CNN-FM: Personalized Content-Aware Image Tag Recommendation

Hanh T. H. Nguyen, Martin Wistuba, Lucas Rego Drumond and Lars Schmidt-Thieme

Abstract Social media services allow users to share and annotate their resources freely with keywords or tags that have valuable information to support organizing or searching uploaded images or videos. Tag recommendation is used to encourage users to annotate their resources. Recommending tags of images to users not only depends on user preference but also strongly relies on the contents of images. In this paper, we propose a method for image tag recommendation using both image visual features and user past tagging behaviours by combining convolutional neural networks (CNN), which are widely used and have achieved high performance in image classification and recognition, and factorization machines (FM), since factorization models are the state-of-the-art approach for tag recommendation. Empirically, we demonstrate that learnable

Hanh T. H. Nguyen
University of Hildesheim, Universitätsplatz 1,
✉ nthhanh@ismll.de

Martin Wistuba
University of Hildesheim, Universitätsplatz 1,
✉ wistuba@ismll.de

Lucas Rego Drumond
University of Hildesheim, Universitätsplatz 1,
✉ ldrumond@ismll.de

Lars Schmidt-Thieme
University of Hildesheim, Universitätsplatz 1,
✉ schmidt-thieme@ismll.de

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING
Vol. 2, No. 1, 2017

DOI 10.5445/KSP/1000058749/16
ISSN 2363-9881

features extracted by CNNs can improve up to 7 percent the performance of FMs in image tag recommendation.

1 Introduction

The development of computer and network technologies promotes the explosion of social media sharing systems with a large amount of digital resources stored, shared and accessed by users around the world through the Internet. The most popular media sharing site is Flickr with more than 3.5 million new uploaded images daily and 87 million registered members in March 2013[1]. To assist organization and later retrieval of images or increasing the access of the community, users are freely able to assign keywords, called tags, to shared resources (Ames and Naaman, 2007).

Although people can easily enter their own keywords, a considerable number of shared resources has a few or no tags, because tagging is a time consuming task that discourages users from annotating their resources. As mentioned by Sigurbjörnsson and Van Zwol (2008), around 64 percent of the photos have 1 to 3 tags and 20 percent have no tags at all. Tag recommendation systems are used to facilitate the tagging task by suggesting relevant tags. These systems can be dependent or independent of authors who tag images. Because different users like to use their own words describing items in their way, it is practical to suggest a personalized list containing their "favorite" tags to assign to an item. Besides that, users often annotate their images with vocabularies relating to the content or context of images (Sigurbjörnsson and Van Zwol, 2008). Therefore, the visual information is an important factor that is able to help to improve the performance of tag recommendation.

In this paper, we propose a personalized tag recommendation model using the visual content of images and users' historical tagging information to evaluate the capability of image features in enhancing recommendation performance. Our approach firstly applies the convolutional neural network to extract image features. Then, these features are fed to the factorization machine to suggest a ranked list of tags that is built upon pairs of users and images.

[1] http://www.theverge.com/2013/3/20/4121574/flickr-chief-markus-spiering-talks-photos-and-marissa-mayer

2 Related work

There is a vast literature on tag recommendation. In the view of content-based recommendation, Li and Wang (2008) propose a model that learns a mapping between visual features of images and concepts and then predicts the tags associated with the concepts. In Li et al (2008), the authors propose a neighbor voting algorithm to accumulate the relevant scores from the votes of similar images. Tagging photos using the users' vocabularies approach collects tags of all neighbors of an input image according to GPS, time and visual features from tagging historical information of the image's owner and selects the most frequent tags from the list to suggest to him (Qian et al, 2013). Another approach is based on collective knowledge to suggest tags to users (Sigurbjörnsson and Van Zwol, 2008). For each user-provided tag, a list of correlated words are selected by measuring the similarity between them and the initial tags based on a co-occurrence metric. Then, all lists are aggregated to generate the expected list. Garg and Weber (2008) combine personalized suggestion and global tag co-occurrence to narrow down the recommendation for individuals.

Factorization models applied for tag recommendation show good performance. The Pairwise Interaction Tensor Factorization (PITF) (Rendle and Schmidt-Thieme, 2010) models all pairwise interactions between users, items and tags. The Factorization Machine (FM) proposed by Rendle (2010) takes advantages of feature engineering flexibility and strong predicting capability of factorization to suggest a ranked list of tags for a user.

The Convolution Neural Network (CNN), a strong model for image classification and recognition, is also applied in image annotation (Gong et al, 2013; Wei et al, 2014) in terms of solving multilabel classification. However, these kinds of methods do not consider personalization when suggesting labels for images. They learn parameters of the predictors based on pairwise or Weighted Approximate Ranking (WARP) loss functions (Gong et al, 2013) or predict labels from arbitrary trained objects (Wei et al, 2014).

3 The proposed model

In our approach for personalized content-aware tag recommendation, we aim at taking advantages of the learning features capability of the CNN and the

effective ability to suggest tags of the FM. We recall the basic principles of CNNs and FMs and then present the details of our proposed model.

3.1 Problem formulation

To formulate the personalized tag recommendation, we use the notation stated by Rendle and Schmidt-Thieme (2010). A social image tagging system consists of a set of users U, a collection of images I and a set of tags T. The set of all observed assignments of tags to images by users is denoted by $S \subseteq U \times I \times T$.

In the case of personalized tag recommendation, each user-image tuple (u,i) indicating a post associates to a set of tags $T_{u,i} := \{t \in T | (u,i,t) \in S\}$. All observed posts P_S are defined as $P_S := \{(u,i) | \exists t \in T : (u,i,t) \in S\}$.

Solving the tag recommendation problem of a post (u,i) means dealing with a ranking problem to produce an ordered tag list for this post. The recommendation model learns a scoring function $\hat{y} : U \times I \times T \to \mathbb{R}$ for each triple (u,i,t).

In this paper, we are interested in RGB square images that have dimension $d \times d$. The collection of images is defined as $I := \{I_i | I_i \in \mathbb{R}^{d \times d \times 3}\}$ and the visual features of the i-th image I_i are coded as a vector $\mathbf{z}_i \in \mathbb{R}^m$. Now, the prediction model that computes the scores of all tags based on the information of the user u and the visual features $\mathbf{z}_i$ is defined by $\hat{y}(u,i,t) : U \times \mathbb{R}^m \times T \to \mathbb{R}$.

The tag recommendation set $\hat{T}_{u,i}$ for a post (u,i) contains the top-K tags retrieved from the list of all tags that is sorted in descending order of their predicted scores (Marinho et al, 2012).

$$\hat{T}_{u,i} := \underset{t \in T}{\overset{K}{\operatorname{argmax}}} \hat{y}(u,i,t)$$

3.2 Factorization machines

Rendle (2010) combines the advantages of factorization models and feature engineering in one model that can be applied for classification, regression and ranking problems. In the case of tag recommendation, the input of a FM model is a sparse vector $\mathbf{x} \in \mathbb{R}^{|U| \times |I| \times |T|}$ describing the triplet (u,i,t).

$$\mathbf{x}_{u,i,t} = \left(\underbrace{0,\ldots,\overbrace{1}^{u},\ldots,0}_{|U|}, \underbrace{0,\ldots,\overbrace{1}^{i},\ldots,0}_{|I|}, \underbrace{0,\ldots,\overbrace{1}^{t},\ldots,0}_{|T|} \right) \quad (1)$$

The scoring function of the FM model is formulated as:

$$\hat{y}(u,i,t) = f(\mathbf{x}_{u,i,t}) = w_0 + \sum_{j=1}^{p} x_j w_j + \sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} x_j x_{j'} \sum_{f=1}^{q} v_{j,f} v_{j',f} \quad (2)$$

where $p = |U| + |I| + |T|$, the global bias $w_0 \in \mathbb{R}$, the user/image/tag bias $w_j \in \mathbb{R}$, the interaction between the j-th and j'-th variables $\langle \mathbf{v}_j, \mathbf{v}_{j'} \rangle$. Each interaction variable $\mathbf{v}_j$ is a q dimensional vector and can be interpreted as a row of a matrix of factors $\mathrm{V} \in \mathbb{R}^{p \times q}$.

3.3 Convolutional neural networks

CNN (LeCun et al, 1998) is a special kind of neural networks that is able to represent a high-level abstraction of image features and demonstrates the strong capability to classify images on multiclass datasets such as MNIST (Wan et al, 2013), CIFAR-10/100 (Lin et al, 2013) or ImageNet (Krizhevsky et al, 2012). The model involves one or more convolutional layers generating 2-dimensional feature maps by sliding learnable filters across the input volume. The subsampling layers following after the convolutional layers pool a rectangular block of previous layers to produce an output. After several convolutional layers, the output of a CNN is a dense feature vector describing the input image. For the purpose of this work, a CNN can be seen as a non-linear function $\mathbf{z}_i = f_{\mathrm{cnn}}(I_i) : \mathbb{R}^{d \times d \times 3} \rightarrow \mathbb{R}^m$. In the case of multi-class classification, several fully-connected layers may follow the last convolutional or pooling layer to identify a relevant label for the input image.

3.4 Convolutional neural network feature factorization machines

In our approach, we aim at taking advantage of the FM that allows encoding any additional data to extend the input features and the strong capability of a CNN to learn visual contents of images. The model architecture is described by Fig. 1. Candidate tags are recommended for a post (u,i) according to historical tagging

information of the user u and visual contents of image i. For each pair (u,i), the image i is passed through a CNN to transform a 3-dimensional matrix to a feature vector $\mathbf{z} \in \mathbb{R}^m$. Then, values of this vector are fed to a FM to calculate all scores of tags directly or indirectly. There are two different approaches for computing the scores.

1. Direct way: The model applied in this case is notated **CNN-FM**. The values of the vector $\mathbf{z}$ replace the part describing image i in the input vector $\mathbf{x}$ of the FM used to predict the scores of tags.

$$\mathbf{x}_{u,i,t} = \left(\underbrace{0,\ldots,\overbrace{1}^{u},\ldots,0}_{|U|}, \overbrace{\underbrace{z_1,z_2,\ldots,\ldots,z_m}_{m}}^{i}, \underbrace{0,\ldots,\overbrace{1}^{t},\ldots,0}_{|T|} \right)$$

   The scoring function in Equation 2 becomes:

$$\hat{y}(u,i,t) = w_0 + w_u + \sum_{j=1}^{m} z_j \cdot w_{|U|+j} + w_{|U|+m+t} + \underbrace{\sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} x_j x_{j'} \langle \mathbf{v}_j, \mathbf{v}_{j'} \rangle}_{\Psi_{u,i,t}} \quad (3)$$

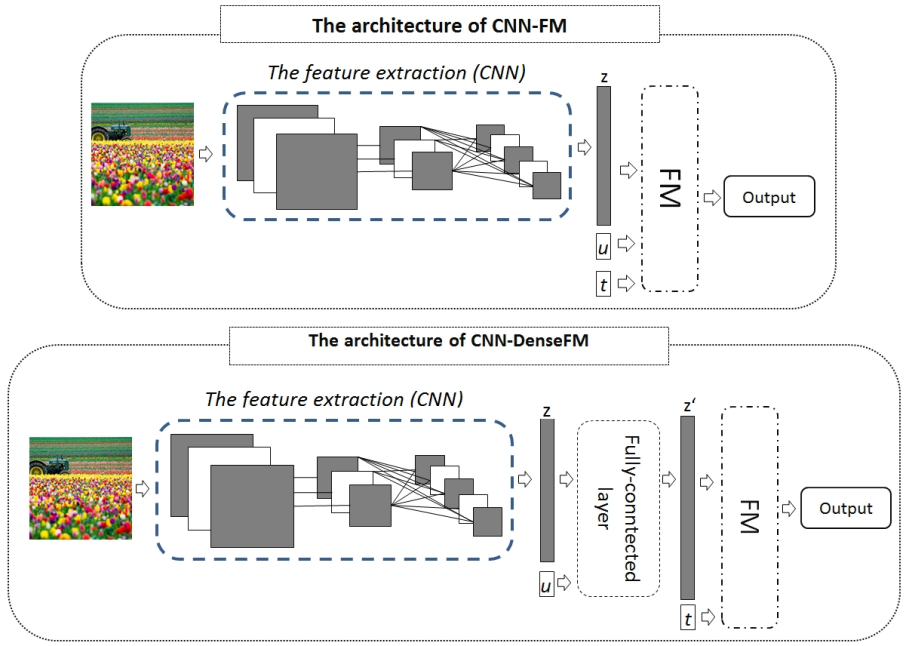
   where $\Psi_{u,i,t} = \Psi_{u,i} + \Psi_{u,t} + \Psi_{i,t} + \Psi_i$ is the interaction between all variables. More specifically:

   - $\Psi_{u,i} = \sum_{j=1}^{m} z_j \langle \mathbf{v}_u, \mathbf{v}_{|U|+j} \rangle$ models the interaction between the user and the visual image content.
   - $\Psi_{u,t} = \langle \mathbf{v}_u, \mathbf{v}_{|U|+m+t} \rangle$ describes the interaction between a user and a tag.
   - $\Psi_{i,t} = \sum_{j=1}^{m} z_j \langle \mathbf{v}_{|U|+j}, \mathbf{v}_{|U|+m+t} \rangle$ models the interaction between visual features and a tag.
   - $\Psi_i = \sum_{j=1}^{m-1} \sum_{j'=j+1}^{m} z_j z_{j'} \langle \mathbf{v}_{|U|+j}, \mathbf{v}_{|U|+j'} \rangle$ models the interaction between the j-th and the j'-th feature of image.

2. Indirect way: The model applied in this case is called **CNN-DenseFM**. The vector $\mathbf{z}$ and a sparse vector representing user u are combined into a vector $\mathbf{h}$ that is passed through a fully-connected layer to extract the interaction features of u and visual feature $\mathbf{z}$.

$$\mathbf{h}_{u,i} = \left(\underbrace{0,\ldots,\overbrace{1}^{u},\ldots,0}_{|U|}, \overbrace{\underbrace{z_1,z_2,\ldots,\ldots,z_m}_{m}}^{i} \right)$$

Fig. 1 The architecture of CNN-FM and CNN-DenseFM



The output of this layer is a dense vector $\mathbf{z}' \in \mathbb{R}^{m'}$ that is defined as a nonlinear function $\mathbf{z}' = f(W^T\mathbf{h}) : \mathbb{R}^{|U|+m} \rightarrow \mathbb{R}^{m'}$ where $W \in \mathbb{R}^{(|U|+m)\times m'}$ and it is provided as a part of the FM input.

$$\mathbf{x}_{u,i,t} = \left(\underbrace{z'_1, z'_2, \ldots, \ldots, z'_{m'}}_{m'}, \underbrace{0, \ldots, \overbrace{1}^{t}, \ldots, 0}_{|T|}\right)$$

Similarly, the scoring function is derived from Equation 2:

$$\hat{y}(u,i,t) = w_0 + \sum_{j=1}^{m'} z'_j \cdot w_j + w_{m'+t} + \underbrace{\sum_{j=1}^{p-1} \sum_{j'=j+1}^{p} x_j x_{j'} \langle \mathbf{v}_j, \mathbf{v}_{j'} \rangle}_{\Psi_{u,i,t}} \tag{4}$$

where the interaction between all variables $\Psi_{u,i,t}$ is the combination between $\Psi_{(u,i),t}$ and $\Psi_{(u,i)}$:

- $\Psi_{(u,i),t} = \sum_{j=1}^{m'} z'_j \langle \mathbf{v}_j, \mathbf{v}_{m'+t} \rangle$ describes the interaction the feature of (u,i) and a tag.
- $\Psi_{(u,i)} = \sum_{j=1}^{m'-1} \sum_{j'=j+1}^{m'} z'_j z'_{j'} \langle \mathbf{v}_j, \mathbf{v}_{j'} \rangle$ models the interaction between the j-th and the j'-th feature of a post (u,i).

3.5 Optimization

The parameters of the prediction model are learned by optimizing the Bayesian Personalized Ranking (BPR) optimization criterion (Rendle et al, 2009). According to this approach, the criterion with respect to the predicted scores of relevant and irrelevant tags is defined as:

$$\text{BPR-OPT}(\hat{\theta}) := \sum_{(u,i) \in P_{Strain}} \sum_{t^+ \in T^+_{u,i}} \sum_{t^- \in T^-_{u,i}} \ln \sigma(\hat{y}(u,i,t^+) - \hat{y}(u,i,t^-)) \quad (5)$$

where the set of tags that the user u has assigned to the image i is denoted as $T^+_{u,i} := \{t \in T | (u,i,t) \in S_{train}\}$ and the set of unobserved tags is defined as $T^-_{u,i} := \{t \in T | (u,i) \in P_{S_{train}} \wedge (u,i,t) \notin S_{train}\}$.

A stochastic gradient descent algorithm adapted by Rendle and Schmidt-Thieme (2010) associated with a back-propagation algorithm in the CNN is applied in the training process where the gradient used for updating parameters is calculated per quadruples (u,i,t^+,t^-) drawn randomly from the training set.

In the case of the CNN-DenseFM, the difference between the relevant and irrelevant tags in Equation 5 is presented as:

$$\hat{y}(u,i,t^+) - \hat{y}(u,i,t^-) = (w_{t^+} - w_{t^-}) + (\Psi_{u,t^+} - \Psi_{u,t^-}) + (\Psi_{i,t^+} - \Psi_{i,t^-}) \quad (6)$$

The interactions including user-image and image-image and the strength of user and image features in the predictor vanish during the optimization process. For this reason, the scoring function from Equation 3 is shortened with less parameters:

$$\hat{y}(u,i,t) = w_t + \langle (\mathbf{v}_u + \sum_{j=1}^{m} z_j \mathbf{v}_{|U|+j}), \mathbf{v}_{|U|+m+t} \rangle \quad (7)$$

Similarly, the scoring function from Equation 4 is denoted as follows:

$$\hat{y}(u,i,t) = w_t + \sum_{j=1}^{m'} z'_j \langle \mathbf{v}_j, \mathbf{v}_{m'+t} \rangle \quad (8)$$

The gradients of the criterion per quadruples (u,i,t^+,t^-) with respect to the visual features are propagated backward to the CNN to calculate the gradients for all weights of the CNN. From Equation 5 and Equation 6, the gradient for weight w of the feature extractor CNN is computed as:

$$\Delta \mathrm{w} = \frac{\partial \,\text{BPR-OPT}}{\partial \hat{y}(u,i,t^+,t^-)} \times \delta_{(t^+,t^-)} \times \frac{\partial \mathrm{z}}{\partial \mathrm{w}} \tag{9}$$

where $\hat{y}(u,i,t^+,t^-)$ is the difference between the relevant and irrelevant tags and $\delta_{(t^+,t^-)}$ is the derivative of $\hat{y}(u,i,t^+,t^-)$ with respect to the output of the CNN.

More specifically:

$$\delta_{(t^+,t^-)} = \frac{\partial \hat{y}(u,i,t^+,t^-)}{\partial \Psi_{i,t^+}} \times \frac{\partial \Psi_{i,t^+}}{\partial \mathrm{z}} + \frac{\partial \hat{y}(u,i,t^+,t^-)}{\partial \Psi_{i,t^-}} \times \frac{\partial \Psi_{i,t^-}}{\partial \mathrm{z}}$$

In the case of the CNN-DenseFM:

$$\Delta \mathrm{w} = \frac{\partial \,\text{BPR-OPT}}{\partial \hat{y}(u,i,t^+,t^-)} \times \delta'_{(t^+,t^-)} \times \frac{\partial \mathrm{z}'}{\partial \mathrm{w}}$$

$$\delta'_{(t^+,t^-)} = \frac{\partial \hat{y}(u,i,t^+,t^-)}{\partial \Psi_{(u,i),t^+}} \times \frac{\partial \Psi_{(u,i),t^+}}{\partial \mathrm{z}'} + \frac{\partial \hat{y}(u,i,t^+,t^-)}{\partial \Psi_{(u,i),t^-}} \times \frac{\partial \Psi_{(u,i),t^-}}{\partial \mathrm{z}'}$$

4 Evaluation

We performed experiments addressing the impact of visual contents on the recommendation process. Section 4.1 describes the dataset exploited in the evaluation. In Sect. 4.2, we describe the CNN architecture adopted for feature extraction, the evaluation methodology and settings of the proposed models. The results of the evaluation are detailed in Sect. 4.3.

4.1 Dataset

We obtained experiments on a subset of the publicly available multilabel dataset NUS-WIDE (Chua et al, 2009) that contains 269,648 images crawled from Flickr. The subset was obtained by ignoring non-tagged images, removing images that were not available for download in April 2015 from all images of

NUS-WIDE and the 1,000 most popular tags. Moreover, we filtered the dataset to generate a 2-core dataset where each user, image or tag occurs at least in 2 posts (Jäschke et al, 2007). The subset contains 3,009 users, 13,334 images and 983 tags that comprise 29,089 posts and 111,407 triplets (user,image,tag). In addition, we used the Flickr API[2] to download square images having size 75×75.

To evaluate the recommendation models, we adapted leave-one-post-out (Marinho et al, 2012) for users with at least k posts (LOPO-k) to split the dataset. With this kind of division, we randomly took, for each user having at least k posts, one of his post and put it into the test set. In this paper, we use LOPO-2 and LOPO-5 to split the training and test set. With LOPO-2, there are 26,080 posts and 98,610 triplets in the training set and 3,009 posts and 12,797 triplets in the test set. With LOPO-5, the training set has 27,946 posts and 106,826 triplets and the test set has 1,143 posts and 4,581 triplets.

4.2 Experimental Setup

The CNN architecture used in the experiments contains 3 convolutional layers. The first layer filters a $75 \times 75 \times 3$ input image with 10 kernels of size $6 \times 6 \times 3$ with a stride of 3 pixels. The input of the second layer is the output of the pooling layer that takes the maximum value of a square block 2×2 from the first convolutional layer to generate one value of each feature map. It uses 30 kernels of size $6 \times 6 \times 10$ with a stride of 2 pixels to produce feature maps for the next connected pooling layer. The following pooling layer also uses max pooling to produce values of feature maps. The last convolutional layer filters 30 feature maps with 128 kernels of size $2 \times 2 \times 30$ to provide a feature vector for the predictor. In the CNN-DenseFM, a fully-connected layer that extracts the combination features of a post (u, i) produces a dense vector having 128 elements for the recommender.

Due to scalability issues, we evaluated the CNN-DenseFM model only on the LOPO-5 protocol. On the other hand, the CNN-FM model was evaluated in both LOPO-2 and LOPO-5 protocols. For the evaluation, we used three metrics to capture the performance of the models as proposed by Rendle and Schmidt-Thieme (2010).

[2] https://www.flickr.com/services/api/

- Precision at rank K

$$\text{Precision@K} = \underset{(u,i)\in S_{test}}{\text{avg}} \frac{|Top(u,i,K) \cap \{t|(u,i,t) \in S_{test}\}|}{K}$$

- Recall at rank K

$$\text{Recall@K} = \underset{(u,i)\in S_{test}}{\text{avg}} \frac{|Top(u,i,K) \cap \{t|(u,i,t) \in S_{test}\}|}{|\{t|(u,i,t) \in S_{test}\}|}$$

- F1-measure at rank K

$$\text{F1@K} = \frac{2 \cdot \text{Precision@K} \cdot \text{Recall@K}}{\text{Precision@K} + \text{Recall@K}}$$

The number and the size of CNN kernels are fixed like above. We applied a grid search mechanism for looking for the best learning rate α, regularization λ and factor dimension q. The learning rate is selected among the range of $\alpha_{cnn/full} \in \{0.01, 0.001, 0.0001\}$ for all convolutional layers and the fully-connected layer, $\alpha_{fm} \in \{0.01, 0.001, 0.0001\}$ for the FM layer. The values of the regularization hyperparameter for all convolutional layers, the fully-connected layer and the FM layer are selected from $\lambda_{cnn/full} \in \{1e-03, 1e-04, 1e-05\}$ and $\lambda_{fm} \in \{1e-05, 1e-06, 1e-07\}$. The factor dimensions are searched among the range of $q \in \{64, 128, 256\}$.

We compare the proposed models with the following tag recommendation models:

- **PITF** (Rendle and Schmidt-Thieme, 2010)
- **FM** (Rendle, 2010)
- Most popular tags (**MP**) (Jäschke et al, 2007)
- Adapted PageRank (**AP**) (Hotho et al, 2006)
- FolkRank (**FR**) (Jäschke et al, 2007)

We used the Tagrec framework built by Kowald et al (2014) to learn MP, FR and AP while the Tag recommender software[3] was used to learn PITF and FM. We also implemented a convolutional neural network with a multilayer perceptron network placed as the last layer (**CNN-MLP**) to recommend tags based on the content of images. The CNN having the same architecture used in the CNN-FM is deployed in the first step to extract a feature vector and then its values are fed to a multilayer perceptron network as the input to calculate scores of all tags. The parameters of this model are learned based on the WARP loss function (Gong et al, 2013; Weston et al, 2011).

[3] http://www.informatik.uni-konstanz.de/rendle/software/tag-recommender/

Table 1 F1-measure at 5 and 10

Models	LOPO-2 (F1@5)	LOPO-2 (F1@10)	LOPO-5 (F1@5)	LOPO-5 (F1@10)
MP	0.1135	0.1059	0.142	0.1197
CNN-MLP	0.1389	0.1181	0.165	0.1344
AP	0.2366	0.2018	0.2686	0.2193
FR	0.24	0.2021	0.2772	0.2356
CNN-DenseFM	(N/A)	(N/A)	0.2778	0.2241
FM	0.2671	0.224	0.3281	0.2611
PITF	0.2786	0.2242	0.3432	0.2675
CNN-FM	0.2866	0.2293	0.3491	0.2771

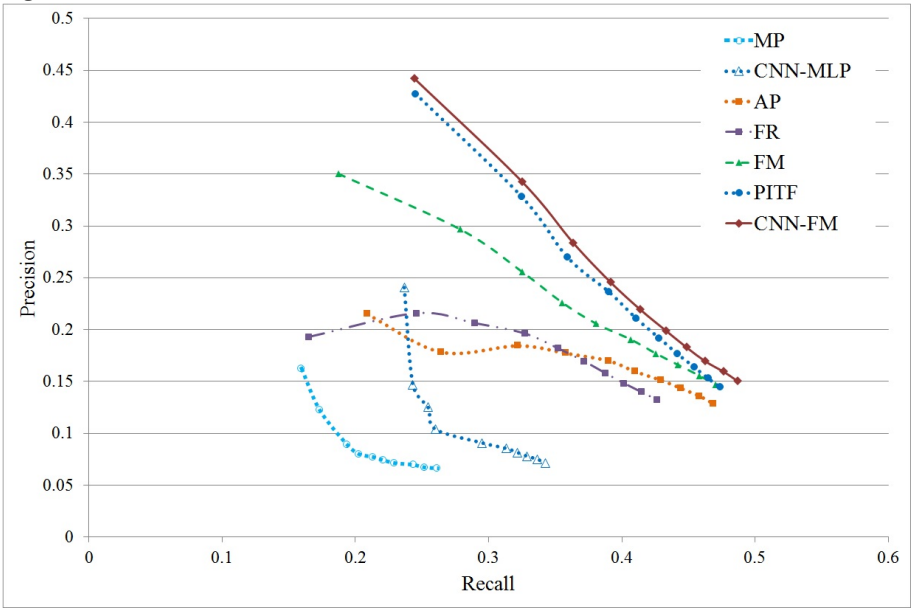
4.3 Results

As shown in Fig. 2 and Fig. 3, the performance of the recommender based only on the visual contents of images (CNN-MLP) nearly equals the performance of the MP. This means that the content-based tag recommendation is not really effective in the case of personalized tag recommendation. In addition, the combination features of the user and the visual content used in the CNN-DenseFM do not increase the performance of the recommender as we can see in Fig. 3. On the contrary, the visual features strongly support the recommendation models achieving higher performances indicated by Figs. 2, 3 and Table 1. For example, the recommender using visual contents improves the performance by 7.2% of precision at 1 in LOPO-2 and 3,8% in LOPO-5 if comparing to FM. Although Rendle (2010) discussed that the performance of the FM and the PITF are comparable in the ECML Discovery Challenge 2009 (task 2), the PITF achieves higher performance than the FM in this narrow folksonomy. However, feeding a FM with CNN-extracted visual features boosts its performance so that it is comparable to a PITF. The result also opens the direction of using visual contents with the PITF model.

5 Conclusion

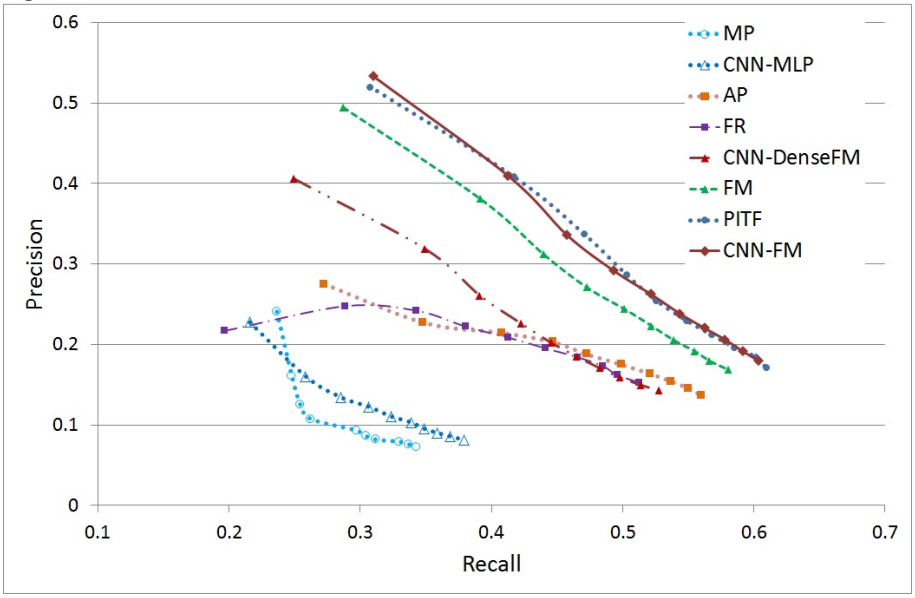
In this paper, we proposed a method to use visual features of images to enhance the performance of a state-of-the-art tag recommendation model. The learnable features of images that are extracted by a CNN are used in a FM to compute the scores of all tags. The experiments show that our proposed method has

Fig. 2 Precision and Recall for LOPO-2 from rank 1 to 10



advantages over the FM that uses only historical tagging information. Through the experiments, the PITF works better than the FM in narrow folksonomy scenarios like Flickr without using contents of images to predict candidate tags. In the future, we plan to investigate another personalized content-aware image tag recommendation that is a combination of a CNN and a PITF and evaluate the effectiveness of its performance on the case of personalized tag recommendation.

Fig. 3 Precision and Recall for LOPO-5 from rank 1 to 10



References

Ames M, Naaman M (2007) Why we tag: Motivations for annotation in mobile and online media. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM, ACM, New York, CHI '07, pp 971–980, DOI 10.1145/1240624.1240772

Chua TS, Tang J, Hong R, Li H, Luo Z, Zheng Y (2009) NUS-WIDE: A real-world web image database from national university of singapore. In: Proceedings of the ACM International Conference on Image and Video Retrieval, ACM, New York, CIVR '09, pp 48:1–48:9, DOI 10.1145/1646396.1646452

Garg N, Weber I (2008) Personalized, interactive tag recommendation for flickr. In: Proceedings of the 2008 ACM Conference on Recommender Systems, ACM, New York, RecSys '08, pp 67–74, DOI 10.1145/1454008.1454020

Gong Y, Jia Y, Leung T, Toshev A, Ioffe S (2013) Deep convolutional ranking for multilabel image annotation. CoRR URL http://arxiv.org/abs/1312.4894

Hotho A, Jäschke R, Schmitz C, Stumme G, Althoff KD (2006) Folkrank: A ranking algorithm for folksonomies. In: Schaaf M, Althoff KD (eds) LWA

2006: Lernen - Wissensentdeckung - Adaptivität, Institut für Informatik, Universität Hildesheim, Hildesheim, Hildesheimer Informatik-Berichte, vol 1, pp 111–114, URL http://www.kde.cs.uni-kassel.de/stumme/papers/2006/hotho2006folkrank.pdf

Jäschke R, Marinho L, Hotho A, Schmidt-Thieme L, Stumme G (2007) Tag Recommendations in Folksonomies, Springer, Berlin, pp 506–514. DOI 10.1007/978-3-540-74976-9_52

Kowald D, Lacic E, Trattner C (2014) Tagrec: Towards a standardized tag recommender benchmarking framework. In: Proceedings of the 25th ACM Conference on Hypertext and Social Media, ACM, ACM, HT '14, pp 305–307, DOI 10.1145/2631775.2631781

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges C, Weinberger K (eds) Proceedings of the 25th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, pp 1097–1105

LeCun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324, DOI 10.1109/5.726791

Li J, Wang JZ (2008) Real-time computerized annotation of pictures. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(6):985–1002, DOI 10.1109/TPAMI.2007.70847

Li X, Snoek CG, Worring M (2008) Learning tag relevance by neighbor voting for social image retrieval. In: Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, ACM, New York, pp 180–187, DOI 10.1145/1460096.1460126

Lin M, Chen Q, Yan S (2013) Network in network. CoRR abs/1312.4400, URL http://arxiv.org/abs/1312.4400

Marinho LB, Hotho A, Jäschke R, Nanopoulos A, Rendle S, Schmidt-Thieme L, Stumme G, Symeonidis P (2012) Recommender Systems for Social Tagging Systems. SpringerBriefs in Electrical and Computer Engineering, Springer, New York, DOI 10.1007/978-1-4614-1894-8

Qian X, Liu X, Zheng C, Du Y, Hou X (2013) Tagging photos using users' vocabularies. Neurocomputing 111:144–153, DOI 10.1016/j.neucom.2012.12.021

Rendle S (2010) Factorization machines. In: 2010 IEEE International Conference on Data Mining, IEEE, pp 995–1000, DOI 10.1109/ICDM.2010.127

Rendle S, Schmidt-Thieme L (2010) Pairwise interaction tensor factorization for personalized tag recommendation. In: Proceedings of the third ACM

international conference on Web search and data mining, ACM, ACM, New York, WSDM '10, pp 81–90, DOI 10.1145/1718487.1718498

Rendle S, Freudenthaler C, Gantner Z, Schmidt-Thieme L (2009) Bpr: Bayesian personalized ranking from implicit feedback. In: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Arlington, UAI '09, pp 452–461

Sigurbjörnsson B, Van Zwol R (2008) Flickr tag recommendation based on collective knowledge. In: Proceedings of the 17th international conference on World Wide Web, ACM, New York, WWW '08, pp 327–336, DOI 10.1145/1367497.1367542

Wan L, Zeiler M, Zhang S, Cun YL, Fergus R (2013) Regularization of neural networks using dropconnect. In: Dasgupta S, Mcallester D (eds) Proceedings of the 30th International Conference on Machine Learning (ICML-13), JMLR Workshop and Conference Proceedings, pp 1058–1066

Wei Y, Xia W, Huang J, Ni B, Dong J, Zhao Y, Yan S (2014) CNN: single-label to multi-label. CoRR abs/1406.5726, URL http://arxiv.org/abs/1406.5726

Weston J, Bengio S, Usunier N (2011) WSABIE: Scaling up to large vocabulary image annotation. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Three, AAAI Press, IJCAI'11, vol 11, pp 2764–2770, DOI 10.5591/978-1-57735-516-8/IJCAI11-460