $\mathcal{V}$ and clothing collocation set X are both contains clothing images, the images in this two datasets are differences and no overlapped items. Furthermore, we can incorporate additional information like category data of products or demographic data about customers. However, we just focus on sales and visual information which is very important in fashion recommendation. Table 1 lists the major notations used throughout this paper.

The fashion recommendation task with visual compatibility relationship to be solved in this paper is to provide a personalized ranking list to each user with the help of visual information. First, given a set of fashion items $X = \{x_1, x_2, ...\}$ and collocations using these items C, learning visual compatibility knowledge is to learn an embedding $\mathcal{F}$ where the distance between item i and j, $d(\mathcal{F}(x_i), \mathcal{F}(x_i))$, is as small as possible if $(x_i, x_j) \in C$. After that, with user interaction records and item images, we try to learn the user's preference towards collaborative information and visual information to generate a ranking list for each user.

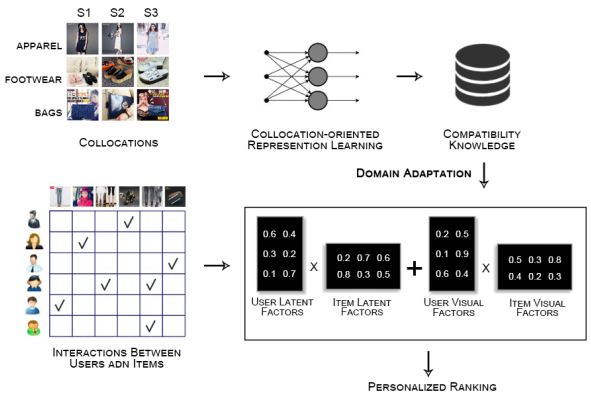


Figure 2: The proposed framework of our method.

4 ENHANCING FASHION RECOMMENDATION WITH VISUAL COMPATIBILITY RELATIONSHIP

We propose a fashion recommendation method which considers compatibility relationship between fashion items. In this method, we combine the collaborative information among users and items with compatibility knowledge. In order to allow the algorithm to understand the aesthetics of humans, we carefully construct an image representation model, through which we can determine what kind of information the resulting representation contains. This makes the computer learn the compatibility knowledge that people understand. After that, we incorporate the generated compatibility knowledge into our recommendation framework with a domain adaptation strategy. The framework of the entire recommender system is shown in Figure 2.

4.1 Learning visual compatibility knowledge from fashion items

In this section, our goal is to learn the visual compatibility relationships between fashion items. Conventional methods are mostly relying on category information to learn image representations. Instead of annotating images with labels or categories, which is costly, we leverage the weakly-labeled web data provided by the external dataset to learn compatibility knowledge.

Since there is no fixed category for the tasks we are going to perform, we cannot use the softmax-based cross entropy loss function for training. So, we chose triplet network to learn the image representation. The advantage of the triplet network is the distinction of details, that is, when the two inputs are similar, the triplet network can better model the details, which is equivalent to adding two measures of the difference of the input differences to learn a better representation of the input. The structure of the network is shown in the Figure 3.

In our task, we take the first item in an item pair in list C as an anchor, the second item as a positive sample, and select an item that is not in the list as the negative sample. More specifically, we can't randomly select negative samples on the entire candidate set, because this will cause $d(A, N)$ to be much larger than $d(A, P)$,

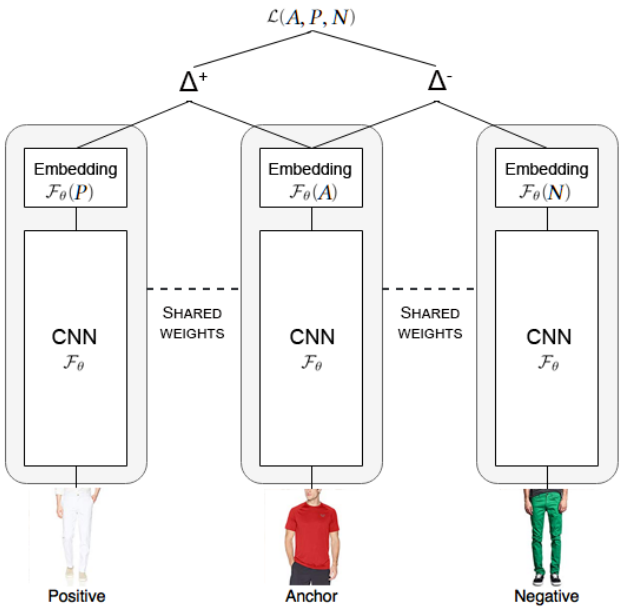


Figure 3: The illustration of our model for learning compatibility knowledge.

which will make the model unable to fully train and enter the prematurely. The state of the joint. So at each training, we need to choose a negative sample with $d(A, P)$ as close as possible to $d(A, N)$. This may make the model as difficult as possible to reduce the risk of overfitting.

Given a fashion image set, $X = \{x_1, x_2, ...\}$, x_i is a picture containing the t-th item. Also give a list $C = \{(x_i, x_j)|x_i \in X\}$ which denotes clothing collocation pairs. We need to learn a x_i to f_i mapping function, so that x_i and x_j are as close as possible, if the two items have good compatibility, and by contrast, the distance between f_i and f_j is as far as possible is i and j are not good compatibility. In other words, we try to learn a new representation of images. More formally, we minimize the following objective function:

$$\mathcal{L}(A, P, N) = max(||\mathcal{F}(A) - \mathcal{F}(P)||_2 - ||\mathcal{F}(A) - \mathcal{F}(N)||_2 + \alpha, 0) \quad (2)$$

where $\mathcal{F}$ is the mapping function. A is the anchor item, P is the item which have good compatibility with A, and N is not a good compatibility. α is the threshold parameter.

As shown in Figure 3, the network contains three sub-networks which shares weights with each other. In each sub-network, we encode the item image with Convolutional Neural Networks. There are many ConvNets architectures to choose from, and we use a variant of AlexNet for simplicity. The AlexNet variant, is the same as the original, except that we do not use pretrained weights and we replace local response normalization with batch normalization. We use the output of the fc6 layer as the encoding feature of the input image. The dimension of the image encoding is 4096.

Using more powerful architectures (e.g. [19]), may achieve better performance; however, we found that AlexNet is sufficient to show the effectiveness of our method. Because this network has fewer parameters, it can be trained more easily and reduces the risk of overfitting during training.

4.2 Fashion recommender system with visual compatibility knowledge

As mentioned above, the recommendation task can be regard as a ranking problem according to the user's preference. Our preference predictor is based on the basis of Matrix Factorization, which is the most promising model for rating prediction as well as modeling implicit feedback. The most related work to this problem is the VBPR model proposed in [5], which learns the visual user preference predictor using a pairwise ranking optimization framework.

We conduct our model based on pairwise learning. We defined the preference predictor as same as VBPR except for the reduce dimension method. To avoid missing information, we use an autoencoder to process dimension reducing. The formulation of encoder and decoder of autoencoder are as follows:

$$Enc(\mathcal{I}) = W_{enc} \cdot \mathcal{F}(\mathcal{I}) + b_{enc} \quad (3)$$

$$Dec(\mathcal{I}) = W_{dec} \cdot Enc(\mathcal{I}) + b_{dec} \quad (4)$$

Thus, the final preference predictor are as follows:

$$\hat{r}_{u,i} = \alpha + \beta_u + \beta_i + \gamma_u^T \gamma_i + \theta_u^T Enc(f_i) \quad (5)$$

For this implicit feedback ranking problem, we conduct the pairwise ranking optimization framework to train the model. The objective is as follows:

$$\max_\theta \sum_{(u,i,j) \in D_S} \ln \sigma\left(\hat{r}_{uij}\right) - \lambda_\Theta ||\Theta||^2 + ||f_i - \hat{f}_i||^2 + ||f_j - \hat{f}_j||^2 \quad (6)$$

where $\hat{r}_{uij} = \hat{r}_{u,i} - \hat{r}_{u,j}$.

Because the compatibility knowledge is learned from an external dataset also called source domain, images from the external dataset and the target dataset belong to different feature space. For example, in our experiments, the backgrounds of images in the source domain are very excursive. But in the target domain, the backgrounds are very clean and neat. Thus, we propose a domain adaption method which uses the domain adaptation technique to ensure that knowledge extracted from the source domain is consistent with the target domain and that knowledge transfer is positive.

When two items are bought by a customer at the same time, it regards as a co-occurrence pair. Most of the time, in the domain of fashion recommendation, we can assumption that the co-occurrence items should be a good clothing matching. Thus, we add the co-occurrence similarity into the objective above as follows to alleviate the distribution gap between source domain and target domain:

$$D_{i,j} = Frequent_{i,j} * ||Enc(f_i - f_j)|| \quad (7)$$

where $Frequent_{i,j}$ is the frequency of co-occurrence in the train set. When item i and item j are bought in a same bundle, we assume

that they are a good match, and the distance between them should be very closer.

In the training procedure, the training set D_S consists of triples in the form of (u, i, j, c), where u denotes the user and item i which they expressed positive feedback, and a non-observed item j. c is the co-occurrence frequency. It can be formalized by:

$$D_S = (u, i, j, c)|i \in I_u^+ \wedge j \in I_u^+ \tag{8}$$

The final formulation is as follows:

$$\sum_{(u,i,j)\in D_S} \ln \sigma\left(\hat{r}_{uij}\right) - \lambda_\Theta \|\Theta\|^2 + ||f_i - \hat{f}_i|| + ||f_j - \hat{f}_j|| + D_{i,j} \tag{9}$$

5 EXPERIMENTS AND ANALYSIS

We perform experiments on several datasets to evaluate the performance of the proposed method. All experiments were conducted on a workstation with a 6-core Intel CPU and two Titan-X (Pascal) graphics cards. Although there is a huge number of images and transaction records, it is still possible to train our model in half of day.

5.1 Datasets and evaluation metrics

The first dataset was provided by *Taobao.com* which is one of the most famous Chinese website for online shopping. It consists of clothing collocation suggestions from fashion experts, image data of Taobao items, and user behavior data. In this dataset, each line represents an item list which delimited by semicolon, every semicolon refers to a Collocation set. Every collocation set includes several goods, delimited by comma. We formatted this dataset into a pair-wise format, which means these two items is a good matching.

Another group of datasets contains user transaction records from two different sources. The first one were introduced in [5] and consist of reviews of clothing items crawled from *Amazon.com*. It was separated into 4 subcategories, named *Amazon Fashion*, *Amazon Women* and *Amazon Men*. The other one was crawled from *Tradesy.com*, which includes serveral kinds of feedback, like clicks, purchases, sales, etc.

The statistical information for the four datasets is provided in Table 2.

Table 2: Dataset statistics

Dataset	Users	Items	Interactions
Amazon Fashion	64,583	234,892	513,367
Amazon Women	97,678	347,591	827,678
Amazon Men	34,244	110,636	254,870
Tradesy.com	33,864	326,393	655,409

We measure recommendation performance of our method by calculating AUC and diversity. The AUC measures the quality of a ranking based on pairwise comparisons. Formally, we have

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u\in\mathcal{U}} \frac{1}{|\mathcal{D}_u|} \sum_{(i,j)\in\mathcal{D}_u} \xi\left(r_{u,i} > r_{u,j}\right) \tag{10}$$

where $\mathcal{D}_u = \{(i,j)|(u,i) \in \mathcal{T}_u \wedge (u,j) \notin (\mathcal{P}_u \cup \mathcal{V}_u \cup T_u)\}$. In other words, we are calculating the fraction of times that the 'observed' items i are preferred over 'non-observed' item j.

AUC is a typical metric to assess the recommendation performance in reproducing known user opinions that have been removed from the test dataset. The risk of such a metric is that, with recommendations based on similarity and overlap, customers will be exposed to a narrowing band of popular commodities. In other words, we also need other metrics to evaluate the recommendation performance.

Personalization, also named Inter-user diversity, considers the uniqueness of different customers' recommendation list. Given two users i and j, the different between their recommendation lists can be measured by the inter-list distance,

$$h_{ij}(L) = 1 - \frac{q_{ij}(L)}{L} \tag{11}$$

where $q_{ij}(L)$ is the number of common items in the top L places of the both lists: if the two lists are identical, $q_{ij}(L) = 0$ will equals 0 whereas completely different lists have $q_{ij}(L) = 1$.

Averaging $h_{ij}(L)$ over all pairs of users we obtain the mean distance $h(L)$, for which greater or lesser values mean, respectively, greater or lesser personalization of users' recommendation lists.

5.2 Experimental settings and baselines

We compared the proposed method in terms of accuracy and diversity against the following baselines:

- PopRank: Always recommends the top-k most popular items to users.
- BPR-MF (2009): This is a content-free algorithm based on matrix factorization which is designed for top-k recommendation tasks [17]. It optimizes pair-wise preferences between observed and unobserved items.
- VBPR (2016): It is a state-of-the-art image-based recommender system proposed by [5] for implicit feedback. The authors incorporate visual information provided by a pre-trained CNN.
- DVBPR (2017): This is the extension of VBPR by learning 'fashion aware' image representations directly [9].
- CO-BPR: Our method proposed in this paper.

We carefully choose the hyper-parameters and tuned them via grid search for each baseline method. For BPR-MF, VBPR, DVBPR and CO-BPR, we used a mini-batch size of 32 for all experiments. The number of latent factor selected from {6, 8, 10, 12, 14, 16, 18, 20}. We set it to be 12 in all experiments.

5.3 Results

We evaluated our proposed method by comparing it to state-of-the-art methods using some real-world datasets. We report recommendation performance in terms of the AUC and diversity in Table 3. Data with three sparsity ratios in target domain are chosen as training set. Comparing all the methods on these four datasets, we make the following observations:

Table 3: Recommendation Performance in Terms of AUC and Diversity with different sparsity

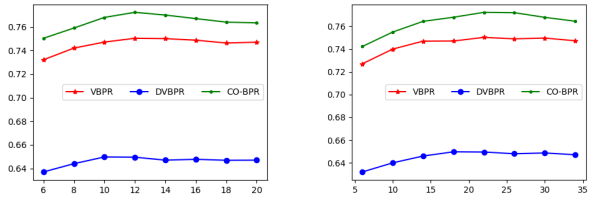
		AUC			Diversity		
		D=0.0010	D = 0.0005	D = 0.0001	D=0.0010	D = 0.0005	D = 0.0001
Amazon Fashion	*POPRANK*	0.5553	0.5627	0.6298	0.0000	0.0000	0.0000
	BPR-MF	0.5866	0.5951	0.6163	0.5621	0.5285	0.3093
	VBPR	0.6953	0.7116	0.7503	0.9715	0.9861	**0.9945**
	DVBPR	0.6134	0.6199	0.6497	**0.9826**	**0.9920**	0.9856
	CO-BPR	**0.7126**	**0.7242**	**0.7723**	0.9806	0.9891	0.9926
Amazon Women	*POPRANK*	0.5534	0.5897	0.6426	0.0000	0.0000	0.0000
	BPR-MF	0.5884	0.6176	0.6437	0.6335	0.5548	0.3786
	VBPR	0.6747	0.6861	0.7161	0.9777	0.9868	**0.9935**
	DVBPR	0.6209	0.6305	0.6714	0.9859	0.9890	0.9871
	CO-BPR	**0.6792**	**0.6901**	**0.7295**	**0.9952**	**0.9921**	0.9898
Amazon Men	*POPRANK*	0.5607	0.6118	0.6538	0.0000	0.0000	0.0000
	BPR-MF	0.5969	0.6269	0.6447	0.5655	0.4583	0.3501
	VBPR	0.6754	0.6857	0.7164	0.9716	0.9760	0.9870
	DVBPR	0.6270	0.6471	0.6726	0.9796	0.9788	0.9797
	CO-BPR	**0.6815**	**0.6982**	**0.7358**	**0.9836**	**0.9902**	**0.9961**
Tradesy.com	*POPRANK*	0.4105	0.3939	0.4756	0.0000	0.0000	0.0000
	BPR-MF	0.5830	0.5763	0.5317	0.8522	0.8813	0.8694
	VBPR	0.6553	0.6819	0.6927	**0.9923**	**0.9964**	**0.9931**
	DVBPR	0.6134	0.6199	0.6497	0.9826	0.9920	0.9856
	CO-BPR	**0.6718**	**0.6873**	**0.7106**	0.9874	0.9901	0.9920

1) Compared with POPRANK method on Amazon datasets, the value of AUC increase with the increase of sparsity ratio. It tells us that customer more likely to purchase popular items on Amazon. However, on Tradesy.com, customers prefer to choose unpopular items.

2) Compared with methods without visual information, we can see that visual information substantially improve recommendation accuracy and diversity.

3) Compared with VBPR and DVBPR, we can see that learning visual compatibility relationship from an external dataset is very effective. Our proposed method CO-BPR outperforms all the comparison methods on all the four datasets. This demonstrates the significant benefits of generating recommendations with visual compatibility relationship.

Furthermore, to investigate the dimensionality sensitivity, we illustrate the performance of VBPR, DVBPR and CO-BPR with varying dimensionality in Figure 4. It is clear that both latent dimensionality and visual dimensionality are note very sensitive.



(a) latent dimensionality (b) visual dimensionality

Figure 4: Performance of VBPR, DVBPR and CO-BPR with varying dimensionality measured by AUC.

6 CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a novel method for the fashion recommendation task with learning compatibility knowledge in visual aspect. A triplet network is used to learn compatibility knowledge from an external dataset. Domain adaptation strategy is used to alleviate the distribution gap between source domain and target domain. The experimental results show that our method is superior to several baselines in the AUC and diversity indicators.

In the future work, we intend to combine the learned image representation with the human understandable semantics to improve the interpretability of the model. In addition, we will use the image segmentation method to learn collocation knowledge from unmarked street photos on the web, thus reducing the reliance on manual labeling datasets.

ACKNOWLEDGMENTS

This research was supported by National Key R&D Program of China (No.2016YFB0801100), Beijing Natural Science Foundation (No.4172054, No.L181010), and National Basic Research Program of China (No.2013CB329605). This work was also supported by the Australian Research Council (ARC) under Grant [DP170101632].

REFERENCES

[1] Anurag Bhardwaj, Vignesh Jagadeesh, Wei Di, Robinson Piramuthu, and Elizabeth Churchill. 2014. Enhancing visual fashion recommendations with users in the loop. *arXiv preprint arXiv:1405.4013* (2014).

[2] Peter Gaspar. 2017. User preferences analysis using visual stimuli. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 436–440.

[3] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huiming Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, et al. 2017. Image matters: Jointly train advertising CTR model with image representation of ad and user behavior. *arXiv preprint arXiv:1711.06505* (2017).

[4] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S Davis. 2017. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 1078–1086.

[5] Ruining He and Julian McAuley. 2016. VBPR: Visual bayesian personalized ranking from implicit feedback.. In *AAAI*. 144–150.

[6] Yang Hu, Xi Yi, and Larry S Davis. 2015. Collaborative fashion recommendation: A functional tensor factorization approach. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 129–138.

[7] Vignesh Jagadeesh, Robinson Piramuthu, Anurag Bhardwaj, Wei Di, and Neel Sundaresan. 2014. Large scale visual recommendations from street fashion images. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 1925–1934.

[8] Shatha Jaradat. 2017. Deep cross-domain fashion recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 407–410.

[9] Wang-Cheng Kang, Chen Fang, Zhaowen Wang, and Julian McAuley. 2017. Visually-aware fashion recommendation and design with generative image models. In *Data Mining (ICDM), 2017 IEEE International Conference on*. IEEE, 207–216.

[10] 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* 19, 8 (2017), 1946–1955.

[11] Qiang Liu, Shu Wu, and Liang Wang. 2017. DeepStyle: Learning user preferences for visual recommendation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 841–844.

[12] 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 20th ACM international conference on Multimedia*. ACM, 619–628.

[13] Corey Lynch, Kamelia Aryafar, and Josh Attenberg. 2016. Images don't lie: Transferring deep visual semantic features to large-Scale multimodal learning to rank. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 541–548.

[14] Mingsong Mao, Jie Lu, Jialin Han, and Guangquan Zhang. 2019. Multiobjective e-commerce recommendations based on hypergraph ranking. *Information Sciences* 471 (2019), 269–287.

[15] Jose Oramas and Tinne Tuytelaars. 2016. Modeling visual compatibility through hierarchical mid-level elements. *arXiv preprint arXiv:1604.00036* (2016).

[16] Charles Packer, Julian McAuley, and Arnau Ramisa. 2018. Visually-aware personalized recommendation using interpretable image representations. *arXiv preprint arXiv:1806.09820* (2018).

[17] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. 2009. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 452–461.

[18] Devashish Shankar, Sujay Narumanchi, HA Ananya, Pramod Kompalli, and Krishnendu Chaudhury. 2017. Deep learning based large scale visual recommendation and search for e-commerce. *arXiv preprint arXiv:1703.02344* (2017).

[19] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning.. In *AAAI*, Vol. 4. 12.

[20] Flavian Vasile, Elena Smirnova, and Alexis Conneau. 2016. Meta-prod2vec: Product embeddings using side-information for recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 225–232.

[21] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*. 4642–4650.

[22] Wei Wang, Guangquan Zhang, and Jie Lu. 2017. Hierarchy visualization for group recommender systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 99 (2017), 1–12.

[23] Dianshuang Wu, Jie Lu, and Guangquan Zhang. 2015. A fuzzy tree matching-based personalized e-learning recommender system. *IEEE Transactions on Fuzzy Systems* 23, 6 (2015), 2412–2426.

[24] Qian Zhang, Dianshuang Wu, Jie Lu, Feng Liu, and Guangquan Zhang. 2017. A cross-domain recommender system with consistent information transfer. *Decision Support Systems* 104 (2017), 49–63.