IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 26, 2018, accepted January 4, 2019, date of publication January 14, 2019, date of current version February 8, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2892984

Learning to Style-Aware Bayesian Personalized Ranking for Visual Recommendation

MING HE, SHAOZONG ZHANG, AND QIAN MENG
Faulty of Information Technology, Beijing University of Technology, Beijing 100124, China

Corresponding author: Ming He (heming@bjut.edu.cn)

This work was supported in part by the Beijing Natural Science Foundation under Grant 4192008, and in part by the General Project of Beijing Municipal Education Commission under Grant KM201710005023.

ABSTRACT Recently, product images have been gaining the attention of recommender system researchers in the field of visual recommendation. This is because the visual appearance of products has a significant impact on consumers' decisions. Extensive studies have been done to integrate the features extracted by convolutional neural networks directly into recommendations. This improves the performance of recommender systems. Style features, an important type of features, are rarely considered. Style features play a vital role in the visual recommendation as a user's decision depends largely on whether the product fits his/her style. However, the representation of the conventional image features fails in capturing the styles of a product. To bridge this gap, we propose introducing style feature modeling, which is highly relevant with user preference, into the visual recommendation model. Furthermore, we propose incorporating the style features into collaborative learning to create awareness pertaining to the preferences of users. The experiments conducted on two public implicit feedback datasets demonstrate the effectiveness of our approach for the visual recommendation.

INDEX TERMS Personalized ranking, recommender systems, deep learning, visual recommendation.

I. INTRODUCTION

While purchasing clothing on the Internet, we usually look through product images before making decisions. In addition, vendors also provide customers with a lot of striking images of their products. References [6] and [22] also show the influence of images in purchasing. Appearance of a product plays a very important role in a Visual Recommendation task.

Due to the fact that a clothing products' image and appearance usually contains useful information, researchers in recommender systems started focusing on visual recommendation systems. As a result, recommender systems started integrating image and appearance information into recommendation processes. There has been extensive research work done in integrating visual features into recommender systems. For example, Visual Bayesian Personalized Ranking (VBPR) [14] incorporates visual features into original Bayesian Personalized Ranking (BPR) and enhances the performance in an implicit feedback situation. Liu *et al.* [23] adopted neural modeling based on visual features to model the style of items.

Style is a vital factor in shopping. For a young female who is going to purchase a skirt, what concerns she is not only "Is the skirt good-like?", but also "Does this colorful skirt match my style?" and "Does this skirt with a linear-designed texture suit my taste?". She will purchase it only if she is satisfied with all her concerns. According to the images of the clothing, we can acquire a lot of style information, such as color schemes, pattern designs, texture designs, and clothing fabric.

Capturing the style information from clothing is a challenging task. Conventional methods, utilized the convolutional neural network features (CNN features) which was extracted directly from the Caffe reference model [17] may not fully discover the latent style-level connections among the items and may not have the ability to extract different styles of items. To address this drawback regarding conventional CNN features, we extract the CNN features from the Caffe reference model [17], [20] and cluster items in the Tradesy dataset [14] in Figure 1.

Intuitively, every category of products is a assigned to corresponding cluster. Despite the CNN features are well-suited for category clustering, products with different styles (e.g., formal, casual and magnificent) cannot be distinguished. However, during shopping, people usually buy clothing with similar styles, and these clothes may not be similar in visual feature space. It poses challenges to the conventional visual recommendation system. Therefore, it is necessary to capture expressive style characteristics to explore taste of a user.

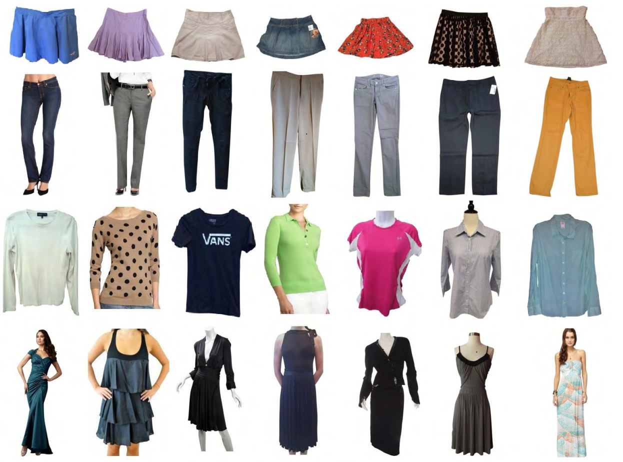


FIGURE 1. Part of the clustering results of items in the clothing subset of the Tradesy dataset [14], measured by the CNN visual features [17], [20]. Each row is a sample of a cluster.

Some researchers such as [23] predicting style features for visual recommendation. In [23], style is defined as conventional CNN features subtracted from categorical information; although, the distinction between style and variety is taken into account, it might be difficult to capture in-depth style information of item images using the model.

In this work, we propose a novel method to incorporate style information into recommender systems. Specifically, we propose to utilize a hierarchical gram matrix based on convolutional neural network, capturing the style of image between different feature spaces, which effectively solves the problem of style extraction and representation. Then, interaction of style feature and user latent factor can help us learn users' style preferences. The main contributions of this work are as follows:

1) To capture style information of items, we employ the hierarchical gram matrix which includes the correlation between different feature spaces. To the best of our knowledge, this is the first work to incorporate the representation of style into the recommender systems. Moreover, we compare the effect with conventional features to demonstrate the necessity of the style features.
2) In order to model users' preferences at style-level, we propose the SBPR (Style-aware Bayesian Personalized Ranking) model to explore the correlation between a user and an item. It can utilize style information in predicting a user's preferences and improving the performance in recommender systems.
3) We perform extensive experiments on two public implicit feedback data to demonstrate the effectiveness and rationality of our SBPR method.

II. RELATED WORK

A. SIDE INFORMATION IN RECOMMENDER SYSTEMS

Recently, we have seen increasing efforts devoted to recommendation models based on additional information, additional information related to users and items including texts and images can effectively improve the recommender system, including but not limited to social connections [15], [25], content [7], [26], and so on.

B. DEEP NEURAL NETWORK AND DEEP LEARNING-BASED RECOMMENDER SYSTEMS

Recently, the revolutionary advances of deep learning in recommender systems have drawn significant attention [36], including but not limited to the textual representations model using CNN [34]. Our proposed method is also a form of deep model using CNNs. Reference [2] applying MLP in YouTube recommendation, which divides the recommendation task into the generation module and ranking module.

C. VISUAL-BASED RECOMMENDER SYSTEMS

Visual features have received significant attention in recent works, with some methods using metrics for visual similarity according to social behavior or activity pattern to identify compatible items [24], [27] and visually improved recommendation [14]. Furthermore, [3] and [13] utilized the features extracted by a deep convolutional neural work (CNN features). Reference [37] leveraged the aesthetic network to extract relevant features and incorporated into the recommendation model, which improved the model performance. Reference [9] adopted image features for recommendations in a social network setting. Reference [35] introduced image features into the POI (point-of-interest) recommendation and proposed a graphical framework to model visual content in the context of POI recommendation. Reference [5] integrated the product images and item descriptions together to make a dynamic Top-N recommendation. Reference [29] designed a unified neural model to build a large-scale visual recommendation system for e-commerce. Reference [4] introduced the attention mechanism into CF to model the item-level as well as the component-level implicit feedback for multiple structural informational recommendations.

D. STYLE AND FASHION IN RECOMMENDER SYSTEMS

Beyond the above methods, style and fashion in item also become a popular task in computer vision [1], [10], [21]. For example, [18] and [33] pursued supervised method to classify people into certain style categories. Reference [30] utilized weak supervision from metadata to learn a latent representation for different style. In contrary to the above works, we apply an unsupervised method for capturing the taste preferences from visual implicit feedbacks.

III. STYLE-AWARE BPR

In this section, we propose our Style-aware Bayesian Personalized Ranking (SBPR) framework. Our model consists of the style feature extraction module to capture image features in different visual spaces, and a collaborative learning component to sense and predict a user's style preference. In this paper, we denoted a set of users as U and a sets of items as I,

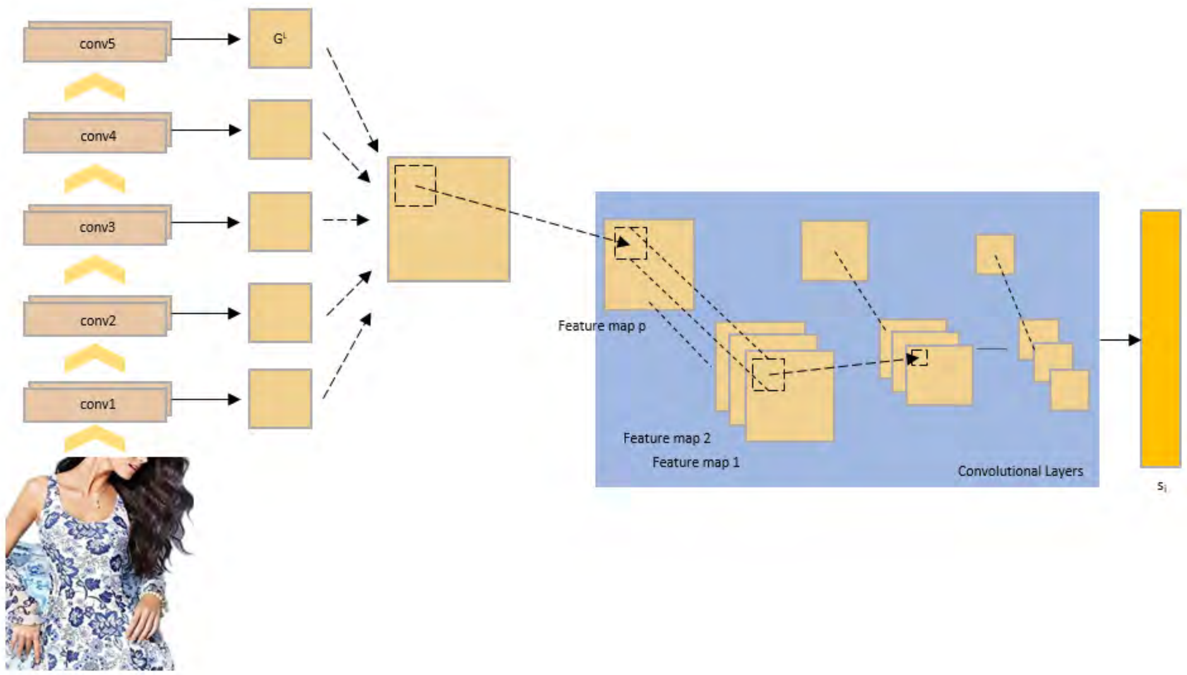


FIGURE 2. Diagrammatic representation of our style feature space modeling structure.

where $|U| = M$ and $|I| = N$. I^u denoted the collection of items that the user associated with in ground truth.

A. STYLE FEATURE SPACE MODELING

To obtain style features, we divide style feature extraction into two components. First, we feed the images of items into a pre-trained convolutional neural network model (CNN) and obtain the correlations between various filter responses in each layer of this network [8]. The style feature space modeling structure is shown in Figure 2. We utilize Gram matrix G^l that represents the feature correlations in certain layer l, where G^l_{fg} is the inner product between the feature maps f and g in the corresponding layer:

$$G^l_{fg} = \sum_k F^l_{fk} F^l_{gk}. \tag{1}$$

Particularly, we feed the corresponding item image into a deep convolutional neural network model. The CNN model applied here is widely used VGG nets [21], which did exceptionally well in ILSVRC14. We used 16 convolutional and 5 pooling layers of VGG-19 network, same as [8], then modified the max pooling layer to the average pooling layer.

1) FEATURE CORRELATIONS

From the perspective of the receptive field, the low-level convolutional neural network contains detailed information, such as the texture and edge of the image. As the network structure becomes deeper, the corresponding receptive field also increases. The high-level convolutional neural network contains the structure of the image. Generally speaking, style features should contain both detailed and structural information. Here we combine the low-level layer's feature correlations and high-level layer's feature correlations as style feature correlations of an item. For item i, we employ weighted responses G_i from various depths' convolutional layers to contain feature correlations:

$$G_i = \sum_k w_k G^k. \tag{2}$$

2) STYLE FEATURE

In order to incorporate the style feature correlations G_i into the collaborative learning model, we conducted experiments in selection of Multi-Layer Perceptron (MLP) and CNN stacks layers. Despite MLP is capable in feature representation theoretically [16], the most important drawback of having large number of parameters cannot be ignored. To address this limitation of MLP, we propose to utilize CNN stacks layers to learn correlations among the embedding dimensions. In rest of the extraction model 2, we design a CNN stacks network to capture style factors of item i. We denoted a feature map c in hidden layer l as a 2D tensor F^{lm}, and all feature maps in layer l as a 3D tensor E^l.

Given the input style feature correlations $G_i \in \mathbb{R}^{d \times d}$ and convolution filter K, we set the stride to 2. We can calculate the feature map as follows:

$$E^l = [m^l_{i,j,n}]_{s \times s \times p}, \tag{3}$$

$$m^l_{i,j,n} = f(\sum_{a=0}^{1}\sum_{b=0}^{1} m_{2i+a,2j+b} \cdot k^l_{l-a,l-b,n} + b_l), \tag{4}$$

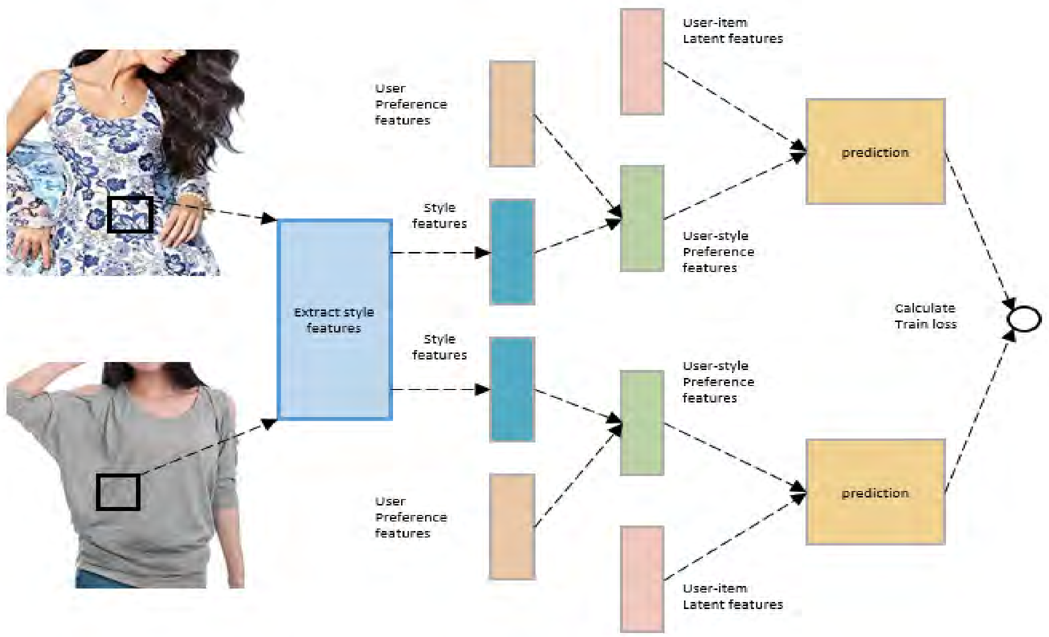


FIGURE 3. Diagrammatic representation of our style-aware BPR (SBPR) model.

where s denotes the input dimensions for Layer l, p denotes the number of feature maps for Layer l, b_l denotes the bias term for layer l, and f denotes nonlinear activation function. Stacks of convolutional layers can be abstracted as $s_i = g_\theta(G_i)$, where g_θ denotes the model of layers containing parameters θ, and $s_i(s_i \in \mathbb{R}^K)$ is the output of last layer, which can be seen as the style vector of item i.

B. COLLABORATIVE LEARNING

Since we got the style features of item i (i.e. s_i), our target is to predict the recommendation score y_{ui}. The collaborative learning structure is shown in Figure 3.

1) INPUT AND EMBEDDING LAYER

Given a user u and item i and its image, we first apply one-hot encoding based on their features. Then we obtain their embeddings by two embedding matrices for inputs of user and item:

$$W_u = R^T v_u^U, \quad W_i = Q^T v_i^I, \; P_u = O^T v_u^U, \tag{5}$$

where $v_u^U \in \mathbb{R}^M$ and $v_i^I \in \mathbb{R}^N$ are the feature encoding for user u and item i. M and N represent the number of users and items respectively, and $R \in \mathbb{R}^{M\times K}$ and $Q \in \mathbb{R}^{N\times K}$ are the embeddings for user and item features respectively. K represents the embedding size. To separate the user's general preferences and user's style preference, we employ P_u to encodes the user's style preference for item i, where $P_u \in \mathbb{R}^K$ denotes the style preference embedding of user u.

2) PREDICTION LAYER

We can predict the recommendation score y_{ui} as

$$\hat{y}_{u,i} = f_1(W_1 \begin{bmatrix} W_u^T \\ W_i \end{bmatrix} + b_1) + f_2(W_2 \begin{bmatrix} P_u \\ s_i^T \end{bmatrix} + b_2), \tag{6}$$

where W_1, W_2 and b_1, b_2 denote the weight matrices and bias vectors, respectively. $f_1(\cdot), f_2(\cdot)$ is set to ReLU function. $P_u \in \mathbb{R}^K$ denote the preference embedding of user u. The score $\hat{y}_{ui}$ characterizes the preference of u to i. In general, the parameters of our model are $\Delta = \{R, Q, P, \theta\}$.

3) OBJECTIVE FUNCTION

Because the model is a personalized ranking task, we consider learning parameters of SBPR with pair-wise ranking. Reference [12] proposes the usage of point-wise classification, which fails to learn models from implicit feedback. However, Bayesian Personalized Ranking (BPR) [28] is a pair-wise ranking optimization framework, which has an effective hypothesis that observed samples should be ranked higher than the unobserved samples. Thus, we represent the training set D as:

$$D \in \{(u, i, j) | u \in U \wedge i \in I^u \wedge j \in I \setminus I^u\}. \tag{7}$$

Furthermore, for a user u, given a non-observed (or negative) item j, we denote the preference score from the mirror structure in Figure 3 as y_{uj}. Ultimately, the model prediction is $\hat{y}_{ui} - \hat{y}_{uj}$. We use the following objective function:

$$J = \sum_{(u,i,j)\in D} \ln \sigma(\hat{y}_{u,i} + \gamma - \hat{y}_{u,j}) - \lambda \parallel \Delta \parallel^2, \tag{8}$$

where D denotes the training set, and λ are the hyper-parameters to prevent overfitting in training. γ is a margin that separates the positive and negative item pairs. It is crucial to impose L2 loss on the model parameter Δ. Moreover, in sparse datasets, we found that dropout can be applied for nonlinear layers to alleviate overfitting.

IV. EXPERIMENT

In this section, we evaluate the proposed SBPR model with existing recommendation methods on widely used real-world datasets. In brief, we aim to answer the following questions via experiments.

RQ1: How does our model perform compared with the state-of-art visual recommender systems?

RQ2: How do the key hyper-parameters affect the performance? Take the latent dimensions for example.

RQ3: How to interpret the advanced performance visually?

A. EXPERIMENTAL SETUP

Our experiments are conducted on two public implicit feedback datasets.

Amazon The Amazon dataset is a purchase datasets that contains a large number of consumption records [27]. We used the clothing shoes and jewelry category to train the model.

Tradesy The Tradesy dataset is from Tradesy.com [13], a second-hand clothing trading website and contains purchase histories of its consumers.

We perform a popular filtering method similar to [11] and [32] to remove users with less than five purchase records in Tradesy dataset. For all dataset, we randomly split records by 8:1:1 for training, validation and test. Furthermore, the cold-start issue will be considered in Tradesy datasets. We process each dataset by extracting implicit feedback as previously described. Table 1 shows statistics of our datasets.

TABLE 1. Post-processed datasets statistics.

Dataset	users	items	feedback
Amazon.com	*39,387*	*23,033*	*278,677*
Tradsy.com	*14,776*	*112,093*	*268,091*
Total	*54,163*	*135,126*	*546,768*

We implemented our model with Tensorflow.[1] For all baseline models, hyper-parameters are tuned using grid-searching with a validation set. We optimize our model with Adaptive Moment Estimation (Adam) [19] with a mini-batch size of 512, which promotes dynamic adjustment of learning parameters. The learning rate of SBPR is set to be 0.005 and use a L-2 regularization with $\lambda = 0.005$. We model the layer "conv4_1" and "conv5_1", where weight of the gram matrix is set to be (0.5, 0.5). Following some previous work [14], [28], we use Area Under the ROC Curve (AUC) to evaluate methods, which can measures the probability of an observed item ranking higher than a unobserved one. There are two types of evaluation settings during the test process: warm-start and cold-start.

In every 5 iteration of training, we enumerate all positive instances to optimize models and perform average AUC on the full test set T (denoted by 'warm-start') on both datasets, as well as a subset T' of T that only consists of items having less than five observed feedback correlations (denoted by "cold-start") on the Tradesy dataset. The items of T' account for around 85% for Tradesy.com. This means for such sparse real-world datasets, a model must make full use of interactive information to guarantee its ability of making a recommendation.

We compare our SBPR with classic models as well as recent state-of-art methods, such as:

1) **BPR-MF** [28]: A Bayesian Personalized Ranking Matrix Factorization method for solving implicit feedback task.
2) **VBPR** [14]: A state-of-the-art visual-based recommendation model[9]. Integrating visual features into BPR model to improve performance of recommender systems.
3) **DEEPSTYLE** [23]: A Bayesian Personalized Ranking method, which adopts neural modeling based on product images to model the style of items.

B. PERFORMANCE COMPARISON (RQ1)

1) MODEL COMPARISON

Table 2 reports the results comparison among SBPR, DEEPSTYLE, VBPR and BPR-MF under warm-start and cold-start settings.

In Figure 4, we can observe that, SBPR, DEEPSTYLE and VBPR achieve better performance than BPR-MF, which verifies that visual information can contribute to the performance in recommendation.

DEEPSTYLE performs better than VBPR. This observation highlights the importance of style information in personalized recommendations, and it is consistent with the intuition that a consumer's decision depends largely on whether the clothing is fit her style taste. This implies, that combining the style information in the model is helpful in sensing users' style preference.

The performance of DEEPSTYLE fails to surpass SBPR. This observation is not surprising because DEEPSTYLE fails in capturing the deep-level style features. Capturing the style of image between different feature spaces, SBPR can aware expressive style characteristics to sense preferences of users. These improvements forecast the advantage of SBPR in performance, which answers **RQ1**.

C. IMPACT OF DIMENSIONALITY (RQ2)

In order to answer RQ2, We evaluate the performance of SBPR, VBPR and BPR-MF using varying dimensionalities $K = [5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]$. Figure 5 shows how the embedding dimension influences

[1] https://www.tensorflow.org/

TABLE 2. AUC performance comparison on the test set T (K = 30).

Dataset	setting	BPR-MF	VBPR	DEEP STYLE	SBPR
Amazon.com	*warm-start*	*0.6804*	*0.7389*	*0.7467*	**0.7592**
Tradsy.com	*warm-start*	*0.6058*	*0.7608*	*0.7653*	**0.7767**
	cold-start	*0.5288*	*0.7436*	*0.7385*	**0.7598**

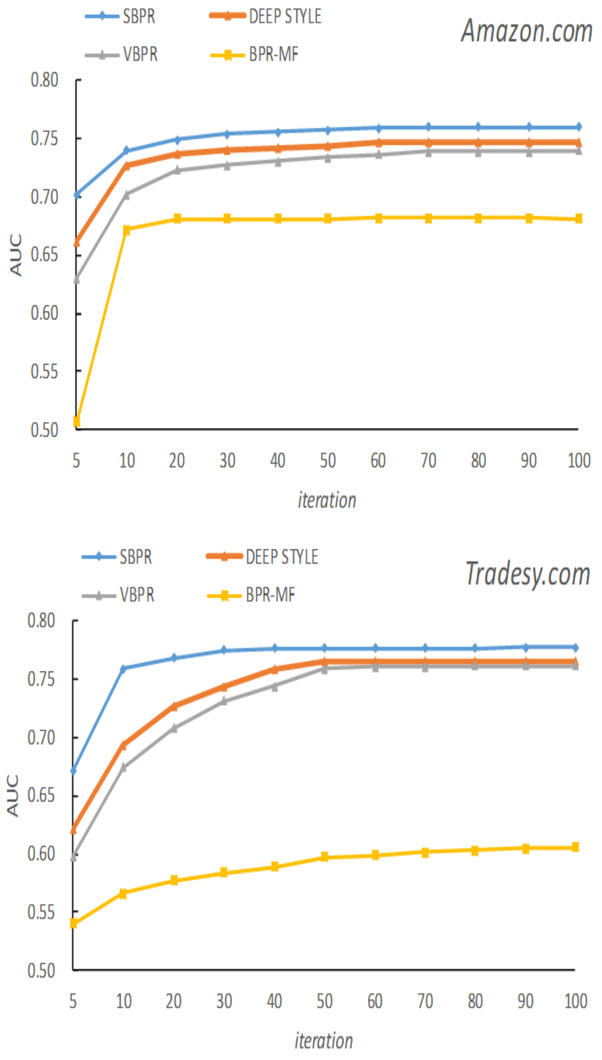


FIGURE 4. AUC performance comparison with training iterations (K = 30).

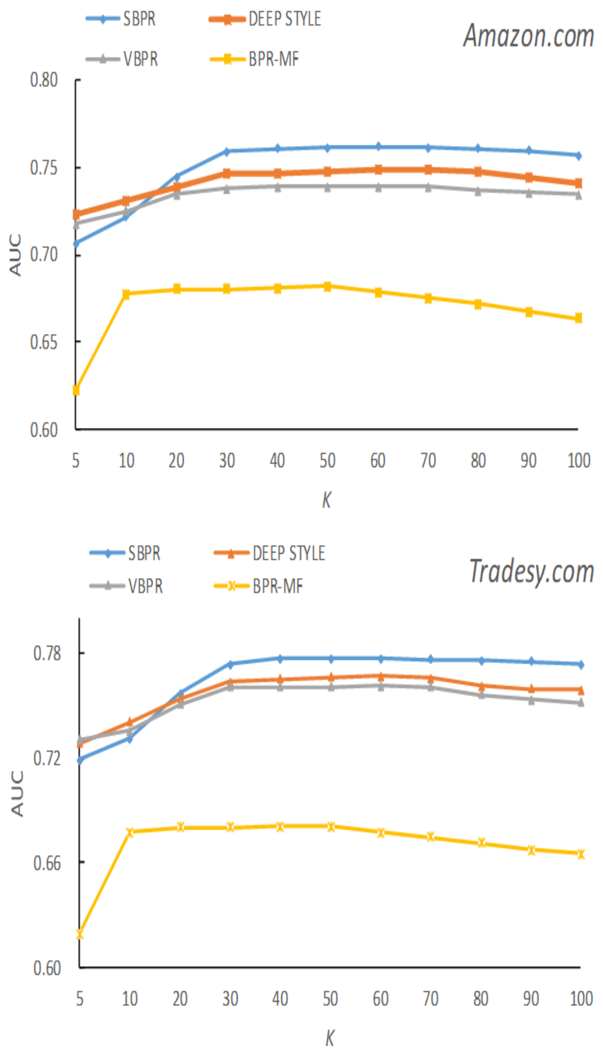


FIGURE 5. AUC with varying dimensions.

the model performance. We can observe, on the Amazon dataset, the performance in SBPR is not as good as in case of VBPR and DEEPSTYLE before the meeting points since style features need to be explored in a higher dimension. All the models exhibit their best performance while using certain dimensions. Similar results can also be observed on the Tradesy dataset. This observation suggests, while expressive ability is increased, using too many latent factors may also increase the model complexity extremely and may lead to over-fitting, and recede the generalization ability of our models on the test dataset. Generally, our method performs better than others.

D. VISUALIZATION (RQ3)

In this part, we visualize several purchased and recommended items in Figure 6. Items in the first row are purchased by certain users in the Tradesy dataset (sample, random sampling). To illustrate the effect of the style features intuitively, we choose the users with an explicit style preference. Items in the second row and third row are recommended by SBPR and DEEPSTYLE respectively. For these two rows, we choose five best items from 50 recommendations to exhibit. Comparing the first row and the second row, we can see that leveraging both semantic and style information, SBPR can recommend the congeneric commodities with similar style.

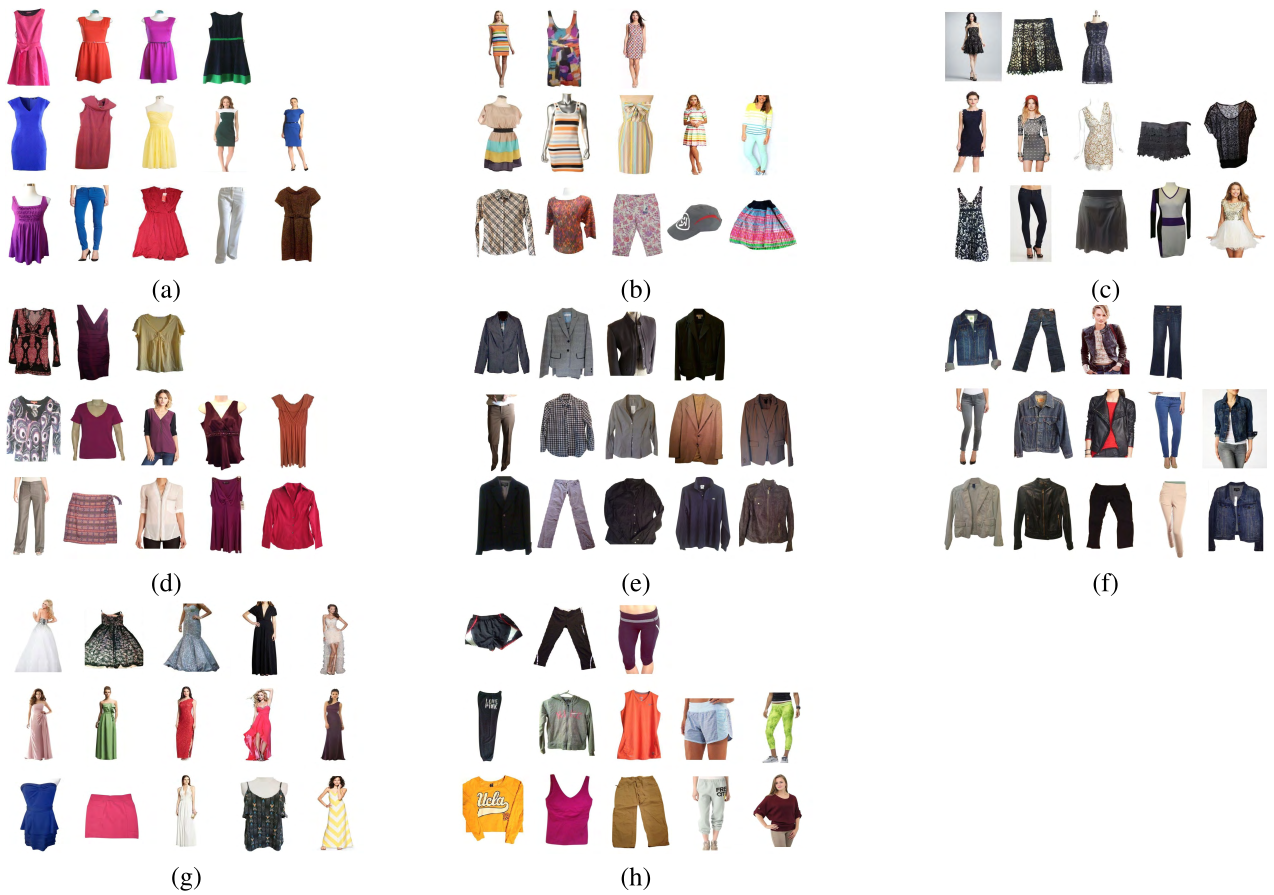


FIGURE 6. Items purchased by consumers and recommended by SBPR and DEEPSTYLE respectively.

Comparing the second row and the third row, we observe that although DEEPSTYLE can recommend different kinds and pertinent products, it did not performed well in style recommendation, especially in texture designs and clothing fabric. Referring to Figure 6 (b) as an example, we can see that what the user likes are closer to colorful style, and items in the second row have more similar styles with the samples than the items in the third row. Items in the second row have similar color schemes and pattern designs, such as colorful designs, linear-designed texture, irregular texture. Moreover, although there are only dresses in the sample, there are skirts and uppers in the recommendations. Note that, SBPR has the ability to recommend items in a similar style but not limited to a single category and shows scalability in multiple types in recommendations.

It is also obvious in Figure 6 (c) that the user in the sample likes background texture style, and the items recommended by SBPR are in the same style, such as slender proportions, exquisite-designed texture, and monotonous designed color. As we can see, our proposed model has the stronger ability to sense various styles of products automatically.

E. CONCLUSION

In this paper, we investigated style features in a personalized recommendation task via the style feature space modeling. To be specific, we propose Style-aware BPR (SBPR) - a novel method based on the personalized pairwise ranking model for implicit feedback, which incorporates style features extracted by convolutional neural network. In addition to the style features, we employed a style representation which includes correlation between different feature spaces. Experiments conducted on challenging public datasets show effectiveness of our proposed method, and success in understanding the style preferences of users.

REFERENCES

[1] Z. Al-Halah, R. Stiefelhagen, and K. Grauman. (2017). "Fashion forward: Forecasting visual style in fashion." [Online]. Available: https://arxiv.org/abs/1705.06394

[2] P. Covington, J. Adams, and E. Sargin, "Deep neural networks for YouTube recommendations," in *Proc. ACM Conf. Recommender Syst.*, 2016, pp. 191–198.

[3] X. Chen, Y. Zhang, Q. Ai, H. Xu, J. Yan, and Z. Qin, "Personalized key frame recommendation," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 315–324.

[4] J. Chen, H. Zhang, X. He, L. Nie, W. Liu, and T.-S. Chua, "Attentive collaborative filtering: Multimedia recommendation with item- and component-level attention," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 335–344.

[5] Q. Cui, S. Wu, Q. Liu, and L. Wang. (2016). "A visual and textual recurrent neural network for sequential prediction." [Online]. Available: https://arxiv.org/abs/1611.06668

[6] M. Dai, L. V. Hove, M. Dai, L. V. Hove, M. Dai, and L. V. Hove, "The impact of customer images on online purchase decisions: Evidence from a Chinese C2C Web site," *First Monday*, vol. 22, no. 10, 2017. [Online]. Available: https://uncommonculture.org/ojs/index.php/fm/article/view/7120/6545

[7] S. Dieleman and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 2643–2651.

[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 262–270.

[9] X. Geng, H. Zhang, J. Bian, and T.-S. Chua, "Learning image and user features for recommendation in social networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 4274–4282.

[10] X. Han, Z. Wu, Y. G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional LSTMs," in *Proc. ACM Multimedia Conf.*, 2017, pp. 1078–1086.

[11] R. He, W. C. Kang, and J. Mcauley, "Translation-based recommendation," in *Proc. 11th ACM Conf. Recommender Syst.*, 2017, pp. 161–169.

[12] X. He, L. Liao, H. Zhang, L. Nie, X. Hu, and T. S. Chua, "Neural collaborative filtering," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 173–182.

[13] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. WWW*, 2016, pp. 507–517.

[14] R. He and J. McAuley, "VBPR: Visual Bayesian personalized ranking from implicit feedback," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 144–150.

[15] M. Jamali and M. Ester, "A matrix factorization technique with trust propagation for recommendation in social networks," in *Proc. ACM Conf. Recommender Syst.*, 2010, pp. 135–142.

[16] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.*, vol. 4, no. 2, pp. 251–257, 1991.

[17] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. MM*, 2014, pp. 675–678.

[18] M. H. Kiapour, K. Yamaguchi, A. C. Berg, and T. L. Berg, "Hipster wars: Discovering elements of fashion styles," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 472–488.

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2014, pp. 1–15.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[21] Y. Li, L. Cao, J. Zhu, and J. Luo, "Mining fashion outfit composition using an end-to-end deep learning approach on set data," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1946–1955, Aug. 2017.

[22] X. Li, M. Wang, and Y. Chen, "The impact of product photo on online consumer purchase intention: An image-processing enabled empirical study," in *Proc. PACIS*, 2014, p. 325.

[23] Q. Liu, S. Wu, and L. Wang, "DeepStyle: Learning user preferences for visual recommendation," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2017, pp. 841–844.

[24] S. Liu, P. Cui, W. Zhu, S. Yang, and Q. Tian, "Social embedding image distance learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 617–626.

[25] H. Ma, D. Zhou, C. Liu, M. R. Lyu, and I. King, "Recommender systems with social regularization," in *Proc. WSDM*, 2011, pp. 287–296.

[26] A. Q. Macedo, L. B. Marinho, and R. L. T. Santos, "Context-aware event recommendation in event-based social networks," in *Proc. ACM Conf. Recommender Syst.*, 2015, pp. 123–130.

[27] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 43–52.

[28] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: Bayesian personalized ranking from implicit feedback," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2009, pp. 452–461.

[29] D. Shankar, S. Narumanchi, H. A. Ananya, P. Kompalli, and K. Chaudhury. (2017). "Deep learning based large scale visual recommendation and search for e-commerce." [Online]. Available: https://arxiv.org/abs/1703.02344

[30] E. Simoserra and H. Ishikawa, "Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 298–307.

[31] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, pp. 1–14.

[32] J. Tang and K. Wang, "Personalized top-n sequential recommendation via convolutional sequence embedding," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 565–573.

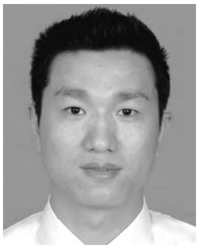
[33] A. Veit, B. Kovacs, S. Bell, J. Mcauley, K. Bala, and S. Belongie, "Learning visual clothing style with heterogeneous dyadic co-occurrences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4642–4650.

[34] X. Wang *et al.*, "Dynamic attention deep model for article recommendation by learning human editors' demonstration," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 2051–2059.

[35] S. Wang, Y. Wang, J. Tang, K. Shu, S. Ranganath, and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 391–400.

[36] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning for recommender systems," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 1235–1244.

[37] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. WWW*, 2018, pp. 649–658.



MING HE received the B.S. degree in computer and application and the M.S. degree in oil-gas well engineering from the Xi'an Petroleum Institute, Shaanxi, China, in 1999 and 2002, respectively, and the Ph.D. degree in computer science and technology from Xi'an Jiaotong University, Shaanxi, China, in 2005. From 2006 to 2011, he was a Research Assistant with the College of Computer Science, Beijing University of Technology, Beijing, China. Since 2012, he has been an Assistant Professor with the Faulty of Information Technology, Beijing University of Technology. He has authored more than 60 articles. His research interests include recommendation systems, data mining, and machine learning. He is a member of the International Association of Computer Science and Information Technology. His awards and honors include the Excellent teachers Award, in 2015, and the Beijing University of Technology Best Academic Paper Award, in 2010.

SHAOZONG ZHANG was born in Handan, Hebei, China, in 1994. He received the B.S. degree in network engineering from Heibei University, Hebei, in 2017. He is currently pursuing the master's degree with the Beijing University of Technology, under the supervision of Prof. M. He. His main research interests include deep learning and recommender systems.

QIAN MENG was born in Zhoukou, Henan, China, in 1996. He received the B.S. degree from the School of Computer Science and Technology, Qilu University of Technology, in 2017. He is currently pursuing the master's degree with the Beijing University of Technology, under the supervision of Prof. M. He. His main research interests include deep learning and recommender systems.