DeepStyle: Learning User Preferences for Visual Recommendation

Qiang Liu[1,3], Shu Wu[1], Liang Wang[1,2,3]
[1]Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition
[2]Center for Excellence in Brain Science and Intelligence Technology
Institute of Automation, Chinese Academy of Sciences
[3]University of Chinese Academy of Sciences
{qiang.liu, shu.wu, wangliang}@nlpr.ia.ac.cn

ABSTRACT

Visual information is an important factor in recommender systems. Some studies have been done to model user preferences for visual recommendation. Usually, an item consists of two fundamental components: *style* and *category*. Conventional methods model items in a common visual feature space. In these methods, visual representations always can only capture the categorical information but fail in capturing the styles of items. Style information indicates the preferences of users and has significant effect in visual recommendation. Accordingly, we propose a DeepStyle method for learning style features of items and sensing preferences of users. Experiments conducted on two real-world datasets illustrate the effectiveness of DeepStyle for visual recommendation.

Keywords

Visual recommendation, user preferences, style features

1. INTRODUCTION

Nowadays, it is important to sense and understand what users prefer and need, which has been the fundamental component of various applications. People always say "Seeing is believing." Accordingly, *visual* information plays an important role in understanding user behaviors, especially in domains such as buying clothes, jewelries, house decorations and so on. It is crucial to investigate the visual dimensions of user preferences and items for better personalized recommendation.

Some studies have been done on investigating visual features for user modeling, including cloth matching [4, 7] and visual recommendation [2, 3]. Functional Pairwise Interaction Tensor Factorization (FPITF) [4] predicts the matching of clothes in outfits with tensor factorization. Personalized matching of items based on visual features has also been investigated [7]. Visual Bayesian Personalized Ranking (VBPR) [3] extends the framework of Bayesian Personalized Ranking (BPR), and incorporates visual features for promoting the performance of item recommendation in implicit feedback scenarios. VBPR is further extended with dynamic dimensions to model the visual evolution of fashion trends in visual recommendation [2].

Above conventional methods modeling items in a common visual feature space, which may fail to capture different styles of items. In Figure 1, we cluster items in the clothing subset of the Amazon dataset[1] [7]. The visual features used here are the Convolutional Neural Networks (CNN) visual features extracted from the Caffe reference model[2] [5, 6], which have been used in several existing works [1, 2, 3, 7]. Intuitively, we can observe that, one category (e.g., ups, dresses, pants, shoes, bags and watches) of items are assigned to one cluster. It is obvious that, items with different styles (e.g., casual, athletic and formal) can not be distinguished in the figure, even between the male styles and the female styles. Items with similar styles are usually bought together, but they are not similar in the visual feature space. Thus, it is hard for a recommender to make reliable prediction in such feature space. For example, in the common visual feature space, the similarity between suit pants and leather shoes is much small than the similarity between suit pants and jeans. However, suit pants and leather shoes are usually bought together by the same user. Obviously, categorical information plays a dominant role in the representation of an item. Recently, the impact of categorical information has been considered in Sparse Hierarchical Embeddings (Sherlock) [1]. In Sherlock, the embedding matrices for transferring visual features to style features vary among different categories. However, one embedding matrix for each category leads to very large amount of parameters to be learned, although a sparse operation on a prior category tree is performed.

Therefore, we need to investigate the properties of items. We can conclude that, an item consists of two components: *style* and *category*. Accordingly, we assume that:

$$item = style + category \ . \tag{1}$$

Based on the above assumption, we can obtain the style features of an item via eliminating the corresponding categorical information. In this work, we propose a novel method

[1]http://jmcauley.ucsd.edu/data/amazon/
[2]bvlc_reference_caffenet from caffe.berkeleyvision.org

SIGIR '17, August 07-11, 2017, Shinjuku, Tokyo, Japan

DOI: http://dx.doi.org/10.1145/3077136.3080658

Figure 1: Part of the clustering results of items in the Clothing subset of the Amazon dataset [7]. It is measured by the CNN visual features [5, 6]. One row is a cluster. We can observe that, items in the same category are assigned to one cluster, and different styles of clothing are not distinguished.

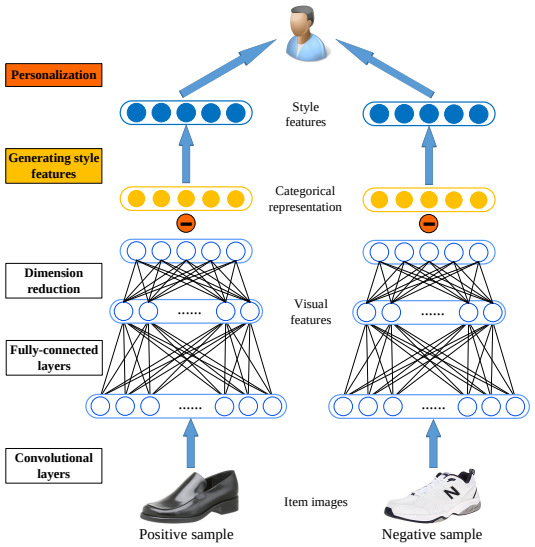


Figure 2: The illustration of DeepStyle for learning styles of items and preferences of users.

called **DeepStyle**. In DeepStyle, images of items are feeded into a deep CNN model. For each item, the output layer of CNN generates its visual feature vector. Then, we subtract a latent representation of the corresponding category from the visual feature vector generated by CNN, and then obtain the style features of items. Finally, we incorporate style features in the widely-used BPR [8] framework for personalized recommendation.

2. NOTATIONS

In this work, we focus on predicting users' implicit feedbacks, i.e., users' selections, on items. We have a set of users denoted as $\mathcal{U}$, and a set of items denoted as $\mathcal{I}$. Users may have selection behaviors on some items, where $\mathcal{I}^u$ denotes the set of items selected by user u. Each item i is associated with an image describing its visual information, and belongs to a specific category l_i.

3. DEEPSTYLE

Conventional methods for visual recommendation are mostly focusing on modeling items in a common visual feature space. This may fail to capture different styles of items. As shown in Figure 1, items with similar styles may be not similar in the visual space at all. And categorical information is dominant in the common visual space. Thus, in visual recommendation, it is vital to eliminate categorical information from representations of items. Accordingly, we propose a DeepStyle method for learning style features of items and preferences of users, as illustrated in Figure 2.

First, for each item i, we feed the corresponding image into a deep CNN model. Following several representative works [1, 2, 3, 7] for visual recommendation, the CNN model applied is the Caffe reference model [7]. It consists of 5 convolutional layers followed by 3 fully-connected layers. The model is pre-trained on 1.2 million ImageNet images[3], for capturing some common visual concepts. On the output layer of the CNN model, there is a 4096 dimensional visual feature vector denoted as $\mathbf{v}_i \in \mathbb{R}^{4096}$.

Then, to obtain style features, according to Equation 1, we subtract items' latent categorical representations from the visual features generated by CNN. For item i, we can calculate its style features as

$$\mathbf{s}_i = \mathbf{E}\mathbf{v}_i - \mathbf{l}_i \ , \tag{2}$$

where $\mathbf{s}_i \in \mathbb{R}^d$ denotes the style features of item i, $\mathbf{l}_i \in \mathbb{R}^d$ denotes the latent categorical representation of the corresponding category l_i, $\mathbf{E} \in \mathbb{R}^{d \times 4096}$ is a matrix for transferring visual features to lower dimensionality on the top layer, and d is the dimensionality of learned representations.

Furthermore, we incorporate the style features in the BPR [8] framework, which is the state-of-the-art method for modeling implicit feedbacks, for sensing preferences of users. The prediction of user u on item i can be made as

$$\hat{y}_{u,i} = (\mathbf{p}_u)^T (\mathbf{s}_i + \mathbf{q}_i) \ , \tag{3}$$

where $\mathbf{p}_u \in \mathbb{R}^d$ denotes the latent representation of user u, and $\mathbf{q}_i \in \mathbb{R}^d$ denotes the latent representation of item i, which can capture the collaborative information among users and items. For user u, with an arbitrary negative sample i', the model needs to fit

$$\hat{y}_{u,i} > \hat{y}_{u,i'} \ , \tag{4}$$

where i is a positive item that $i \in \mathcal{I}^u$, and i' is a negative item that $i' \notin \mathcal{I}^u$. Then, in the BPR framework, we need to maximize the following probability

$$p\left(u, i > i'\right) = g\left(\hat{y}_{u,i} - \hat{y}_{u,i'}\right) \ , \tag{5}$$

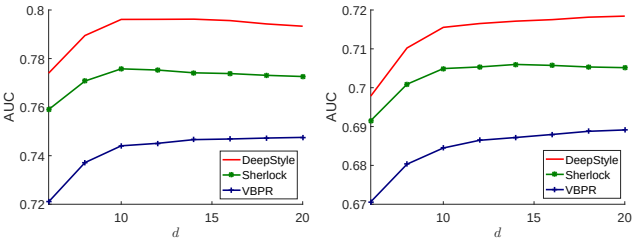
where the activation function $f(x)$ is usually chosen as $g(x) = 1/(1+e^{-x})$. Incorporating the negative log likelihood, we can minimize the following objective function equivalently

$$J = \sum_{u,i} \ln\left(1 + e^{-\left(\hat{y}_{u,i} - \hat{y}_{u,i'}\right)}\right) + \lambda \|\theta\|^2 \ , \tag{6}$$

[3]http://image-net.org/

Table 1: Performance comparison on predicting users preferences on items measured by AUC. The dimensionality is $d = 10$ on both datasets.

dataset	setting	BPR	VBPR	Sherlock	DeepStyle
Clothing	warm-start	0.6243	0.7441	0.7758	**0.7961**
	cold-start	0.5037	0.6915	0.7167	**0.7317**
Home	warm-start	0.5848	0.6845	0.7049	**0.7155**
	cold-start	0.5053	0.6140	0.6322	**0.6396**



(a) Clothing. (b) Home.

Figure 3: Performance of DeepStyle, Sherlock and VBPR with varying dimensionality under the warm-start setting measured by AUC.

where θ denotes all the parameters to be estimated in DeepStyle, and λ is a hyper-parameter to control the power of regularization. Then, the derivations of J with respect to all the parameters in DeepStyle can be calculated, and we can employ Stochastic Gradient Descent (SGD) to estimate the model parameters.

4. EXPERIMENTS

In this section, we introduce our experiments to evaluate the effectiveness of DeepStyle. First, we introduce our experimental settings. Then, we give comparison among some state-of-the-art methods and analyze the impact of dimensionality. Finally, we demonstrate the clustering visualization of the style features.

4.1 Experimental Settings

Our experiments are conducted on two subsets of the Amazon dataset [7]. In particular, we adopt the "Clothing, Shoes and Jewelry" subset and the "Home and Kitchen" subset, which are named as the **Clothing** dataset and the **Home** dataset for short. Visual features are important in buying things such as clothes, shoes, jewelries, house decorations and so on. For example, visual features have been proven to be useful in cloth recommendation [1, 2, 3, 7]. The Clothing dataset consists of 74 categories, e.g., jeans, pants, shoes, shirts and dresses. The home dataset contains 86 categories, e.g., sheets, furniture, pillows and cups.

In our experiments, we empirically set the regulation parameter as $\lambda = 0.01$, and the learning rate for SGD is set to be 0.01. For each dataset, we use 80% instances for training, and remaining 20% instances for testing. Moreover, we remove users with less than 5 records and more than 100 records. There are two types of evaluation settings on both datasets during the testing procedure: **warm-start** and **cold-start**. The former focuses on measuring the overall ranking performance, while the latter captures the capability to recommend cold-start items, i.e., items with less than 5 records during training, in the system.

Then, following some previous works [3, 8], for evaluating the performance of all the methods, we apply the Area Under the ROC Curve (**AUC**) metric:

$$AUC = \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} \frac{1}{|set\left(i \in \mathcal{I}^u, i' \notin \mathcal{I}^u\right)|} \sum_{i \in \mathcal{I}^u, i' \notin \mathcal{I}^u} \delta\left(p_{u,i} > p_{u,i'}\right),$$

where $\delta(.)$ is the Dirac delta function, which outputs 1 when the condition is met, and 0 otherwise. The larger the AUC value, the better the performance.

Moreover, to investigate the performance on predicting users preferences on items, some state-of-the-art methods are compared: **BPR** [8], **VBPR** [3] and **Sherlock** [1]. BPR is a widely-used method for modeling implicit feedbacks. Based on BPR, VBPR incorporates visual features of items. Sherlock extends VBPR, and takes categorical information into consideration. As in [1, 3], visual features used in VBPR and Sherlock are CNN features extracted from the Caffe reference model [5, 6].

4.2 Performance Comparison

Table 1 illustrates the performance comparison among DeepStyle, Sherlock, VBPR and BPR under warm-start and cold-start settings, where the dimensionality is $d = 10$. We can clearly observe that, methods incorporating visual features can outperform the baseline method BPR with relatively large advantages on both datasets. The advantages comparing with BPR are even larger under the cold-start setting, which indicates that visual features can model properties of cold-start items when observations are not enough, and promote the performance. Moreover, methods modeling categorical effects on styles of items, i.e., Sherlock and DeepStyle, have better performance than VBPR on both datasets under both settings. DeepStyle outperforms VBPR by 5.2% and 3.1% on Clothing and Home respectively under the warm-start setting, and 4.1% and 2.6% under the cold-start setting. This shows it is vital to take categorical information into consideration for modeling styles of items. Moreover, Sherlock is clearly the best one among all the compared methods in visual recommendation, and outperforms all the compared methods. Comparing with Sherlock, DeepStyle improves AUC values by 2.1% and 1.1% on Clothing and Home respectively under the warm-start setting, and 1.5% and 0.7% under the cold-start setting. These improvements indicates the superiority of DeepStyle for learning style features of items and preferences of users.

4.3 Impact of Dimensionality

Furthermore, to investigate the dimensionality sensitivity, we illustrate the performance of DeepStyle, Sherlock and VBPR under the warm-start setting with varying dimensionality $d = [6, 8, 10, 12, 14, 16, 18, 20]$ in Figure 3. It is clear that, DeepStyle can consistently outperform Sherlock and VBPR. On both datasets, the performance of DeepStyle stays stable after $d = 10$. This indicates that, DeepStyle is not very sensitive with the dimensionality, and shows the flexility of DeepStyle. Accordingly, the performance with $d = 10$ is reported in the rest of our experiments. Moreover, Comparing with VBPR and DeepStyle, Sherlock has tendency to overfit the data when the dimensionality is large.

Figure 4: Visualization of part of the clustering results of items in the Clothing dataset. It is measured by the style features, which are learned in the proposed DeepStyle model. Items in one square belong to the same cluster. It is obvious that, different styles of items can be distinguished by using DeepStyle.

This may indicates that, in Sherlock, one embedding matrix for each category requires to estimate too many parameters.

4.4 Visualization

Based on the 10-dimensional style features learned in DeepStyle, items in the Clothing dataset are clustered into several distinct styles. The visualization of part of the clustering results is shown in Figure 4. It is obvious that, one category of items are assigned to different clusters, and items in one cluster have very similar styles. Female items are in the top two rows, and male items are in the bottom row. The left column covers formal and official styles of clothing, in which the middle square is closer to the banquet-style. Items in the middle column are mostly casual, school-style or street-style clothing for women and men. In the right column, items somehow belong to the old-style, and the middle square is more likely the clothing style of middle-aged women. Each cluster clearly covers a distinct style of clothing. Note that, during the training of DeepStyle, there is absolutely no supervision on styles. Obviously, our proposed method is able to automatically capture different styles of items.

5. CONCLUSIONS

In this paper, we propose a novel method, i.e., DeepStyle, for learning styles of items and preferences of users. DeepStyle subtracts categorical information from visual features of items generated by CNN, and style features are obtained. Based on the learned style features and the BPR framework, personalized recommendation can be performed. Experimental results demonstrate the successful performance of DeepStyle for visual recommendation.

6. ACKNOWLEDGMENTS

This work is jointly supported by National Key Research and Development Program (2016YFB1001000), National Natural Science Foundation of China (61403390, U1435221), CCF-Tencent Open Fund and CCF-Venustech Hongyan Research Fund.

7. REFERENCES

[1] R. He, C. Lin, J. Wang, and J. McAuley. Sherlock: Sparse hierarchical embeddings for visually-aware one-class collaborative filtering. In *IJCAI*, pages 3740–3746, 2016.

[2] R. He and J. McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *WWW*, pages 507–517, 2016.

[3] R. He and J. McAuley. Vbpr: visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[4] Y. Hu, X. Yi, and L. S. Davis. Collaborative fashion recommendation: A functional tensor factorization approach. In *MM*, pages 129–138, 2015.

[5] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, pages 675–678, 2014.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[7] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel. Image-based recommendations on styles and substitutes. In *SIGIR*, pages 43–52, 2015.

[8] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. In *UAI*, pages 452–461, 2009.