CCF Transactions on Pervasive Computing and Interaction (2019) 1:275–284
https://doi.org/10.1007/s42486-019-00017-y

REGULAR PAPER

Learning from visual content and style: an image-enhanced recommendation model

Suchang Luo[1] · **Lei Chen**[1] · **Le Wu**[1]

Received: 17 June 2019 / Accepted: 16 October 2019 / Published online: 6 November 2019

Abstract
Image based platforms are popular in recent years. With a large number of images in these image based platforms, how to properly recommend images that suit each user's interest is a key problem for recommender systems. While a simple idea is to adopt collaborative filtering for image recommendation, it does not fully utilize the visual information and suffers from the data sparsity issue. Recently, with the huge success of *C*onvolutional *N*eural *N*etworks (CNN) for image analysis, some researchers proposed to leverage image content information for recommendation. Specifically, Visual Bayesian Personalized Ranking (VBPR) (He and McAuley, in: The association for the advancement of artificial intelligence, 2016) is a state-of-the-art visual based recommendation model, which proposed to learn users' preferences to items from two spaces: a visual content space learned from CNNs, and a latent space learned from classical collaborative filtering models. VBPR and its variants showed better recommendation performance with image content modeling. In the real-world, when browsing visual images, users not only care the image content, but also concern the matching degree of the image style. Compared to image content, the role of visual styles has been largely ignored in the image recommendation community. Therefore, in this paper, we study the problem of learning both the visual content and style for image recommendation. We leverage advances in computer vision to learn the visual content and style representation, and propose to how to combine visual signals with users' collaborative data. Finally, experimental results on a real-world dataset clearly show the effectiveness of our proposed model.

1 Introduction

Visual signals are playing more and more important roles for users' daily life (Wu et al. 2019). Especially in modern fast-paced society, users are reluctant to spend much time to read literal content, but prefer to browser visual images (Gelli et al. 2018). To adapt to the modern society, more and more image based applications, such as *Instagram*, *Flickr* and *Pinterest*, have emerged. In the meantime, most of these platforms allow users to upload images to further increase the prosperity. With millions of images uploaded in the image platforms every day, how to design an image recommendation algorithm has become a key issue. Accurately, predicting users' visual preferences and making personalized recommendation could increase the users' satisfaction and loyalty, which is beneficial for both users and platforms.

As Collaborative Filtering (*CF*) is one of the most popular approaches for recommender systems (Koren 2008; Sarwar et al. 2001), a natural idea is to perform image recommendation with these CF models. For example, with user-image implicit feedbacks, Bayesian Personalized Ranking (*BPR*) (Rendle et al. 2009) is a pair-wise based ranking model for CF. BPR projects users and items in a latent space, and learns user and item latent representation with a ranking loss function from users' behavior. Though successful, these kind of models suffer from the data sparsity issue of user behavior. Recently, the development of convolutional neural networks has dominated the computer vision community, and showed state-of-the-art performance on many image classification tasks (Radenović et al. 2018; Ren et al. 2015; Simonyan and Zisserman 2015). Given the superior performance of CNNs for image content modeling, researchers proposed to leverage the image content from CNNs to alleviate the data sparsity in CF and enhance image recommendation. E.g., VBPR is one of the first few attempts that

✉ Le Wu
lewu.ustc@gmail.com

Suchang Luo
luosuchang.edu@gmail.com

Lei Chen
chenlei.hfut@gmail.com

[1] Hefei University of Technology, Hefei, China

combine BPR and image content learned from CNNs for image recommendation (He and McAuley 2016). Furthermore, some researchers also extended VBPR and combined additional data, such as the uploader information, the social network for image recommendation.

Besides the image content modeling for recommendation, image styles also play an important role for users' visual experience. When browsing images, users' preferences are not only decided by "what is the content of this information", but also "does the image style matches my taste" (Wu et al. 2019). While visual content features are empirically effective for image based recommendation, the visual styles are largely ignored in the image recommendation research community. Compared to image content, the image style is more abstract and hard to be represented. Therefore, how to learn visual content and visual style for images, and leverage them for image recommendation is quite challenging. To this end, in this paper, we propose to learn image content and style for better image recommendation performance. Specifically, as previous works, we use the last fully connected layer in CNNs for image content modeling. For image style representation, we propose to borrow the state-of-the-art computer vision community and use a feature space designed to capture the textual information from the computer vision community (Gatys et al. 2016, 2017). After that, we design a user preference learning function that considers users' latent preferences, the visual content and style preferences. To show the effectiveness of the proposed model, we conduct experiments on a real-world image recommendation dataset. The experimental results clearly show the effectiveness of our proposed model.

2 Related work

As one of the state-of-the-art personalized ranking methods, Bayesian Personalized Ranking (BPR) proposed by Rendle et al. is based on Latent Factor Model and adopts pairwise approach to make implicit feedback as relative preferences rather than absolute one. Specifically, a user u is assumed to prefer an item i to an item j. Thus, the user-item pair (u, i) is observed and (u, j) is non-observed (Rendle et al. 2009). It shows a good performance and has been a popular strong baseline in researches concerned. In the real-world applications, instead of the explicit action or inaction, the multimedia information of items is usually implicitly related to users' opinions (Bartolini et al. 2016; Canini et al. 2012; Chen et al. 2017). Specifically, in image recommendation, user preferences are mainly reflected by the vision feature of items (e.g., some users like science fiction images, others like landscape images).

In image recommendation, in order to utilize rich content information of images, researchers proposed some recommendation models by considering the rich context information (Chen et al. 2016; He and McAuley 2016). Typically, Visual Bayesian Personalized Ranking (VBPR) (He and McAuley 2016) is a breakthrough algorithm which is the first time to contain the visual information into BPR. It extracts image content features via pre-trained convolutional neural network and adopts embedding method to reduce the dimensions of visual item factors (Fan et al. 2008; Krizhevsky et al. 2012). But, VBPR not measures users' preference to image of item and it has weak interpret ability (Liu et al. 2017; Luo et al. 2018). Simultaneously, there are some other image recommendation models (Guo et al. 2019; Hsiao and Grauman 2017; Shankar et al. 2017) that trying to consider fine-grained features in order to enhance the accuracy and interpretability of VBPR. E.g., Guo et al. proposed a model that combining the feature embeddings of the fine-grained image objects, and considering the relative weights which may be distinct for different users. Also, some researchers showed that image feature with others rich information can better reflect users' personality. E.g., visual factors combined with user tags, geographic features can effectively improve the quality of image recommendations (Niu et al. 2018) Besides, in social contextual recommendation, social links, dialog history and images uploaded are the most important factors that should be considered (Jiang et al. 2014; Wang et al. 2017).

Usually, image recommendation develops with the pace of computer vision. Recently Gatys et al. proposed a new model of extracting image styles based on the feature maps of convolutional neural networks (Gatys et al. 2016). The proposed model showed high perceptual quality for extracting image style, and has been successfully applied to related tasks, such as image style transfer, and high-resolution image stylization (Gatys et al. 2015). In this way, we think that visual style is also a considerable character that cannot be ignored. Thus, we take visual content and visual style into account for image recommendation.

3 Visual embedding and problem formulation

Users' preference, especially in image oriented platforms, is changeable but analyzable, which can be decided by both non-visual factor and visual factor. The non-visual factor can be regarded as latent factor, which can be extracted via matrix factorization. As for visual factor, apart from the basic content of images, the style of images is also concluded as an important factor. In this section, we illustrate how to extract visual content feature and visual style feature. Then, we use two extracted features to model user's visual preferences. Finally, problem formulation is presented.

3.1 Visual embedding

3.1.1 Visual content embedding

Image always plays an essential role in social platforms, especially in image oriented platforms. Users show more interest in visual information than text information as image can directly show the information publisher's ideas and intentions. Totally, image content feature is extracted from convolutional neural network, which can reduce processing data volume, simultaneously, preserve visual information as more as possible.

In this paper, we apply VGG19 to help us extract visual content feature (Simonyan and Zisserman 2015), which is a common method to extract feature. VGG19 is capable of extracting visual content feature in highly efficiency and quality. According to the ability of CNN model for visual tasks (Donahue et al. 2014; Gatys et al. 2016; Tzelepi and Tefas 2018), the nearer the feature map to the output layer, the more information the factor representation contains. Thus, we use the f_i^c dimensional representation in the last connected layer in VGG19 as visual content feature vector representation of one image. In several popular models like VBPR, embedding is widely used to reduce dimensions. Therefore, we suggest an embedding matrix E multiplied by visual content feature matrix, which is capable of reducing dimension from a high dimension to a lower dimension in linearly transformation (He and McAuley 2016).

$$w_i^c = E_c f_i^c, \tag{1}$$

where E_c is a $C \times F_c$ embedding matrix (C is far less than the size of f_i^c, like 32 or so), f_i^c ($F_c - dimensional$) denotes the visual content feature vector obtained from the last connected layer in VGG19. After multiplication of embedding matrix and visual content feature vector, the visual content feature is extracted into C-dimension vector (Fig. 1).

3.1.2 Visual style embedding

In image oriented platforms, users no more keep eyes on image content only, besides, their preferences are also highly influenced by image style. In a real world, if one user is into Van Gogh style images, we have a very high level of confidence that this user is interested in similar style images uploaded by other publishers. In the same way, the one who likes uploading a sort of color assortment is fond of the same color assortment images uploaded by others. Therefore, visual style should be taken into consideration as one of attribution in social recommendation. In accordance to the ability of CNN model for visual tasks (Chu and Wu 2018; Donahue et al. 2014; Gatys et al. 2016), the feature map nearer to input layer contains more visual style feature information. In this paper, we choose a common practice of image style extraction (Gatys et al. 2016). This method has a powerful ability to perception and as an image style extraction method, it has been widely used in many image-style based researches.

This method is aimed to measure which features in the style layer l activate simultaneously for the style image. A Gram matrix is proposed for tensors output by style layers l. It calculates dot products for the vector of the feature F_{ik}^l, which denotes the j-th filter at position k of a style layer l.

$$G_{ij}^l = \sum_k F_{ik}^l F_{jk}^l, \tag{2}$$

where G is a Gram matrix, denoting the correlation between feature map i and j in style layer l. Researches show that

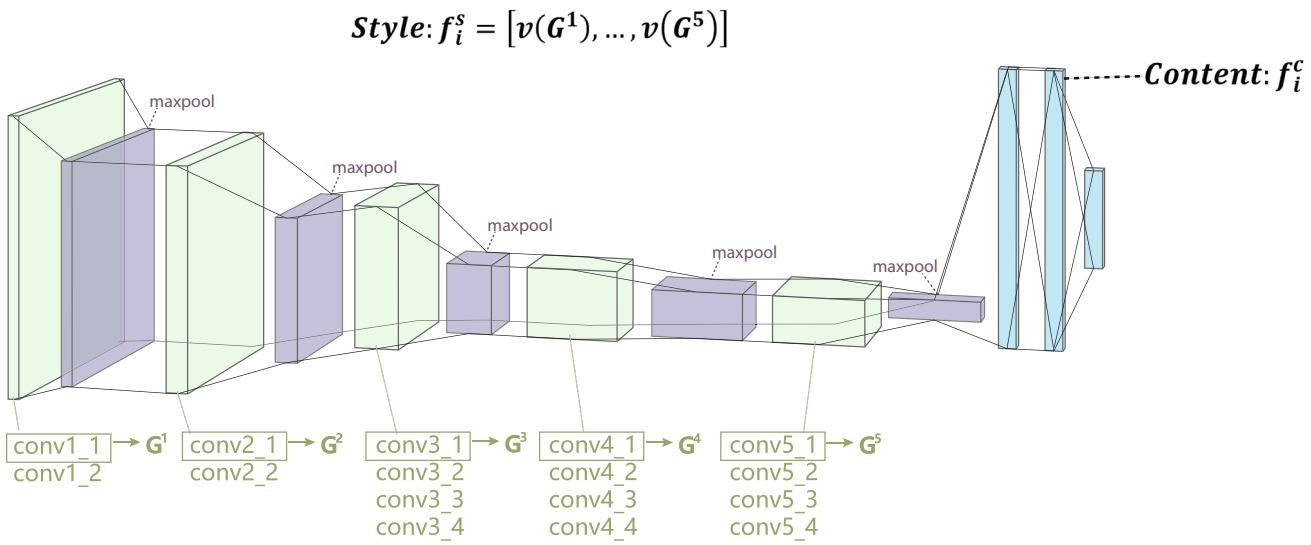


Fig. 1 Diagram of VGG19 convolutional neural network. It shows the process of visual content and visual style extraction

the visual style representations on layers $'con1\_1'$, $'con2\_1'$, $'con3\_1'$, $'con4\_1'$ and $'con5\_1'$ can represent the textures of an image better. Thus Gram matrices G^1, G^2, G^3, G^4, G^5 can be obtained from these five layers, which provides descriptions of the image style (Gatys et al. 2017, 2016). However, the size of Gram matrix is very large, we downsample each Gram matrix into a fixed size of 32×32, and concatenate the style vectors of the downsampled Gram matrix of the five layers. Finally, we obtain a 1024 dimensions vector for an image, representing the visual style.

After we obtain the visual style features for all images, we have to process the vectors because they are too large to handle. Similarly to visual content, we choose a embedding matrix E multiplied by visual style vectors. In this way, lower dimensional visual style representations are obtained.

$$w_i^s = E_s f_i^s, \tag{3}$$

where w_i^s denotes the low dimensional style representation after processed, E_s is the embedding matrix and f_i^s is original visual style representation of image i. The inner product $E_s f_i^s$ transforms the original F_s-dimensions into S-dimensions, which is lower and convenient to calculate.

3.2 Problem formulation

We use U and I to denote the set of users and the set of items, and every user has a positive feedback item set I_u^+ (e.g. ratings) and implicit feedback $I \backslash I_u^+$ (e.g. browsing image dialog). Besides, we identity two aspects of visual information, i.e., visual content and visual style. For one image i, it has the visual content f_i^c and the visual style f_i^s (Table 1).

Table 1 Mathematical Notations

Notations	Description
U	User set, $\|U\| = M$
I	Item set, $\|I\| = N$
u	User
i, j	Image
I_u^+	Positive item set of user u
M_p	The base embedding matrix of users
M_l	The base embedding matrix of items
M_c	The content embedding matrix of users
M_s	The style embedding matrix of users
f_i^c	The visual content representation of item i
f_i^s	The visual style representation of item i

4 The proposed model

In this section, we propose our ranking model (visual *C*ontent and *S*tyle based *B*ayesian *P*ersonalized *R*anking (CSBPR)). First, prediction function is illustrated to explain how we combine the non-visual factors and visual factors. Then, we introduce Bayesian Personalized Ranking, which sorts the scores given by prediction function, and ultimately obtain the recommendation results.

4.1 Prediction function

Given the rating matrix R and the corresponding images of items, we identify that a users' preferences are influenced by three aspects. Specifically, the ratings matrix R from each user's feedback is well recognized as a latent and an important factor in the image recommendation (He and McAuley 2016; Wu et al. 2019). When a user sees a new image, it's natural to pay attention to the style and content of the image. So, we design the two visual aspects in users' preference decision process: the *visual content* aspect that explains the user preference for the content of image, and the *visual style* aspect that shows the user preference for the style of image. These three aspects characterize each user's implicit feedback to images, so we define three embedding space to catch the three aspects characterize (the latent embedding space, the content space and the style space). Specifically, each user associated with three embedding (p_u, c_u and s_u), and each item associated with a base embedding l_i and two visual representation (f_c^i and f_s^i). The content user embedding vector c_u characterizes each user's preference from the content of images. Similarly, the style user embedding vector s_u characterizes each user's preference from the style of images. Thus, by combining the visual content and style of image, we can better model each user u's predicted preference to image i as shown in Fig. 2.

As shown in Fig. 2, our model is based on latent factor model which remarkably avoid the data sparsity and has ability to learn from implicit feedback. The basic prediction formulation of latent factor model is shown as follows:

$$\hat{r}_{u,i} = \alpha + b_u + b_i + p_u^T l_i, \tag{4}$$

where α is presupposed average score of user-item rating matrix, b_u and b_i denote the bias of users and items, p_u and l_i are lower-dimensional vectors respectively representing the latent factors of user u and item i. The inner product $p_u^T l_i$ is in order to connect the latent interest of user u and item i attributions.

Adding the visual content and style of images in our model, the predicted preference score is obtained:

$$\hat{r}_{u,i} = \alpha + b_u + b_i + p_u^T l_i + c_u^T w_i^c + s_u^T w_i^s, \tag{5}$$

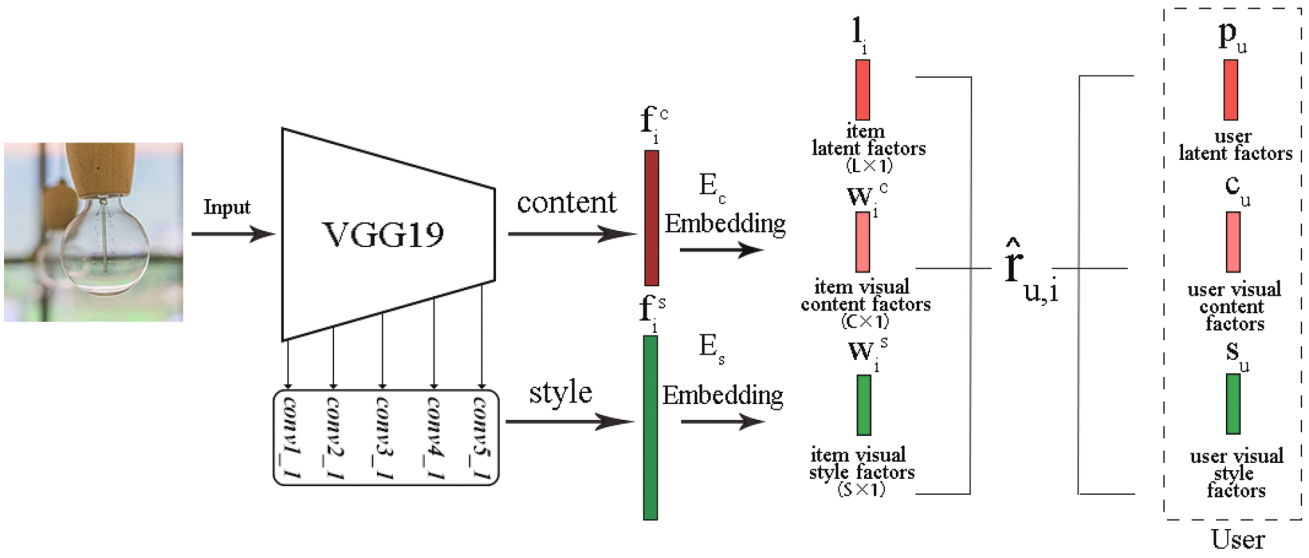


Fig. 2 The overall architecture of the proposed model. Rating dimension is composed by latent factors, visual content factors and visual style factors. Inner products between users and item factors model the compatibility between users and items. For avoiding overfitting, we also introduce bias terms and normalization

where w_i^c and w_i^s are lower-dimensional vectors respectively denoting the visual content and style features of image i. The inner product $c_u^T w_i^c$ contains both the user u interest forward image content and item i visual content feature. It reflects the preference degree which user u gives to item i in visual content. Similarly, the inner product $s_u^T w_i^s$ possesses the user u interest to image style and item i visual style feature, which can represent the preference degree of user u given to item u in visual style. According to Eqs. (1) and (2), the prediction function Eq. (5) can be rewritten as:

$$\hat{r}_{u,i} = \alpha + b_u + b_i + p_u^T l_i + c_u^T (E_c f_i^c) + s_u^T (E_s f_i^s), \tag{6}$$

where the f_i^c and f_i^s are the content representation and style representation which are extracted by pre-trained VGG19 (Simonyan and Zisserman 2015). The E_c and E_s is an embedding kernel which linearly transforms.

4.2 Model learning

Usually, according to data among the existing users and items, we can obtain all item ratings given by users and recommend the several highest ratings items to users, such as funkSVD (Bell et al. 2009). However, under the background of recommending the low single-digit items from tens of millions of items, the rank of high score items appears to be very important. Bayesian Personalized Ranking is such a method, which adopts pairwise approach, sorting among all items. A training set D_S is composed of triples of the form (u, i, j), where user u provides positive feedback to item i and no feedback to item j.

$$D_S = \{(u, i, j) | i \in I_u^+ \wedge j \in I \backslash I_u^+\}. \tag{7}$$

Bayesian Personalized Ranking is based on maximum a posteriori probability. A posteriori probability is proportional to the likelihood probability multiplied by the prior probability. Thus, for every user u:

$$P(\theta | >_u) \propto P(>_u | \theta) P(\theta), \tag{8}$$

where $\theta = [M_p, M_l, M_c, M_s, E_c, E_s]$ denotes the parameters in our model. The $>_u$ represents the preference relation of user u to all of items. The right side of this equation can be divided into two parts, $P(>_u | \theta)$ and $P(\theta)$. As for the first part, Rendle et al. (2009) propose assumptions that the preference relations of every user is independent and one user's preference to every item is independent. Thus, the first part can be rewritten as:

$$\prod_{u \in U} P\left(>_u | \theta\right) = \prod_{(u,i,j) \in D_S} \sigma(\hat{r}_{uij}), \tag{9}$$

where $\hat{r}_{uij}$ can be defined as $\hat{r}_{ui} - \hat{r}_{uj}$ according to latent factor model. For second part $P(\theta)$, Rendle et al. (2009) use Bayesian assumption, assuming that the probability belongs to normal distribution $P(\theta) \sim N(0, \sum_\theta)$. Under this assumption, $lnP(\theta)$ and $\|\theta\|^2$ are in direct proportion. In order to

reduce the unknown hyperparameters and make subsequent model learning easier, we use $\|\theta\|^2$ to replace $lnP(\theta)$. Finally, the optimization criterion is used for Bayesian Personalized Ranking:

$$\begin{aligned} lnP(\theta|>_u) &\propto lnP(>_u|\theta)P(\theta) \\ &= ln\prod_{(u,i,j)\in D_S}\sigma(\hat{r}_{ui}-\hat{r}_{uj}) + lnP(\theta) \\ &= \sum_{(u,i,j)\in D_S} ln\sigma(\hat{r}_{ui}-\hat{r}_{uj}) + lnP(\theta) \\ &= \sum_{(u,i,j)\in D_S} ln\sigma(\hat{r}_{ui}-\hat{r}_{uj}) + \lambda\|\theta\|^2, \end{aligned} \tag{10}$$

where σ is the logistics sigmoid function, in practice we set $\sigma(x)$ as a sigmoid function that transforms the input into range (0, 1). λ is model specific regularization term that regularizes the user and item embeddings. This method can be efficiently learned via stochastic gradient ascent. There are four sets of parameters should be updated in the above loss function: (1) the sampled triple *(u,i,j)* (2) the latent factors parameters, (3) the visual content parameters, and (4) the visual style parameters. We train the above parameters with Adam in TensorFlow.

4.3 Scalability

Due to the fact that our proposed loss function is on the base of BPR model, we compare the time complexity with BPR model. Due to the fact that the number of positive samples is far less than the negative samples. Thus, according to the Bayesion hypothesis, the BPR model costs $\mathcal{O}(MCL)$, where M is the number of users, C is the positive samples of items, and L is the dimension of latent factors. Similarly in our proposed CSBPR model, for latent factors, visual content factors and visual style factors, the M and C are same and the dimensions of them are L, C and S respectively. Therefore, the proposed CSBPR model cost $\mathcal{O}(MC(L+S+C))$.

5 Experiments

5.1 Experimental settings

Dataset In order to demonstrate the real performance of our proposed model better, we select a large dataset from the biggest image sharing social platform *Flickr*, which is an extension on the basis of the widely used NUS-WIDE dataset (Chua et al. 2009). NUS-WIDE contains approximately 270,000 images which are classified by 81 humans. In the process of data preprocessing, we filter out those users and images that have less than 10 rating records. In this way, we draw a smaller but in higher sparsity dataset with 4418 users, 31,460 images and 761,812 rating records. In data splitting procedure, we adopt the leave-one-out strategy (Chen et al. 2017; He et al. 2017), which is mentioned in several research works. To be specific, we choose the last rating record of users as test data and the remaining are used as train data. In order to tune algorithm parameters, we randomly select 5% data from train set as validation set. Table 2 shows the statistics of the datasets after splitting.

Table 2 The statistics of the dataset after splitting

Dataset	Users	Images	Ratings
Train	4418	31,460	752,948
Validation	4418	1573	37,647
Test	4418	4418	4418

Evaluation metric In order to evaluate our model performance more scientifically, we use two the most widely used metrics for top-K ranking recommendation: the Hit Ratio (HR) and the Normalized Discounted Cumulative Gain (NDCG) (Chen et al. 2017; He and McAuley 2016). HR calculates the percentage of images that user likes in top-K recommend, which reflects the accuracy of recommendation. The NDCG values the gain between all images position in top-K list and the hit images list, which represents the precision of top-K recommend.

Baselines We compare our proposed model Visual Content and Style based Image Recommendation with the following baselines:

- *BPR*: This method is a classical ranking method based latent factor model, which is extended on the basic matrix factorization, adding the bias terms and regularization. This method uncovers the latent features of users and items and comes up with more personalized rankings for each user. It is acknowledged recognized as a strong baseline for personalized recommendation (Rendle et al. 2009).
- *VBPR*: This method takes visual content into recommendation, which considering both latent dimensions of users' preference and visual dimensions. The visual content is extracted from pretrained VGG19 network, same as our proposed model (He and McAuley 2016).
- *SBPR*: where S denotes visual style. This baseline only considers the influence of visual style factors and latent factors, excluding visual content, where visual style feature is extracted from the same method as our proposed model.
- *Content+Style*: This method is a modification of our proposed model, which only considers the influence of visual content and visual style. The latent factors are not

taken into account. Similarly, the visual content features and style features are as same as our proposed model.

Parameter setting The one important parameter in our model is the dimension D of the user and image embedding, including visual content embedding, visual style embedding and latent factor embedding. We test the performance of our model under D in [16, 32, 64, 128]. For convenience, we set the same dimension for these three embedding in order to obtain the best performance of our model. We find that when $D = 32$, our proposed model reaches the best performance.

Another main parameter is coefficient of normalization λ. Normalization is aimed to avoid overfitting. We test the λ in [0.00001,0.0001,0.001,0.01] with D = 32 to find the best state of our model showed as Fig. 4. And the best set is $\lambda = 0.0001$.

There are several parameters in baselines. For fair comparison, all the model are test on the same dataset. We set the dimensions D of user and image embedding as 32, as same as our proposed model and all the baselines are in the best performance. For all models, we stop training when both HR@K and NDCG@K (K = [5, 10, 15, … , 50]) in validation set begin to decrease.

5.2 General performance

There are two subfigures in Fig. 3 showing the overall performance of our proposed model and baselines on HR@K and NDCG@K in top-K (K = [5, 10, 15, … , 50]) and results are shown in Table 3, where the results of our CSBPR are in bold showing its superiority. According to the graphs, it is noticeable that our proposed model is outstanding among all the models in both HR@K and NDCG@K. And the VBPR and SBPR illustrate the almost same trend in both HR@K and NDCG@K, except for the results when top-20, which are better than the results of BPR with the same top-K. VBPR improves over BPR about average 5% on NDCG and average 8% on HR by taking visual image information into the model, which makes up for the lack of data sparsity. As for our proposed model, no matter in NDCG or HR, it is always in the best position. On average, our proposed model CSBPR gradually rises to approximately 20% over

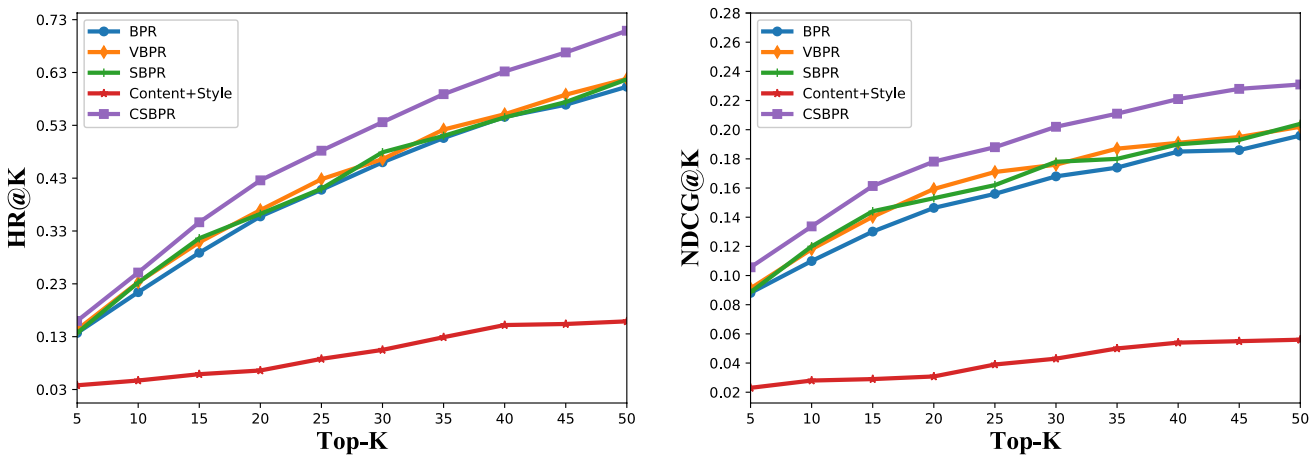


Fig. 3 Overall performance of different models on NDCG@K and HR@K

Table 3 HR@K and NDCG@K comparisons for different models

Metric	Models	K = 5	K = 10	K = 15	K = 20	K = 25	K = 30	K = 35	K = 40	K = 45	K = 50
HR	BPR	0.136	0.214	0.288	0.357	0.408	0.460	0.506	0.546	0.569	0.603
	VBPR	0.143	0.232	0.308	0.369	0.428	0.466	0.522	0.551	0.588	0.618
	SBPR	0.136	0.232	0.316	0.362	0.410	0.479	0.510	0.545	0.574	0.617
	Content+Style	0.038	0.047	0.059	0.066	0.088	0.105	0.129	0.152	0.154	0.159
	CSBPR	**0.159**	**0.215**	**0.346**	**0.425**	**0.482**	**0.536**	**0.589**	**0.632**	**0.668**	**0.709**
NDCG	BPR	0.088	0.109	0.130	0.146	0.156	0.168	0.174	0.185	0.186	0.196
	VBPR	0.091	0.118	0.140	0.153	0.170	0.171	0.187	0.191	0.198	0.202
	SBPR	0.087	0.120	0.144	0.153	0.162	0.178	0.180	0.190	0.193	0.204
	Content+Style	0.023	0.028	0.029	0.030	0.039	0.043	0.050	0.054	0.055	0.056
	CSBPR	**0.105**	**0.134**	**0.161**	**0.178**	**0.188**	**0.202**	**0.211**	**0.221**	**0.228**	**0.231**

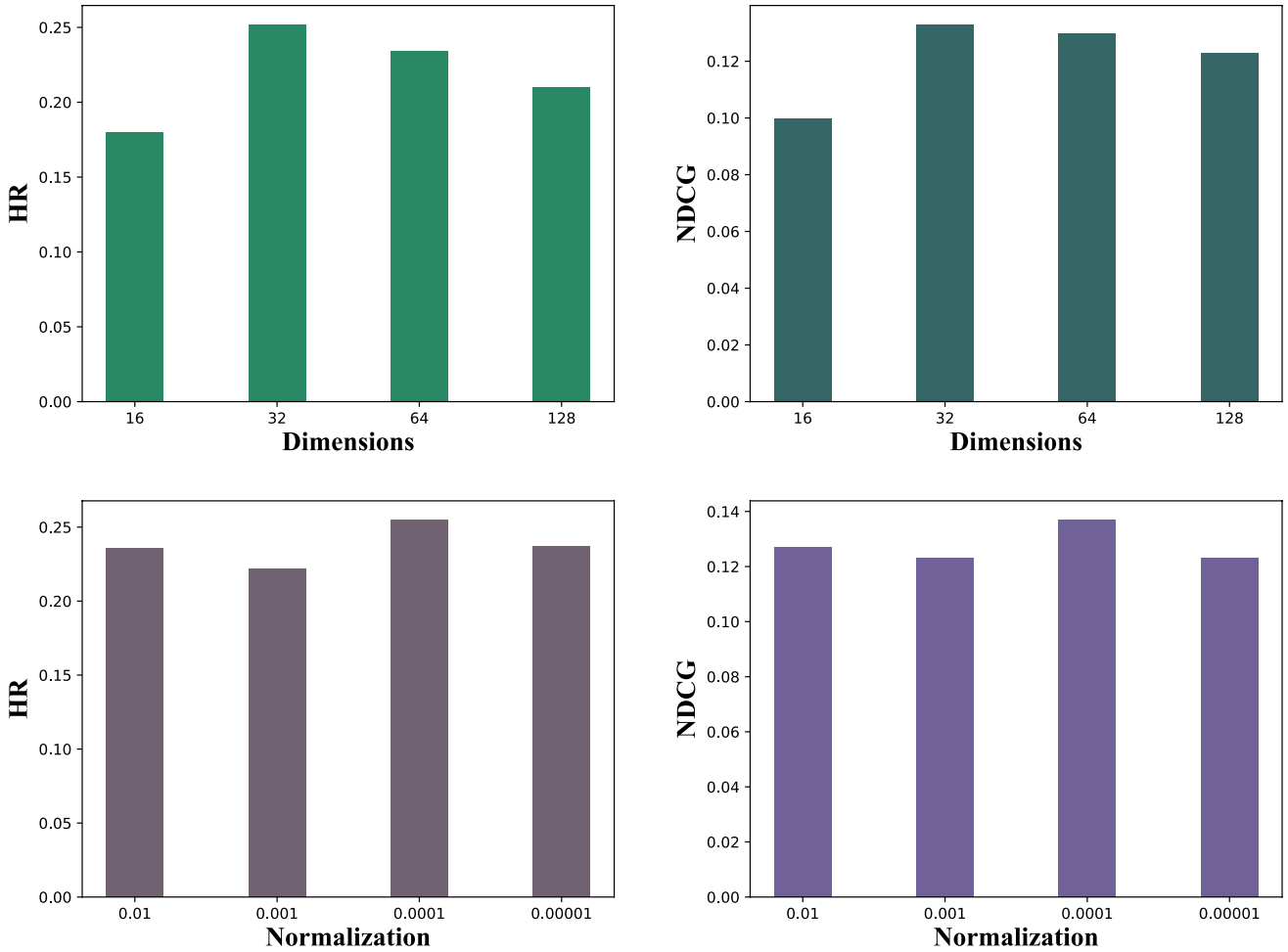


Fig. 4 Diagram of the performance of our proposed model in diverse embedding dimensions in top-10 and varying normalization in same embedding dimensions 32

BPR baseline, and more than 15% improvement over the VBPR on both NDCG and HR. Last but not the least, our proposed model stands out apparently with the increasing of value of top-K.

5.3 Parameter experiments

In this part, we introduce the process of the parameters tuning. There are two main parameters that should be tuned particularly: the dimensions of the user and image embedding D and the coefficient of normalization λ. Firstly, we test our embedding dimensions in [16,32,64,128] in top-10 recommendation showed as Fig. 4. During the experiments, we found that the performance of recommendation not improves along with the increasing of embedding dimensions. A large dimension can contribute to overfitting and improve unobviously due to the data sparsity. We found that when embedding dimension = 32, the proposed model gets the best performance relatively. Next, we conducted experiments on coefficient of normalization under embedding dimension = 32, which is designed to avoid overfitting. We set normalization in the range [0.01, 0.001, 0.0001, 0.00001]. According to the chart showed as (Fig. 4), even though the results are not obvious under these four dimensions, the performance when $\lambda = 0.0001$ is better slightly.

6 Conclusion

Visual factors play an essential role in people's daily life. Except for visual content, which can directly express the meaning and opinion of images, visual style also influences users' choices and preference deeply. In this paper, we proposed a scalable method (CSBPR) that incorporates the visual content and visual style into latent factor model. Our model is trained with bayesian personalized ranking with stochastic gradient ascent. Experiments on the large real-world datasets demonstrate that our proposed model significantly outperforms other baselines. As part of our future work, we will consistently extend our model by considering more effective information.

Funding Funding was provided by NSAF Joint Fund (Grant No. 61602147).

References

Bartolini, I., Moscato, V., Pensa, R.G., Penta, A., Picariello, A., Sansone, C., Sapino, M.L.: Recommending multimedia visiting paths in cultural heritage applications. Multimedia Tools Appl. **75**(7), 3813–3842 (2016)

Bell, R., Koren, Y., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **8**, 30–37 (2009)

Canini, L., Benini, S., Leonardi, R.: Affective recommendation of movies based on selected connotative features. IEEE Trans. Circuits Syst. Video Technol. **23**(4), 636–647 (2012)

Chen, J., Zhang, H., He, X., Nie, L., Liu, W., Chua, T.S.: Attentive collaborative filtering: multimedia recommendation with item-and component-level attention. In: The ACM Conference on Research and Development in Information Retrieval, ACM, pp. 335–344 (2017)

Chen, T., He, X., Kan, M.Y.: Context-aware image tweet modelling and recommendation. In: ACM Multimedia, ACM, pp. 1018–1027 (2016)

Chu, W.T., Wu, Y.L.: Image style classification based on learnt deep correlation features. IEEE Trans. Multimedia **20**(9), 2491–2502 (2018)

Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: Nus-wide: a real-world web image database from national university of singapore. In: The ACM International Conference on Image and Video Retrieval, ACM, p. 48 (2009)

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: a deep convolutional activation feature for generic visual recognition. In: International Conference on Machine Earning, pp. 647–655 (2014)

Fan, J., Keim, D.A., Gao, Y., Luo, H., Li, Z.: Justclick: personalized image recommendation via exploratory search from large-scale flickr images. IEEE Trans. Circuits Syst. Video Technol. **19**(2), 273–288 (2008)

Gatys, L.A., Ecker, A.S., Bethge, M.: A neural algorithm of artistic style. (2015) arXiv preprint arXiv:150806576

Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (2016)

Gatys, L.A., Ecker, A.S., Bethge, M., Hertzmann, A., Shechtman, E.: Controlling perceptual factors in neural style transfer. In: The IEEE Conference on Computer Vision and Pattern Recognition, pp. 3985–3993 (2017)

Gelli, F., Uricchio, T., He, X., Del Bimbo, A., Chua, T.S.: Beyond the product: discovering image posts for brands in social media. In: The ACM Conference on Multimedia Conference (2018)

Guo, G., Meng, Y., Zhang, Y., Han, C., Li, Y.: Visual semantic image recommendation. IEEE Access **7**, 33424–33433 (2019)

He, R., McAuley, J.: VBPR: visual bayesian personalized ranking from implicit feedback. In: The Association for the Advancement of Artificial Intelligence, pp. 144–150 (2016)

He, X., Liao, L., Zhang, H., Nie, L., Hu, X., Chua, T.S.: Neural collaborative filtering. In: The International Conference on World Wide Web, pp. 173–182 (2017)

Hsiao, W.L., Grauman, K.: Learning the latent "look": Unsupervised discovery of a style-coherent embedding from fashion images. In: IEEE International Conference on Computer Vision, IEEE, pp. 4213–4222 (2017)

Jiang, M., Cui, P., Wang, F., Zhu, W., Yang, S.: Scalable recommendation with social contextual information. IEEE Trans. Knowl. Data Eng. **26**(11), 2789–2802 (2014)

Koren, Y.: Factorization meets the neighborhood: a multifaceted collaborative filtering model. In: The International Conference on Knowledge Discovery and Data Mining, ACM, pp. 426–434 (2008)

Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)

Liu, Q., Wu, S., Wang, L.: Deepstyle: learning user preferences for visual recommendation. In: The ACM Conference on Research and Development in Information Retrieval, ACM, pp. 841–844 (2017)

Luo, H., Zhang, X., Chen, B., Guo, G.: Multi-view visual bayesian personalized ranking from implicit feedback. In: The Conference on User Modeling, Adaptation and Personalization, ACM, pp. 361–362

Niu, W., Caverlee, J., Lu, H.: Neural personalized ranking for image recommendation. In: ACM International Conference on Web Search and Data Mining, ACM, pp. 423–431 (2018)

Radenović, F., Tolias, G., Chum, O.: Fine-tuning CNN image retrieval with no human annotation. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1655–1668 (2018)

Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: The Conference on Uncertainty in Artificial Intelligence, pp. 452–461 (2009)

Sarwar, B.M., Karypis, G., Konstan, J.A., Riedl, J., et al.: Item-based collaborative filtering recommendation algorithms. Int. World Wide Web Conf. **1**, 285–295 (2001)

Shankar, D., Narumanchi, S., Ananya, H., Kompalli, P., Chaudhury, K.: Deep learning based large scale visual recommendation and search for e-commerce. (2017) arXiv preprint arXiv:170302344

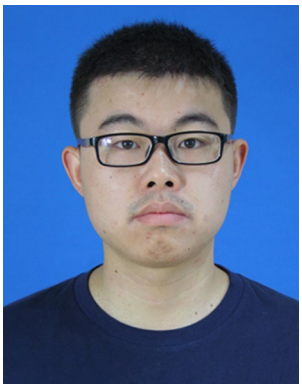
Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)

Tzelepi, M., Tefas, A.: Deep convolutional learning for content based image retrieval. Neurocomputing **275**, 2467–2478 (2018)

Wang, S., Wang, Y., Tang, J., Shu, K., Ranganath, S., Liu, H.: What your images reveal: exploiting visual contents for point-of-interest recommendation. In: The International Conference on World Wide Web, pp. 391–400 (2017)

Wu, L., Chen, L., Hong, R., Fu, Y., Xie, X., Wang, M.: A hierarchical attention model for social contextual image recommendation. IEEE Trans. Knowl. Data Eng. (2019)

Suchang Luo is currently working towards the M.S. degree at Hefei University of Technology, China. Her research interests include recommender systems and data mining.



Lei Chen is currently working towards the Ph.D. degree at Hefei University of Technology, China. He received the M.S. degree from Hefei University of Technology in 2019. His research interests include multimedia analysis and data mining.

Le Wu is currently an associate professor at the Hefei University of Technology (HFUT), China. She received the Ph.D. degree from the University of Science and Technology of China (USTC). Her general area of research interests is data mining, recommender systems and social network analysis. She has published more than 40 papers in referred journals and conferences. Dr. Le Wu is the recipient of the Best of SDM 2015 Award, and the Distinguished Dissertation Award from China Association for Artificial Intelligence (CAAI) 2017.