Double Deep Features for Apparel Recommendation System

Yichi Lu
Osaka Prefecture University
mcb04031@edu.osakafu-u.ac.jp

Yufeng Duan
Osaka Prefecture University
mbb04016@edu.osakafu-u.ac.jp

Ryosuke Saga
Osaka Prefecture University
saga@cs.osakafu-u.ac.jp

Abstract

This study describes a recommendation system embedded in the double features extracted by convolutional neural networks (CNNs). Several probabilistic models, such as probabilistic matrix factorization (PMF)-based approaches, have been utilized for recommendation systems based on a CNN model. Each recommendation algorithm utilizes a single CNN model to extract precise features about documents and pictures, and the systems with CNN have contributed in improving the performance in rating prediction. Meanwhile, the systems for some items should consider at least two precise features simultaneously, and the extension to embed multiple CNN models is necessary. However, methods that integrate multiple CNN-based features into existing recommendation systems, such as PMF, are not available. Thus, this study proposes a novel probabilistic model that integrates double CNNs into PMF. For apparel goods, two trained CNNs from document and image shape features are combined, and the latent variables of users and items are optimized based on the vectorized features of CNNs and rating. Extensive experiments demonstrate that our model outperforms other recommendation models.

Index Terms – recommender system, deep learning, image shape feature,convolutional neural network, probabilistic matrix factorization

1. Introduction

The sparseness of useritem rating in e-commerce services is a major data type. Traditional recommendation systems have been committed to predict the rating prediction accuracy based on several supplemental features, such as user demographics and social networks [1–3].

The target of this study is apparels. Apparels are different from other goods, such as books and movies. Visual information is more important than document information because generally,document information is useful in improving recommendation accuracy [4–7]. Several methods, such as TFIDF, latent semantic analysis, and latent Dirichlet allocation (LDA), have been utilized for treating item descriptions. Collaborative topic regression is used together with LDA and probabilistic matrix factorization (PMF) for recommendations, and their variants have been proposed [7].Wang et al. proposed collaborative deep learning that integrates PMF and deep learning to learn hidden representation [6]. However, this method utilizes bag-of work, which cannot treat contextual information. Kim et al. addressed this problem and proposed the combination of convolutional matrix factorization (ConvMF) with PMF and CNN [4].

For apparels, item descriptions are regarded as a sub information. Thus, pictures and images are main contents to express such items. Item descriptions are insufficient for explaining the shape, texture, and design of items because of the limitation of human imagination. By contrast, images and pictures can deliver visual features at ease. For example, in apparel and fashion magazines, their contents are mostly pictures and images that convey the features of apparels to readers, such as in the saying, A picture is worth a thousand words.

Several years ago, there were several fashion recommender systems that did not use deep learning but used visual feature [8, 9]. Liu et al. proposed a latent SVM based recommendation model to incorporate the matching rules among visual feature, attribute and occasion within a unified framework [8]. Hu et al. proposed a functional tensor factorization approach to generate an outfit by modeling the interactions between user and fashion items [9]. Subsequently, deep learning has made rapid progress in the fields of image processing and natural language processing. It can extract high-level features and achieve better performance than traditional methods.

Recently, several approaches have been proposed for deep learning-based recommendation [4, 10–13]. As previously mentioned, ConvMF [4] is a recommen-

URI: https://hdl.handle.net/10125/64282
978-0-9981331-3-3

Table 1. comparison of typical recommendation system combining matrix factorization with deep learning

	ConvMF [4]	VBPR [15]	ISFMF [16]	Proposed
Relationship between model and deep learning	Tight	Loose	Tight	Tight
Recommended object	Movie	Apparel	Apparel	Apparel
Side information	Text	Image	Image	Text + Image
Prediction	Rating(1-5)	Binary(0-1)	Rating(1-5)	Rating(1-5)

dation model that uses CNN. Although the document information and user-item rating of the ConvMF model are tight, they do not fully utilize the document and visual information. Here, tightly indicates that the parameters of CNN and PMF are optimized simultaneously. Hidasi et al. [13] proposed a session-based recommendation model based RNN. Wang et al. used CNN as a component for image feature extraction combined with traditional POI recommender systems [12]. There are several fashion recommender systems with deep learning [14–17]. He et al. investigated the usefulness of visual features for retrieval [14] and personalized ranking tasks [15]. Duan et al. proposed a recommender system that integrated the CNN into PMF to use image shape feature of items [16]. Han et al. propose to jointly train a Bi-LSTM model and a visual-semantic embedding for fashion outfit recommendation [17].

However, We can know from Table I that there is no tight recommendation models can utilize accurate side information from visual and document information.

Thus, we propose a framework called double deep feature matrix factorization (DDFMF), which is tightly embedded in double CNNs with PMF for document and visual information. We optimize the parameters of PMF and the document and visual information simultaneously to integrate user-item rating into the information seamlessly. We also utilize DeepContour, which is a deep learning-based edge extraction algorithm similar to Carny [18], to extract the visual information. We confirm the accuracy of our proposed method via an experiment using Amazon dataset.

The key contributions of this study are as follows. This work introduces the first approach that models not only user-item rating but also visual and document information tightly. In addition, this work shows the usability of visual information and its context extracted by CNN.

The remainder of this paper is organized as follows. Section II describes the preliminary knowledge about PMF, CNN, and visual features. Section III introduces the graphical model of DDFMF and the architecture of our CNN. Section IV presents the experimental results and model performance evaluation. Section V concludes the study and presents the future works.

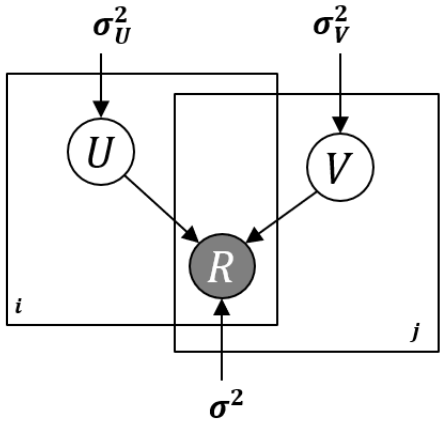

Figure 1. Graphical model of PMF

2. Preliminary

2.1. Probabilistic Matrix Factorization(PMF)

Salakhutdinov et al. [19] proposed the PMF, which is a well-known approach for recommendation systems. Table II summarizes the notations of PMF, and Fig. 1 shows the overview of the graphical model of PMF. We suppose that M users, N items, and a rating matrix $R \in \mathbb{R}^{N \times M}$ exist. Also, we demand the user latent matrix $R \in \mathbb{R}^{k \times N}$ and item latent matrix $R \in \mathbb{R}^{k \times M}$ to reconstruct the rating matrix R. The goal of the PMF is to determine the optimal matrix U. V minimizes the loss function $\mathcal{E}$, as shown as follows:

$$\begin{aligned} \min \mathcal{E}(U,V) = &\sum_{i}^{N}\sum_{j}^{M} \frac{I_{ij}}{2}\left(r_{ij} - u_i^T v_j\right)^2 \\ &+ \frac{\lambda_U}{2}\sum_{i}^{N}\|u_i\|^2 + \frac{\lambda_V}{2}\sum_{j}^{M}\|v_j\|^2 \end{aligned} \tag{1}$$

2.2. Deep Features

In the last few years, deep learning has made significant progress in natural language processing, image processing and object detection, and we have been

Table 2. Notations

Notation	Description
R	Rating matrix
N	Number of users
u_i	Latent factors of user i
M	Number of items
v_j	Latent factors of item j
r_{ij}	Rating of item j given by user i
$\widehat{r}_{ij}$	Predicted rating of item j given by user i
U	User latent factor
V	Item latent factor
k	Size of latent factor
I	Indicator,$I_{ij}=1$ if $r_{ij}\neq 0$,otherwise $I_{ij}=0$
σ^2,σ_U^2,σ_V^2	Variance
S_j	Image shape feature of item j
D_j	Contextual feature of item j
s_j	Complex feature latent factor of item j
W	Internal weights in the CNNs
w_d	Each weight in the CNNs

able to extract high-level features using deep neural networks. Meanwhile, these features are being applied to the recommender system, and these deep features are useful in improving recommendation accuracy [4].

Kim et al. [4] addressed limitations of the bag-of-words model-based approaches and proposed a novel document context-aware recommendation model(ConvMF) that integrates CNN into PMF. This models CNN can capture the contextual meaning of words in documents and can even distinguish the subtle contextual difference of the same word via different shared weights. Duan et al. [16] used the deep feature of the product image to improved the recommendation accuracy of the apparel recommender system. This deep feature is extracted by DeepContour. This method is proposed by Shen et al. [18, 20–22], and compared to the traditional method(Canny [23], Sketch tokens [24]), it can extract more accurate image shape features. They used CNN in learning contour features to improve the accuracy of contour detection. They divided the contour data into subclasses based on the contour shape, thereby converting the contour versus non-contour classification problem into a multi-class classification problem. They also proposed a new loss function called positive sharing loss function. This function focuses on the loss of contours and non-contours rather than the loss of each subclass and helps explore more discriminative features compared with the softmax loss function.

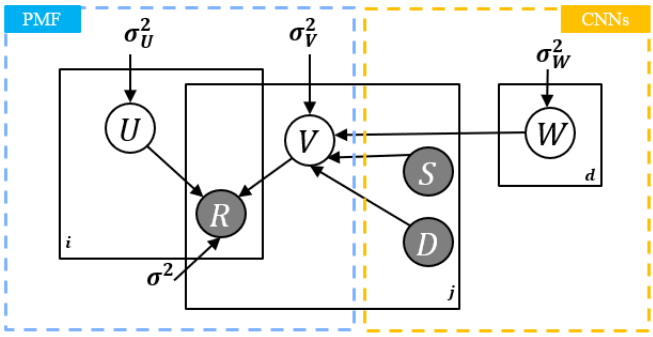

Figure 2. Graphical model of DDFMF

3. Double Deep Features Matrix Factorization(DDFMF)

3.1. Probabilistic Model of DDFMF

DDFMF is modified based on ConvMF [4]. We simultaneously use the contextual information of items and their image shape feature. Table II summarizes the notations of the DDFMF, and Fig. 2 shows the overview of the probabilistic model of the DDFMF. First, suppose that N users,M items, and a user-item rating matrix ($R \in \mathbb{R}^{N\times M}$) exist. In a probabilistic point of view, the conditional probability of the observed rating matrix R is expressed as follows:

$$p\left(R|U,V,\sigma^2\right)=\prod_{i}^{N}\prod_{j}^{M}\left[N\left(r_{ij}|u_i^T v_j,\sigma^2\right)\right]^{I_{ij}} \quad (2)$$

where $N\left(X|\mu,\sigma^2\right)$is a Gaussian distribution of X with mean μ and variance σ^2. Subsequently, R is matrix-decomposed into user latent model $U \in \mathbb{R}^{\boldsymbol{k}\times\boldsymbol{N}}$ and item latent model $V \in \mathbb{R}^{\boldsymbol{k}\times\boldsymbol{M}}$. However, different from the traditional PMF, we assume that our item latent matrix V depends on the following four variables:

These four variables allow the item latent model to be further optimized for the ratings. Therefore, the final item latent model consists of the item latent, image shape latent factors, and contextual latent factors. The item latent factor is obtained by PMF decomposition, and the image shape and contextual latent factor are obtained by our CNNs. The final item latent model is expressed as follows:

$$v_j=\mathrm{cnn}\left(W,S_j,D_j\right)+\varepsilon_j \quad (3)$$

$$\varepsilon_j=N\left(0,\sigma_V^2 I\right) \quad (4)$$

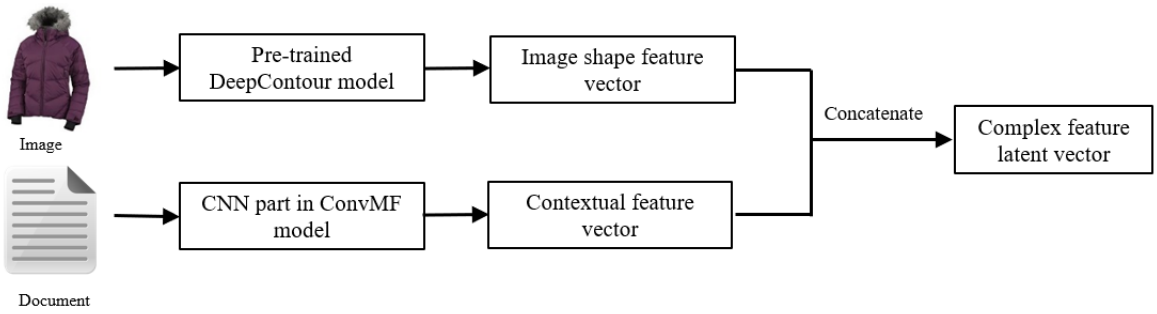

Figure 3. The architecture of our CNNs

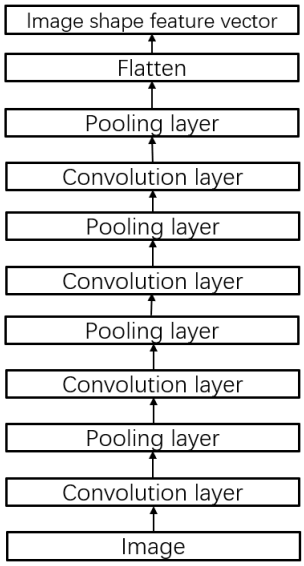

Figure 4. The CNN for extracting image shape feature

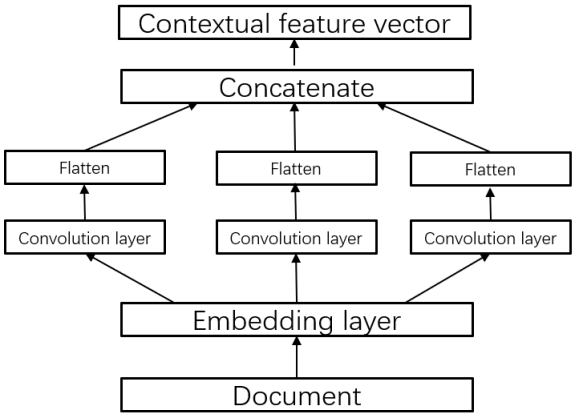

Figure 5. The CNN for extracting contextual feature

where $cnn()$ represents the output of our CNNs architecture. For each weight w_k in W, we place zero-mean spherical Gaussian prior, the most commonly used prior.

$$p\left(W|\sigma_W^2\right) = \prod_d N\left(w_d|0,\sigma_W^2\right) \tag{5}$$

The prior distributions of latent models U and V are

$$p\left(U|\sigma_U^2\right) = \prod_i^N N\left(u_i|0,\sigma_U^2\right) \tag{6}$$

$$p\left(V|W,S,D,\sigma_V^2\right) = \prod_j^M N\left(v_j|\,\mathrm{cnn}\left(W,S_j,D_j\right),\sigma_V^2 I\right) \tag{7}$$

3.2. Architecture of Our CNNs

Our CNNs generate an image shape and contextual latent vectors from the images and documents of items to compose high-precision item latent models with epsilon variables. Fig.3 shows our CNN architecture that comprises two networks. The first CNN is for extracting image shape features. We use transfer learning to place a pre-trained CNN model (DeepContour [18]) into our structure for contour detection. This pre-trained model has four convolutional layers and three fully-connected layers. For our proposed model, we consider the output of the first fully-connected layer. Then, we use a three-channel image patch with a size of 45 $\times$ 45 as input and obtain a 128-dimensional feature vector of the image shape (see Fig.4). The second CNN is for extracting contextual features. We use the CNN model from ConvMF and consider the output of the first fully-connected layer. Then, we can use a raw document as input and obtain a 300-dimensional contextual feature vector (see Fig.5).

Finally, we concatenate the feature vector of the image shape via the contextual feature vector and use a conventional nonlinear projection to obtain a

k-dimensional latent vector of complex features. Our CNN model becomes a function that utilizes the images and documents of item j as input and returns the latent factor of the complex feature of item j as output, as expressed as follows:

$$s_j = \mathrm{cnn}\left(W, S_j, D_j\right) \tag{8}$$

3.3. Optimization

To optimize the variables such as U, V, W, we use maximum a posteriori estimation as follows:

$$\begin{aligned} &\max_{U,V,W}\left(U,V,W|R,S,D,\sigma^2,\sigma_U^2,\sigma_V^2,\sigma_W^2\right) = \\ &\max_{U,V,W}\left[p\left(R|U,V,\sigma^2\right)p\left(U|\sigma_U^2\right)\right. \\ &\qquad\left. p\left(V|W,S,D,\sigma_V^2\right)p\left(W|\sigma_W^2\right)\right] \end{aligned} \tag{9}$$

If we give a negative logarithm on Eq.(9), it can be reformulated as follow:

$$\begin{aligned} \min \mathcal{E}(U,V,W) = &\sum_i^N \sum_j^M \frac{I_{ij}}{2}\left(r_{ij} - u_i^T v_j\right)^2 \\ &+ \frac{\lambda_U}{2}\sum_i^N \|u_i\|^2 \\ &+ \frac{\lambda_V}{2}\sum_j^M \|v_j - \mathrm{cnn}\left(W, S_j, D_j\right)\|^2 \\ &+ \frac{\lambda_W}{2}\sum_d^{|w_d|} \|w_d\|^2 \end{aligned} \tag{10}$$

where λ_U is $\frac{\sigma^2}{\sigma_U^2}$,$\lambda_V$ is $\frac{\sigma^2}{\sigma_V^2}$, and λ_W is $\frac{\sigma^2}{\sigma_W^2}$.

Our model aims to find user latent and item latent models to infer user preferences by minimizing function $E(U,V,W)$. Partially differentiate Eq.(10) with u_i and v_j respectively. Subsequently, by updating u_i and v_j with the coordinate descent method, we can obtain the optimum user latent matrix (U) and the item latent matrix (V). The update formulas for u_i and v_j are shown in Eqs.(11) and (12):

$$u_i = \left(V I_i V^T + \lambda_U I_K\right)^{-1} V R_i \tag{11}$$

$$\begin{aligned} v_j = &\left(U I_j U^T + \lambda_V I_K\right)^{-1}\left(U R_j\right. \\ &\left. + \lambda_V \,\mathrm{cnn}\left(W, S_j, D_j\right)\right) \end{aligned} \tag{12}$$

where I_i is a diagonal matrix with I_{ij} ,j = 1;...;M as its diagonal elements and R_i is a vector with $\left(r_{ij}\right)_{j=1}^M$ for user i. For item j, I_j and R_j are similarly defined as I_i and R_i, respectively.

However, W cannot be optimized as we can do for U and V. Because W is the weights and biases of each layer, and it is closely related to the features in our CNNs architecture such as max-pooling layers and non-linear activation functions. Fortunately, when U, V are temporarily fixed, loss function $\mathcal{E}$ becomes an error function with regularized terms of neural net.

$$\begin{aligned} \mathcal{E}(W) = &\frac{\lambda_V}{2}\sum_j^M \|v_j - \mathrm{cnn}\left(W, S_j, D_j\right)\|^2 \\ &+ \frac{\lambda_W}{2}\sum_d^{|w_d|} \|w_d\|^2 + \text{constant} \end{aligned} \tag{13}$$

To optimize W, we use the back propagation algorithm with given target value v_j (v_j is temporarily fixed). The overall optimization process (U, V and W are alternatively updated) is repeated until convergence. Finally, we can predict rating of user i on item j as follow:

$$\widehat{r}_{ij} = u_i^T v_j = u_i^T\left(\mathrm{cnn}\left(W, S_j, D_j\right) + \varepsilon_j\right) \tag{14}$$

4. Experiment

4.1. Goal, Dataset, Environments, and Criteria

In this experiment, we evaluate our proposed model for apparels. We use the clothes, shoes, and accessories as the category data from the Amazon product dataset [14, 25]. The dataset consists of rating data and the images and documents of items. We pre-processed the dataset for the experiment as follows:

- We divide the dataset into three, namely clothes, shoes and accessories to investigate the performance of our model for different product types.
- We remove the items that do not have their images or description documents from the three datasets.
- The matrix size of the training data is set similarly to that of the original data to divide the training and test data. However, the three original datasets are not used in this experiment because they are large and extremely sparse and cannot be split into training and test data. Consequently, we remove users that have less than two ratings. The statistics of each datum validate that the three datasets possess different characteristics (Table III).

Table 3. Detail Statistics of the Three Datasets

Dataset	Users	Items	Ratings	Density
Clothes	8222	8513	19654	0.0281%
Accessories	3889	3507	8544	0.0626%
Shoes	622	749	1334	0.286%

Table 4. Hyperparameters

Model	Clothes		Accessories		Shoes	
	λ_U	λ_V	λ_U	λ_V	λ_U	λ_V
PMF	0.1	0.1	0.1	0.1	0.1	0.1
ConvMF	1	10	1	10	0.1	10
ISFMF	1	1	1	0.1	1	1
DDFMF	1	1	1	10	1	1

We compare the DDFMF with the following baselines:

- PMF [19]: PMF is a standard rating prediction model for user ratings only.
- ConvMF [4]: This model was introduced by Kim et al. It uses the contextual features of items to improve the rating prediction accuracy.
- ISFMF [16]: This model is that we used transfer learning to integrate DeepContour into the model proposed by Duan et al.

4.2. Experiment Setup

Our experimental environment uses the Keras Python library [26] with NVIDIA GeForce Titan X. The parameter settings are as follows:

- The size of the latent dimension of U and V is set at 50.(we set the same value of U and V for all models)
- The size of the input image is set at 45×45.(same size as the pre-trained Deepcontour model and ISFMF)
- The maximum length of documents and set the size of vocabulary is set at 300 and 8000, respectively.(same value as the convMF)
- Each dataset is randomly split into training, validation,and test sets.

We select the hyperparameters (λ_U, λ_V) via grid search; for different models, these hyperparameters may vary. Table IV lists the best combination of hyperparameters (λ_U, λ_V). For the evaluation measure, we use the root mean squared error (RMSE), as shown as follows:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i,j}^{N,M} \left(r_{ij} - \widehat{r}_{ij}\right)^2}{\text{ratings}}} \tag{15}$$

For the reliability of our results, we repeat the evaluation procedure 10 times and report the mean test errors of each model.

4.3. Experimental Results

Table 5. Overall RMSE Test on Clothes Dataset

Model	Ration of training data		
	70%	80%	90%
PMF	1.646	1.603	1.592
ConvMF	1.327	1.272	1.261
ISFMF	1.21	1.165	1.177
DDFMF	**1.168**	**1.131**	**1.129**
Imp.1	12%	11.10%	10.50%
Imp. 2	3.50%	2.90%	4.10%

Table 6. Overall RMSE Test on Accessories Dataset

Model	Ration of training data		
	70%	80%	90%
PMF	1.724	1.602	1.641
ConvMF	1.459	1.398	1.450
ISFMF	1.363	1.302	1.389
DDFMF	**1.342**	**1.283**	**1.331**
Imp.1	8%	8.2%	8.1%
Imp. 2	1.5%	1.4%	4.1%

Table 7. Overall RMSE Test on Shoes Dataset

Model	Ration of training data	
	80%	90%
PMF	1.882	1.623
ConvMF	1.263	1.261
ISFMF	1.186	1.341
DDFMF	**1.179**	**1.297**
Imp.1	6.7%	0.7%
Imp. 2	0.6%	3.3%

Table 8. Results of T-test on Shoes Dataset (90% training data)

Models	P value
DDFMF versus ConvMF	0.143547 >0.05
DDFMF versus ISFMF	0.009217 <0.05
ISF versus ConvMF	0.028917 <0.05

4.3.1 Results on Clothes Dataset

Table V shows the overall RMSE of the PMF, ConvMF, ISFMF, and DDFMF on the clothes dataset. Imp. 1 and Imp. 2 denote the improvement, in which our model outperforms the ConvMF and the ISFMF, respectively. In comparison with the other three models, DDFMF achieves significant performance on the clothes dataset. The improvement of our model over the best competitor, namely, ISFMF, increases consistently from 4.1% to 2.9%. Using document and visual information simultaneously can produce accurate results. In addition, using visual information alone is better than with document information.

4.3.2 Results on Accessories Dataset

Table VI shows the overall RMSE of the PMF, ConvMF, ISFMF, and DDFMF on the accessories dataset. Imp. 1 and Imp. 2 denote the improvement, in which our model outperforms the ConvMF and the ISFMF, respectively. In comparison with the other three models, DDFMF attains significant performance on the accessories dataset. The improvement of our model over the best competitor, namely, ISFMF, increases consistently from 4.1% to 1.4%. However, when the training set is 90%, the accuracy of all the models decrease.

4.3.3 Results on Shoes Dataset

Table VII shows the overall RMSE of the PMF, ConvMF, ISFMF, and DDFMF on the shoes dataset. Imp. 1 and Imp. 2 represent the improvement, in which our model outperforms the ConvMF and the ISFMF, respectively. The performance of our model on the shoes dataset is poor compared with the other three models, as shown by the p value between the three models (Table VIII). When the training set is 90%, the accuracy of our model is almost the same as that of ConvMF. Moreover, using visual information alone is worse than with document information.

5. Conclusion

In this study, we determine whether the use of image shape and contextual features of items can effectively improve the rating prediction accuracy of an apparel recommendation system.We also propose a novel probabilistic model embedded in double CNNs for document and visual information with PMF. Our experimental results corroborate that the DDFMF significantly outperforms the other competitors. However, our model does not perform well in the shoes dataset. Possibly, the image shape feature is not suitable for this dataset.

For future works, we intend to use different visual information to investigate the accuracy of the recommendation system. [27]

6. Acknowledgement

We gratefully acknowledge the support of JSPS KAKENHI Grant Numbers 16K01250 to use GPU and NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

7. References

References

[1] R.Saga, Y.Hayashi, and H.Tsuji, "Hotel recommender system based on user's preference transition," *In Proceedings of the 2008 IEEE International Conference on Systems, Man and Cybernetic*, pp. 2437–2442, Oct. 2008.

[2] J. Liu, C. Wu, and W. Liu, "Bayesian probabilistic matrix factorization with social relations and item contents for recommendation," *Decision Support Systems*, vol. vol.55, pp. 838–850, June 2013.

[3] Y.-D. Kim and S. Choi, "Scalable variational bayesian matrix factorization with side information," *in Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 493–502, 2014.

[4] D. Kim, C. Park, J. Oh, S. Lee, and H. Yu, "Convolutional matrix factorization for document context-aware recommendation," *in Proceedings of the 10th ACM Conference on Recommender Systems Artificial Intelligence and Statistics*, pp. 233–240, 2016.

[5] J. Liu, D. Wang, and Y. Ding, "Phd: A probabilistic model of hybrid deep collaborative ltering for recommender systems," *in Proceedings of the 9th Asian Conference on Machine Learning*, pp. 224–239, 2017.

[6] H. Wang, N. Wang, and D.-Y. Yeung, "Collaborative deep learning forrecommendersystems," *in Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1235–1244, 2015.

[7] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientic articles," *in Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 448–456, 2011.

[8] Z. S. T. Z. H. L. C. X. S. Liu, J. Feng and S. Yan, "Hi,magic closet, tell me what to wear!," *in Proceedings of the 20th ACM International Conference on Multimedia*, p. 619628, 2012.

[9] X. Y. Y. Hu and L. S. Davis, "Collaborative fashion recommendation:a functional tensor factorization approach," *in Proceedings of the 23rd ACM international conference on Multimedia*, pp. 129–138, 2017.

[10] L. Y. S. Zhang and A. Sun, "Deep learning based recommender system: A survey and new perspectives," *arXiv preprint arXiv:1707.07435*, July 2017.

[11] M. W. H. T. H. Nguyen and L. Schmidt-Thieme, "Personalized tag recommendation for images using deep transfer learning," *in Machine Learning and Knowledge Discovery in Databases*, pp. 705–720, Sept. 2017.

[12] J. T. K. S. S. R. S.Wang, Y.Wang and H. Liu, "What your images reveal: Exploiting visual contents for point-of-interest recommendation," *in Proceedings of the 26th International Conference on World Wide Web*, p. 391400, 2017.

[13] L. B. B. Hidasi, A. Karatzoglou and D. Tikk, "Session-based recommendations with recurrent neural networks," *arXiv preprint arXiv:1511.06939*, 2017.

[14] J. McAuley, C. Targett, Q. Shi, , and A. van den Hengel, "Imagebased recommendations on styles and substitutes," *in Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 43–52, 2015.

[15] R. He and J. McAuley, "Vbpr: Visual bayesian personalized ranking from implicit feedback," *in Proceedings of the Thirtieth AAAI Conference on Articial Intelligence*, pp. 144–150, 2016.

[16] Y. Duan and R. Saga, "Apparel goods recommender system-based image shape features extracted by a cnn," *in Proceedings ofthe 2018 IEEE International Conference on Systems, Man and Cybernetics*, 2018.

[17] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional lstms," *in Proceedings of the 2017 ACM on Multimedia Conference*, pp. 1078–1086, 2017.

[18] W. Shen, X. Wang, Y. Wang, X. Bai, and Z. Zhang, "Deepcontour: A deep convolutional feature learned by positive-sharing loss for contour detection," *in Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3982–3991, June 2015.

[19] R. Salakhutdinov and A. Mnih, "Probabilistic matrix factorization," *in Proceedings of the 20th International Conference on Neural Information Processing Systems*, pp. 1257–1264, 2007.

[20] P. Dollar and C. L. Zitnick, "Structured forests for fast edge detection," *in Proceedings of the 2013 IEEE International Conference on Computer Vision*, pp. 1841–1848, Dec. 2013.

[21] C. L. Zitnick and P. Dollar, "Edge boxes: Locating object proposals from edges," *in Proceedings of the 2014 European Conference on Computer Vision*, pp. 391–405, Sept. 2014.

[22] P. Dollar and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. vol.37,no.8, pp. 1558–1570, Aug. 2015.

[23] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. vol. PAMI8, no. 6, pp. 679–698, Nov. 1986.

[24] J. J. Lim, C. L. Zitnick, and P. Dollar, "Sketch tokens: A learned mid-level representation for contour and object detection," *in Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3158–3165, 2013.

[25] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative ltering," *in Proceedings of the 25th International Conference on World Wide Web*, pp. 507–517, 2016.

[26] F.Choll, "Keras," *https://github.com/fchollet/keras*, 2015.

[27] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," *in Proceedings of the 31st International Conference on International Conference on Machine Learning*, vol. vol.32, pp. I–647–I–655, 2014.