An Item Recommendation Approach by Fusing Images based on Neural Networks

Weibin Lin
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
linweibin@whut.edu.cn

Lin Li
College of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
cathylilin@whut.edu.cn

Abstract—There are rich formats of information in the network, such as rating, text, image, and so on, which represent different aspects of user preferences. In the field of recommendation, how to use those data effectively has become a difficult subject. With the rapid development of neural network, researching on multi-modal method for recommendation has become one of the major directions. In the existing recommender systems, numerical rating, item description and review are main information to be considered by researchers. However, the characteristics of the item may affect the user's preferences, which are rarely used for recommendation models. In this work, we propose a novel model to incorporate visual factors into predictors of peoples preferences, namely MF-VMLP, based on the recent developments of neural collaborative filtering (NCF). Firstly, we get visual presentation via a pre-trained convolutional neural network (CNN) model. To obtain the nonlinearities interaction of latent vectors and visual vectors, we propose to leverage a multi-layer perceptron (MLP) to learn. Moreover, the combination of MF and MLP has achieved collaborative filtering recommendation between users and items. Our experiments conduct Amazon's public dataset for experimental validation and root-mean-square error (RMSE) as evaluation metrics. To some extent, experimental result on a real-world data set demonstrates that our model can boost the recommendation performance.

Index Terms—Recommender System, Neural Network, Matrix Factorization, MLP

I. INTRODUCTION

With the massive amount of different source data generated by online services, including e-commerce, social media applications and online news, recommender systems are playing an increasingly important role among them. Matrix factorization (MF) [12], [14], is one of the most popular collaborative filtering (CF) [22] techniques, using the numerical rating properly for predicting missing ratings. MF method has achieved great success in the Netflix Prize contest. It casts users and items into a shared latent space, using a latent vector to represent users' interests or items' features. Hence, the result of inner product of these two latent vectors means a user's preferences on a item.

Because of the sparsity of user-item interactions, MF method usually suffers from limited performance. Moreover, without explaining the basic principles, a rating only reflects a user's overall satisfaction towards an item. To improve the accuracy of recommender system, hybrid recommendation [5], [6] are proposed, which includes two mainly research lines the

hybridization of algorithms and the hybridization of multi-modal data. For the first research line, researchers attempt to integrate different recommendation methods to improve performance. There are various strategies to combine two methods, such as weighting, switching, mixing, cascading feature combination [6], [8] and so on. The combination of content-based [17] and CF-based algorithms is the most common use for research and application. For the other research line, to a certain extent, using more than one information can improve the prediction accuracy of recommender systems, such as rating plus review or rating plus image [9], [10], [16]. Researchers propose to integrate numerical rating with textual review for recommendation, including topic modeling [2], [23], and neural network approaches for review modeling [32], [33]. In order to fuse the characteristics of multi-source data, multi-view learning is a common solution [26]–[28]. Although researchers use a variety of data to improve recommendation performance or interpretability, most of them have ignored image information. Currently, the picture is mainly used in the field of retrieval [25] and recognition. The appearance of the item is one of the most intuitive and most influential factors of a user's hobby.

Different from the predecessors' works, He et al. [10] integrates image with implicit feedback based on bayesian personalized ranking [19]. In order to reduce the feature dimension of the picture, authors propose to learn an embedding kernel which linearly transforms the high-dimensional features into a much lower-dimensional space. Compared with the rapid development of neural networks, this method not only has a slower training speed, but also the result is not well. Although Zhang et al. [32] used deep neural networks (DNNs) [29], [30] to model auxiliary information, such as visual content of images. When building models for key collaborative filtering effects, they still use MF method, using an inner product to combine the latent vectors of the users and the items. However, we use neural networks to train model with non-linear method.

In our model, we use explicit feedback as the input, which represents the user's overall satisfaction with the item. The rmse is an evaluation index of our model to predict the missing ratings and train the model. First of all, we extract the visual representations by a pre-trained convolutional neural network (CNN) [24] named ImageNet. Secondly, image features are

combined with traditional MF method, which can be deemed as a linear transformation model to learn features. However, He et al. [11] use a MLP model to replace the interaction function and show a well performance gains over traditional methods such as MF. Unlike it's approach, we consider the visual features and plus the features into the model. The last job is to fuse the MF method and MLP, taking into account the image characteristics. The experiments on authoritative dataset like clothing from Amazon.com show that our model outperforms baseline in prediction accuracy. Specifically, our main contributions are listed as follows:

- We propose a recommended method for fusing images and ratings.
- We propose a suitable model to incorporate image into predictors of people's preferences based on NCF framework.

In the rest of the paper, we first review the related work in Section 2, and present the proposed model in Section 3. Section 4 gives the experimental results and analysis, and Section 5 concludes the paper.

II. RELATED WORK

As we all know, collaboration filtering (CF) [22] algorithm is the most commonly used algorithm in recommendation system. It is divided into three categories, including user-based, item-based and model-based. Among these recommendation algorithms, the model-based recommendation algorithm has become the mainstream of research for a long time. It mainly uses the existing feedback data, such as click, rating, etc, to learn the parameters of model, and then uses this model to predict and recommend. However, MF approach is the most important one in the model-based CF methods. Koren et al. [14] proposes the MF approach based on latent factor model (LFM). The algorithm has an excellent performance in the Netflix competition, which provides a reference for future works.

Because of the sparsity of user-item interactions, CF method usually suffers from limited performance. User-based model is good for recommending popular items in a wide range of interests, but lacks personalization. Item-based model is able to discover long tail items from the users' personal interests, but lacks diversity. From this point of view, each method has its own advantages and disadvantages. In order to combine the benefits of various methods, hybrid model is proposed to make use of different information, including text, context, social [2], [15], [18], and so on.

The recommended tasks for CF method are mainly divided into two types, one is to predict missing values, and the other is to recommend a short list of items to users named top-k. In the early days, although literature on recommendations primarily focus on explicit feedback [21], recent attention has increasingly turned to implicit data [3], [12]. Depending on the data format used, it can be roughly divided into two ways, which are using implicit feedback to predict the recommended list and using explicit feedback to predict missing scoring values. In our model, we select the latter one.

With the development of neural network, to make use of deep learning, many recent researches have developed non-linear neural network models for CF method [1], [4], [7], [11], [31]. In particular, instead of using fixed interaction function (i.e., inner product) in MF, He et al. [11] propose a neural collaborative filtering (NCF) framework to learn the user-item interaction function from data. And then, they propose a model NCF model named NeuMF, which fuse the MF and MLP to learn the interaction function. After that, based on the NCF framework, Bai et al. [1] incorporate the neighborhoods of users and items, Cheng et al. [7] model aspects in textual reviews, and so on.

The appearance of item, such as clothes, is one of the most important factors to influence user's opinions to item. However, in the past, image factors have been ignored because feature extraction methods are failed to achieve effective performance in visual machine learning. However, while deep learning technology has been greatly improved, we can extracted a visual feature from a pre-trained convolutional neural network (CNN) effectively to represent the latent feature of image. Based on the MF method, He et al. [10] propose visually-aware recommender systems. They embedding the high-dimensional image into low-dimensional image by a pre-trained CNN model, which can extract the high performance visual features to use. In the neural network model, they seldom consider the incorporation of rating and image. As a comparison, our focus is to consider visual features based on NCF framework for rating prediction.

III. METHODOLOGY

In this section, we introduce three models used in this paper in turn and explain how to generate our models. In order to fuse image and numerical rating for recommendation, we propose the joint model. Table I summarizes the key notations used in the models.

TABLE I
NOTATIONS

Notation	Explanation
U, V	User set ($|U| = n$), Item set ($|V| = m$)
R	Rating matrix ($n \times m$)
y_{ui}	Rating assigned by user u to item i
K	Number of latent features in MF
p_u, q_i, v_i	Latent factors of user u, item i, and image i, respectively
λ_u, λ_v	Regularization parameters
$\varnothing_x$	x-th neural network layer
z_i	New item representation

A. Visual Matrix Factorization (VMF)

MF method relates each user and item to a real-valued vector of latent features. Given n users and m products, to decompose rating matrix $R \in \mathbb{R}^{n \times m}$ into two low rank-K matrices $U \in \mathbb{R}^{K \times n}$ and $V \in \mathbb{R}^{K \times m}$, is the goal of MF approach. Let p_u and q_i denote the latent vector for user us and item i, respectively. MF estimates an interaction y_{ui} as the inner product of p_u and q_i:

$$\hat{y}_{ui} = p_u^T q_i = \sum_{k=1}^{K} p_{uk} q_{ik} \qquad (1)$$

Where K denotes the dimension of the latent space. However, The basic MF approach has a over-fitting problem. The regularized MF method add a normalization factor to the loss function as follows:

$$\min_{U,V} \frac{1}{2} \sum_{u=1}^{n} \sum_{j=1}^{m} (y_{ui} - p_u^T q_i)^2 + \frac{\lambda_u}{2} ||U||_F^2 + \frac{\lambda_v}{2} ||V||_F^2 \qquad (2)$$

λ_u and λ_v are regularization parameters for user embedding matrix U and item embedding matrix V, respectively, and $|| \cdot ||_F^2$ denotes the Frobenius norm. The gradient descent-based optimization technique is generally applied to find the local minimum solution for Eq. (2).

Considering that image characteristics can affect user preferences, users who buy clothes will care about its style, color, etc. In order to improve the accuracy of the recommended model, we have added image parameters. The newly loss function as follows:

$$\min_{U,V} \frac{1}{2} \sum_{u=1}^{n} \sum_{j=1}^{m} (y_{ui} - p_u^T q_i - \theta_u^T \theta_i)^2$$
$$+ \frac{\lambda_u}{2} ||U||_F^2 + \frac{\lambda_v}{2} ||V||_F^2 \qquad (3)$$

θ_u and θ_i are newly introduced D-dimensional visual factors whose inner product models the visual interaction between u and i. To some extent, the user u pays attention to D visual dimensions. How to use image features effectively in a model has become a research problem. A simple method is to extract features directly from Deep CNN as item features θ_i. However, due to the extracted image features are too high in dimension, the results of the model are unsatisfied. Another approach is to reduce the dimensions of image features. Although some reduction techniques like PCA and CCA are possible to solve this problem, experiments show that these methods lose too much useful information of the original features. Nevertheless, we adopt an embedding kernel [10] to linearly transforms such high-dimensional features into a much lower-dimensional space:

$$\theta_i = E f_i \qquad (4)$$

Here E is a D $\times$ F matrix embedding Deep CNN feature space (F-dimensional) into visual space (D-dimensional), where f_i is the original visual feature vector for item i.

B. Visual Multi-Layer Perceptron (VMLP)

Our model VMLP is based on NCF model, which combine the several pathways to model users and items. As we can see from the first layer of the model in Fig. 1, original visual features has extracted from pre-trained CNN model. At the same time, we use the look-up layer to project the one-hot

Fig. 1. VMLP model

input of user and item into low-dimensional embedding. The same as VMF, we need to reduce the dimensional of original image features. To address this issue, we propose to add hidden layers on the concatenated vector. In order to integrate the latent vectors of items and image vector, we concatenate these vectors into item enhanced factor. However, to learn and predict users' preferences for items, we use a standard MLP model to train parameters. Unlike the VMF model, we can endow the model a large level of flexibility and nonlinearity to learn the interactions between users and items, rather than a fixed element-wise method. Precisely, the VMLP model is defined as follows

$$z_1 = a_1(p_u, q_i)$$
$$\phi_2(z_1) = a_2(W_2^T z_1 + b_2)$$
$$... \qquad (5)$$
$$\phi_L(z_{L-1}) = a_L(W_L^T z_{L-1} + b_L)$$
$$\hat{y}_{ui} = \delta(h^T \phi_L(z_{L-1}))$$

Where p_u, q_i, W_l, b_l, and a_l represent users' factors, items' enhanced factors, the weight matrix, bias vector, and activation function for the l-th layer's perceptron, respectively. The latent vector of user and item are fed into a multi-layer neural architecture, which we term as neural network layers, to map the latent vectors to prediction scores. Each layer can be customized to discover certain latent structures of user-item interactions. The dimension of the last hidden layer X determines the model's capability. The final output layer is the predicted score $\hat{y}_{ui}$, and training is performed by minimizing the point-wise loss between $\hat{y}_{ui}$ and its target value y_{ui}.

C. Fusion of MF and VMLP (MF-VMLP)

Till this moment, we have proposed two model of VMF and VMLP. Through above introduction, VMF uses a linear kernel to model the latent feature interactions, and VMLP applies a nonlinear kernel to learn the interaction function from data. And then, an idea is born that how can we fuse the MF and

Fig. 2. MF-VMLP model

VMLP based on NCF framework, so as to learn the user-item interactions better. There are two possible ways to solve the issue.

Firstly, one of the easiest ways to work is to share the same input and embedding layers between them, and then combine the outputs of their interaction functions. However, the performance of the fused model might be limited by sharing embedding layers. Once sharing embedding layers, MF and VMLP should use the same size of embedding. While MF and VMLP have their own optimal effect, it's most likely that the embedding layers of both of them are different in size. So this solution is no feasible.

Secondly, to solve the different size from above problem, we let MF and VMLP have their own embedding layers, and concatenate their last hidden layer for combining the two methods. A layer of fully connected layer is added. The model we propose is named MF-VMLP, whose formulations are given as follows

$$
\begin{aligned}
\phi^{MF} &= p_u \odot q_i \\
\phi^{VMLP} &= a_L(W_L^T(a_{L-1}(...a_2(W_2^T \begin{bmatrix} p_u^V \\ q_i^V \end{bmatrix} \\
&\quad + b_2)...)) + b_L) \\
\hat{y}_{ui} &= \delta(h^T \begin{bmatrix} \phi^{MF} \\ \phi^{VMLP} \end{bmatrix}
\end{aligned}
\tag{6}
$$

Where p_u and p_u^V represent the user embedding for MF and VMLP, respectively, and q_i and q_i^M denote item embeddings. Among several activation functions, we choose the ReLU for MLP layers.

IV. EXPERIMENTS

In this section, we introduce the dataset and perform the validation results of several models in various datasets. Theoretically, visual appearance is expected to have an influence on users' decision-making process.

TABLE II
DATASET STATISTICS (AFTER PREPROCESSING)

Dataset	#users	#items	#feedback
Amazon Women	99,748	331,173	854,211
Amazon Men	34,212	100,654	260,352
Amazon Phones	113,900	192,085	964,477

A. Datasets

The datasets we use are from Amazon.com which introduced by McAuley et al. [16]. We select two categories that are proven to be useful in visual features, named Women's and Men's Clothing. In addition, we also consider Cell Phones, which are expected to paly a little role for models. Each dataset has been process by extracting implict feedback and visual features. We only use the data of user u who has scored more than four items. Table II shows the datasets.

B. Visual Features

Recall that each item i is associated with a visual feature vector, denoted by v_i. To set v_i, we use a pre-trained approach to generate visual features from raw product images using the deep learning framework of CAFFE [13]. Following [10], we adopt the CAFFE reference model with five convolutional layers followed by three fully-connected layers that has been pre-trained on 1.2 million IMAGENET [20] images. For item i, the second fully-connected layer is taken as the visual feature vector f_i, which is a feature vector of length 4096.

C. Evaluation Methodology

We split our dataset into three parts, such as training, validation and test sets. The article uses the root-mean-square error (RMSE) to evaluate the model. The formula is as follows

$$
RMSE = \sqrt{\frac{\sum_{u,i \in TestSet}(r_{u,i} - \hat{r}_{u,i})^2}{TestSet}}
\tag{7}
$$

Where r_{ui}, $\hat{r}_{ui}$ and TestSet denote real score, prediction score and test set, respectively. As we all know, the lower the rmse, the higher the accuracy of the model.

TABLE III
RMSE ON THE TEST SET

Dataset	MF	VMF	VMLP	MF-VMLP	improvement
Amazon Women	1.1303	1.0886	1.0639	1.0575	7.3%
Amazon Men	1.0579	1.0287	1.0193	1.0034	5.2%
Amazon Phones	1.0883	1.0794	1.0854	1.0879	0%

D. Result Analysis

We can find some details from Table 2. Toward these three datasets, the results show image features affect the users' preferences to some extent. The hybrid recommendation model MF-VMLP gets the best results, it's predicted performance is 7.3% higher than MF in Amazon Women data. Visual features show greater benefits on clothing than cellphone datasets. We

Fig. 3. Experimental comparison

consider that the functional items are not much different in appearance, such as phones, which play a little role on the model. More importantly, neural networks have a large impact on the models.

E. Conclusion And Future Work

The recommendation system combined with deep learning has become a hot research topic at present. With the rapid development of deep learning, image information becomes more and more important. Visual features influence many of the choices people make. In this paper, we proposed a suitable model to fuse image features and ratings for recommendation. For future research work, due to research on recommendation methods for multi-modal data fusion is also a research hot-pot, we will consider more side information to improve the accuracy of recommendation model.

REFERENCES

[1] Ting Bai, Ji-Rong Wen, Jun Zhang, and Wayne Xin Zhao. A neural collaborative filtering model with interaction-based neighborhood. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1979–1982. ACM, 2017.

[2] Yang Bao, Hui Fang, and Jie Zhang. Topicmf: Simultaneously exploiting ratings and reviews for recommendation. In *AAAI*, volume 14, pages 2–8, 2014.

[3] Immanuel Bayer, Xiangnan He, Bhargav Kanagal, and Steffen Rendle. A generic coordinate descent framework for learning from implicit feedback. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1341–1350. International World Wide Web Conferences Steering Committee, 2017.

[4] Alex Beutel, Paul Covington, Sagar Jain, Can Xu, Jia Li, Vince Gatto, and Ed H Chi. Latent cross: Making use of context in recurrent recommender systems. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pages 46–54. ACM, 2018.

[5] Robin Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[6] Robin Burke. Hybrid web recommender systems. In *The adaptive web*, pages 377–408. Springer, 2007.

[7] Zhiyong Cheng, Ying Ding, Xiangnan He, Lei Zhu, Xuemeng Song, and Mohan S Kankanhalli. Aˆ3ncf: An adaptive aspect attention model for rating prediction. In *IJCAI*, pages 3748–3754, 2018.

[8] Asela Gunawardana and Christopher Meek. A unified approach to building hybrid recommender systems. In *Proceedings of the third ACM conference on Recommender systems*, pages 117–124. ACM, 2009.

[9] Ruining He and Julian McAuley. Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering. In *proceedings of the 25th international conference on world wide web*, pages 507–517. International World Wide Web Conferences Steering Committee, 2016.

[10] Ruining He and Julian McAuley. Vbpr: Visual bayesian personalized ranking from implicit feedback. In *AAAI*, pages 144–150, 2016.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering. In *Proceedings of the 26th International Conference on World Wide Web*, pages 173–182. International World Wide Web Conferences Steering Committee, 2017.

[12] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 549–558. ACM, 2016.

[13] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014.

[14] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[15] Zhongqi Lu, Zhicheng Dou, Jianxun Lian, Xing Xie, and Qiang Yang. Content-based collaborative filtering for news topic recommendation. In *AAAI*, pages 217–223, 2015.

[16] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52. ACM, 2015.

[17] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[18] Zhi Qiao, Peng Zhang, Yanan Cao, Chuan Zhou, Li Guo, and Binxing Fang. Combining heterogenous social and geographical information for event recommendation. In *AAAI*, volume 14, pages 145–151, 2014.

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

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[21] Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton. Restricted boltzmann machines for collaborative filtering. In *Proceedings of the 24th international conference on Machine learning*, pages 791–798. ACM, 2007.

[22] Xiaoyuan Su and Taghi M Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in artificial intelligence*, 2009, 2009.

[23] Yunzhi Tan, Min Zhang, Yiqun Liu, and Shaoping Ma. Rating-boosted latent topics: Understanding users and items with ratings and reviews. In *IJCAI*, pages 2640–2646, 2016.

[24] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions: Robust landmark retrieval. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 79–88. ACM, 2015.

[25] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Transactions on Image Processing*, 26(3):1393–1404, 2017.

[26] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, Qing Zhang, and Xiaodi Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, 2015.

[27] Yang Wang, Lin Wu, Xuemin Lin, and Junbin Gao. Multiview spectral clustering via structured low-rank matrix factorization. *IEEE transactions on neural networks and learning systems*, (99):1–11, 2018.

[28] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, Meng Fang, and Shirui Pan. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. *arXiv preprint arXiv:1608.05560*, 2016.

[29] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE transactions on cybernetics*, (99):1–12, 2018.

[30] Lin Wu, Yang Wang, Ling Shao, and Meng Wang. 3-d personvlad: Learning deep global representations for video-based person reidentification. *IEEE transactions on neural networks and learning systems*, 2019.

[31] Yao Wu, Christopher DuBois, Alice X Zheng, and Martin Ester. Collaborative denoising auto-encoders for top-n recommender systems. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, pages 153–162. ACM, 2016.

[32] Wei Zhang, Quan Yuan, Jiawei Han, and Jianyong Wang. Collaborative multi-level embedding learning from reviews for rating prediction. In *IJCAI*, pages 2986–2992, 2016.

[33] Lei Zheng, Vahid Noroozi, and Philip S Yu. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 425–434. ACM, 2017.