Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Weakly-Supervised Deep Nonnegative Low-Rank Model for Social Image Tag Refinement and Assignment

Zechao Li, Jinhui Tang*
School of Computer Science and Engineering, Nanjing University of Science and Technology
No. 200 Xiaolingwei Road, Nanjing, China 210094
{zechao.li, jinhuitang}@njust.edu.cn

Abstract

It has been well known that the user-provided tags of social images are imperfect, i.e., there exist noisy, irrelevant or incomplete tags. It heavily degrades the performance of many multimedia tasks. To alleviate this problem, we propose a Weakly-supervised Deep Nonnegative Low-rank model (WDNL) to improve the quality of tags by integrating the low-rank model with deep feature learning. A nonnegative low-rank model is introduced to uncover the intrinsic relationships between images and tags by simultaneously removing noisy or irrelevant tags and complementing missing tags. The deep architecture is leveraged to seamlessly connect the visual content and the semantic tag. That is, the proposed model can well handle the scalability by assigning tags to new images. Extensive experiments conducted on two real-world datasets demonstrate the effectiveness of the proposed method compared with some state-of-the-art methods.

Introduction

Image Tagging (or automatic image annotation) is an essential component of image search systems by estimating the semantic relationships between tags and images. Traditional methods (Barnard et al. 2003; Wong and Leung 2008; Makadia, Pavlovic, and Kumar 2010; Li et al. 2010; Yang, Jing, and Ng 2015) are always based on huge human-labeled training images and difficult to scale. Recent years have witnessed the proliferation of digital images on the social media websites, which poses a great challenge for the traditional image tagging. This challenge can be somewhat alleviated by the user-provided tags. Unfortunately, these tags are often incomplete or inaccurate in describing the visual content of images. That is, the user-provided tags are weakly supervised. What is more, a large fraction (over 50% in Flickr) of images have no tags at all (Chen, Zheng, and Weinberger 2013). Therefore, it is necessary to improve the quality of tags by complementing relevant tags and removing irrelevant tags, and assign tags to new images.

Some works (Li, Snoek, and Worring 2009; Zhu, Yan, and Ma 2010; Qi et al. 2012; Wu, Jin, and Jain 2013; Feng et al. 2014; Johnson, Ballan, and Fei-Fei 2015; Tang et al. 2016a) have been proposed to refine tags of social images. The tag relevance is estimated based on the neighbor voting model (Li, Snoek, and Worring 2009). The low rank model is exploited in (Zhu, Yan, and Ma 2010) by considering the semantic consistency and the visual consistency while a latent space is uncovered based on low rank approximation to connect the visual features of images and tags in (Qi et al. 2012). In (Wu, Jin, and Jain 2013), the relevance between images and tags is estimated by considering the observed image-tag relation and the visual information. Tag completion is implemented by two types of linear sparse reconstructions in (Feng et al. 2014). Social image metadata is used to find neighborhoods and a deep neural network is utilized to blend visual information from the image and its neighbors in (Johnson, Ballan, and Fei-Fei 2015). This work focuses on the low-rank model for social image tag refinement and assignment. The most works are ones in (Zhu, Yan, and Ma 2010; Feng et al. 2014). However, they are not able to link the visual information and the learned low-rank space. That is, the scalability of these methods is limited.

Towards this end, this work tries to simultaneously addresses the problems of image tag refinement and tag assignment assigning tags to new images, and proposes a novel Weakly-supervised Deep Nonnegative Low-rank Model (WDNL) as shown in Figure 1. The proposed model can seamlessly bridge the visual information and the semantic tags. To learn better relationships between images and tags, an ideal image-tag relation matrix is introduced by requiring that it should be low rank and each element is nonnegative. It is natural and reasonable because tags as one kind of text information are subject to the low-rank property (Zhao and Grosky 2002) and the intrinsic relationships between images and tags are nonnegative according to its definition. Thus, a nonnegative low-rank model is proposed to formulate the above motivation. To link the visual information and the ideal tags, a deep architecture is leveraged by learning discriminative features from images, and the ideal tags can be easily predicted with the learned features using a linear transformation matrix. The above components are formulated into a joint deep learning framework. Extensive experiments are conducted on two widely used datasets, and the results compared with some state-of-the-arts verify the effectiveness of the proposed method. Generally, the main contributions of this work are summarized as follows.

*Corresponding author.

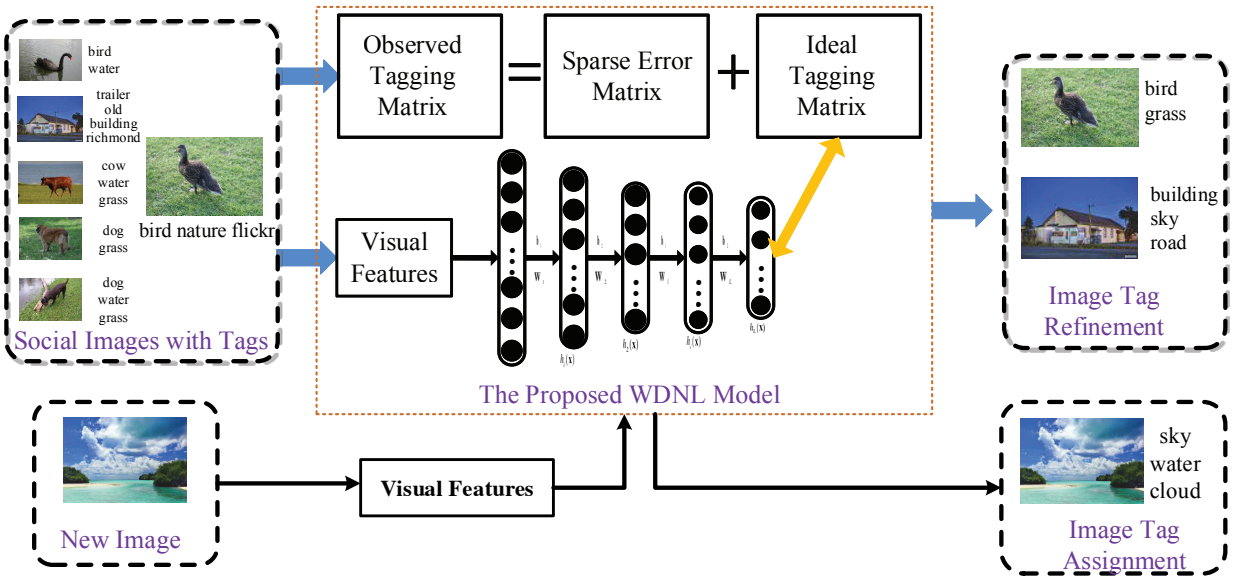

Figure 1: Illustration of the proposed method for social image tag refinement and assignment.

- We propose a weakly-supervised deep nonnegative low-rank framework for the tasks of image tag refitment and tag assignment simultaneously. The ideal image-tag relation matrix and the deep architecture are jointly leaned.
- The learned image-tag relation matrix can provide better information to learn the parameters of the deep architecture.
- The visual contents of images and the high-level tags are seamlessly linked, which can give a direct solution to the prediction of new images.

Previous Work

Many efforts have been devoted to improving the quality of tags (Li and Tang 2015b). Web images are used to refine annotation of an image. Wang et al. (Wang et al. 2010) proposed to exploit the surrounding text of web images to annotate images. In (Li, Snoek, and Worring 2009), tag relevance to the visual content of images is computed based on the neighbor voting strategy. The visual neighbor voting and belief theory in (Znaidia, Borgne, and Hudelot 2013) are used to refine tags of images. Label propagation over noisily-tagged images refines tags based on the kNN-sparse graph on the labeled and unlabeled images in (Tang et al. 2011). In (Wu, Jin, and Jain 2013), the relevance between images and tags is learned by considering the consistency between the observed tags and the visual similarity. Some researchers focus on low-rank models to refine tags of social images. Zhu et al. (Zhu, Yan, and Ma 2010) proposed to refine tags by decomposing the image-tag matrix into a low rank matrix and a sparse matrix, and considering the content consistency and tag correlation. In (Qi et al. 2012), a latent space is uncovered based on low rank approximation by linking the visual features of images and tags. In (Feng et al. 2014), the missing tags are complemented and the noisy tags are de-emphasized by combing the low rank matrix recovery and maximum likelihood estimation. Different from the above models, a new nonnegative low-rank model with a deep architecture is proposed, which can learn better relationships between images and tags.

Deep models have been proved to be potentially effective in addressing complex tasks (Bengio 2009). Recently, some deep models are proposed for image tagging. Gong et al. (Gong et al. 2013) proposed to use convolutional architectures for multilabel tagging. Images are annotated using the Canonical Correlation Analysis (CCA) model using CNN features and textual features in (Murthy, Maji, and Manmatha 2015). In (Li and Tang 2016), a deep matrix factorization method is proposed to improve the quality of tags.

Different from the above methods, this work proposes a weakly-supervised deep nonnegative low-rank model to refine tags of social images and assign tags to new images. The proposed model can seamlessly connect the visual information and the high-level semantic tags.

The Proposed WDNL Model

In this section, we first introduce the motivation of this work and then elaborate the proposed WDNL model.

Motivation

The essential problem of image tag refinement and image tag assignment is how to uncover the intrinsic relevance of tags to the visual content of images with the help of available resources. In this work we focus on how to address this issue based on the low-rank framework.

For social images, users often provide some tags to tag them. It has been demonstrated in (Zhao and Grosky 2002)

that the semantic space spanned by tags can be approximated by a subset of salient tags from the original tag space. And users usually tag images with the semantically correlated tags synchronously. Consequently, tags of social images are subject to the low-rank property. That is, the intrinsic image-tag relation matrix is a low-rank one. On the other hand, the user-provided tags are reasonably accurate to certain level and the number of tagged tags is essentially sparse with respect to the number of total tags. As a consequence, the error of the observed image-tag relation matrix is sparse. Our goal is to uncover the intrinsic low-rank matrix by decomposing the observed image-tag relation matrix into its sparse and low-rank components. On the other hand, although there exist noisy or irrelevant tags, the visual contents of images and tags are correlated. It is necessary to develop a scheme to link them, which can alleviate the well-known semantic gap. To seamlessly bridge the visual contents and the high-level tags, a deep architecture is introduced. What is more important, the introduced deep architecture makes the proposed method have good scalability. That is, it can assign tags to any coming image.

The Proposed Formulation

Without loss of generality, in this work lowercase italic letters (i.e., i, j, n, etc.) and uppercase italic letters (i.e., A, B, M, etc.) denote scalars while bold uppercase characters (i.e, $\mathbf{W}$, $\mathbf{X}$, etc.) and bold lowercase characters (i.e, $\mathbf{a}$, $\mathbf{x}$, etc.) are utilized to denote matrices and vectors, respectively. For any matrix $\mathbf{A}$, $\mathbf{a}^i$ means the i-th column vector of $\mathbf{A}$, $\mathbf{a}_i$ means the i-th row vector of $\mathbf{A}$, A_{ij} denotes the (i,j)-element of $\mathbf{A}$ and Tr[$\mathbf{A}$] is the trace of $\mathbf{A}$ if $\mathbf{A}$ is square. $\mathbf{A}^T$ denotes the transposed matrix of $\mathbf{A}$. The Frobenius norm of a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ is defined as $\|\mathbf{A}\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n A_{ij}^2 = \text{Tr}[\mathbf{A}^T\mathbf{A}]$. The nuclear norm of $\mathbf{A}$ is denoted as $\|\mathbf{A}\|_*$. The ℓ_1 norm of $\mathbf{A}$ is defined as $\|\mathbf{A}\|_1 = \sum_{i=1}^m \sum_{j=1}^n |A_{ij}|$.

Consider a social image set consisting of n images $\{\mathbf{x}^i\}_{i=1}^n$ assigned with m user-provided tags $\mathcal{C} = \{t_1, t_2, \cdots, t_m\}$. For each image $\mathbf{x}^i$, the observed relationships between this image and tags can be represented as a m-dimensional binary-valued vector $\{\mathbf{f}^i\}$. The visual feature matrix is denoted as $\mathbf{X} = [\mathbf{x}^1, \cdots, \mathbf{x}^n]$, in which $\mathbf{x}^i \in \mathbb{R}^d$ is the feature vector of the i-th image while $\mathbf{F} = [\mathbf{f}^1, \cdots, \mathbf{f}^n] \in \mathbb{R}^{m\times n}$ is the observed tagging matrix, in which $F_{ji} = 1$ indicates that $\mathbf{x}^i$ is associated with the j-th tag, and $F_{ji} = 0$ otherwise. The ideal tagging matrix is denoted as $\mathbf{Y} = [\mathbf{y}^1, \cdots, \mathbf{y}^n] \in \mathbb{R}^{m\times n}$. Under the low-rank framework, the essential purpose of tag refinement is to uncover $\mathbf{Y}$ and the error matrix $\mathbf{E}$.

$$\mathbf{F} = \mathbf{Y} + \mathbf{E} \tag{1}$$

As aforementioned as well as introduced later, in the proposed scheme, there are three components for the objective to optimize. Specially, we utilize $rank(\mathbf{Y})$ to characterize the rank of the matrix $\mathbf{Y}$, $S(\mathbf{E})$ to measure the sparsity of the tagging error matrix $\mathbf{E}$, $loss(\mathbf{Y}, f(\Theta; \mathbf{X}))$ to measure the loss function of the tag prediction, and $\Omega(\Theta)$ to represent the regularization term. Here, $f(\Theta; \mathbf{X})$ is the prediction function based on the deep architecture and Θ is the set of parameters in the deep architecture. The proposed method is formulated as follows.

$$\begin{aligned}\min_{\mathbf{Y},\Omega}\ & rank(\mathbf{Y}) + \lambda_1 S(\mathbf{F}-\mathbf{Y}) \\ & +\lambda_2 loss(\mathbf{Y}, f(\Theta;\mathbf{X})) + \lambda_3\Omega(\Theta)\end{aligned} \tag{2}$$

Here λ_1, λ_2 and λ_3 are three nonnegative parameters to balance these terms. As discussed above, $\mathbf{Y}$ is subject to the low-rank property and the error matrix $\mathbf{E}$ is sparse. Thus we have

$$rank(\mathbf{Y}) = \|\mathbf{Y}\|_* \tag{3}$$

$$S(\mathbf{E}) = \|\mathbf{E}\|_1 \tag{4}$$

According to the definition of the tagging matrix $\mathbf{Y}$, each element Y_{ji} denotes the relevance of the j-th tag to the i-th image, which is nonnegative in nature. Unfortunately, in the above formulation, we do not impose any constraint on the signs of elements of $\mathbf{F}$. It leads to that the learned $\mathbf{F}$ has mixed signs, which violates its definition. To well address this problem, it is natural and reasonable to impose nonnegative constraints on $\mathbf{Y}$. Thus, we have the following objective function.

$$\begin{aligned}&\min_{\mathbf{Y},\Omega} \|\mathbf{Y}\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}\|_1 + \lambda_2 loss(\mathbf{Y}, f(\Theta;\mathbf{X})) + \lambda_3\Omega(\Theta) \\ &\text{s.t.}\quad \mathbf{Y} \geq 0\end{aligned} \tag{5}$$

Here $\mathbf{Y} \geq 0$ denotes that each element of $\mathbf{Y}$ is nonnegative.

To well bridge the visual features and the high-level tags, the deep architecture is introduced. The proposed deep framework contains L layers of nonlinear transformations, and there are $r_l (l = 1, \cdots, L)$ units in the l-th layer. The output of the most top layer is the expected representation $g(\mathbf{x}_i)$. The input of the $(l+1)$-th layer is the output of the l-th layer. Thus, we have

$$g(\mathbf{x}_i) = h_L(\mathbf{x}_i), \tag{6}$$

$$h_l(\mathbf{x}_i) = s(\mathbf{W}_l h_{l-1}(\mathbf{x}_i) + \mathbf{b}_l), l = 1, \cdots, L, \tag{7}$$

where $s(\cdot)$ is a nonlinear activation function and $h_l(\cdot)$ is the output of the l-th layer. $\mathbf{W}_l$ and $\mathbf{b}_l$ are the projection matrix and the bias vector to be learned in the l-th layer, respectively. The input of the first layer is the original visual feature: $h_0(\mathbf{x}_i) = \mathbf{x}_i$. And the output of the first layer is

$$h_1(\mathbf{x}_i) = s(\mathbf{W}_1\mathbf{x}_i + \mathbf{b}_1). \tag{8}$$

The learned representation $g(\mathbf{x}_i)$ is expected to have the ability to bridge visual contents and tags. That is, it should be discriminative to well predict the proper tags. For this purpose and for simplicity, a linear prediction function is introduced to predict tags.

$$f(\Theta;\mathbf{X}) = \mathbf{W}g(\mathbf{X}) \tag{9}$$

Hence $\Theta = \{\mathbf{W}_1, \cdots, \mathbf{W}_L, \mathbf{b}_1, \cdots, \mathbf{b}_L, \mathbf{W}\}$. To avoid the problem of overfitting, it is necessary and reasonable to impose regularization terms on the parameters of the proposed method. For simplicity, the regularization function $\Omega(\Theta)$ is defined as follows.

$$\Omega(\Theta) = \frac{1}{2}(\|\mathbf{W}\|_F^2 + \sum_{l=1}^{L}(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2)) \tag{10}$$

By incorporating the above-described properties, the proposed method is formulated as the following optimization problem.

$$\min_{\mathbf{Y}\geq 0,\mathbf{W},\mathbf{W}_l,\mathbf{b}_l} \|\mathbf{Y}\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}\|_1 + \lambda_2 loss(\mathbf{Y},\mathbf{W}g(\mathbf{X})) + \frac{\lambda_3}{2}(\|\mathbf{W}\|_F^2 + \sum_{l=1}^{L}(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2)) \quad (11)$$

Optimization

In order to solve the proposed problem (11), the loss function $loss(\cdot,\cdot)$ should be defined in advance. This work utilizes the function $loss(x,y) = \frac{1}{2}(x-y)^2$ to measure the prediction error. Consequently, we obtain the following problem.

$$\min_{\mathbf{Y}\geq 0,\mathbf{W},\mathbf{W}_l,\mathbf{b}_l} \|\mathbf{Y}\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}\|_1 + \frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 + \frac{\lambda_3}{2}(\|\mathbf{W}\|_F^2 + \sum_{l=1}^{L}(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2)) \quad (12)$$

An iterative algorithm is developed to optimize the above problem similar to (Li and Tang 2015a).

With $\mathbf{W}$,$\mathbf{W}_l$ and $\mathbf{b}_l$ ($1 \leq l \leq L$) fixed, we first introduce two auxiliary variables $\mathbf{Y}_1$ and $\mathbf{Y}_2$ to make the objective function separable. The following problem is obtained.

$$\min_{\mathbf{Y},\mathbf{Y}_1,\mathbf{Y}_2} \|\mathbf{Y}_1\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}_2\|_1 + \frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 \quad \text{s.t.} \quad \mathbf{Y}_1 = \mathbf{Y},\ \mathbf{Y}_2 = \mathbf{Y},\ \mathbf{Y}_2 \geq 0 \quad (13)$$

Then, the inexact augmented Lagrangian method (IALM) (Lin, Chen, and Ma 2009) is used to solve the above low-rank problem. The augmented Lagrangian function of problem (13) is as follows.

$$\begin{aligned}\mathcal{L}(\mathbf{Y},\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Z}_1,\mathbf{Z}_2,\eta) &= \|\mathbf{Y}_1\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}_2\|_1 \\ &+\frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 + <\mathbf{Z}_1,\mathbf{Y}-\mathbf{Y}_1> \\ &+<\mathbf{Z}_2,\mathbf{Y}-\mathbf{Y}_2> + \frac{\eta}{2}(\|\mathbf{Y}-\mathbf{Y}_1\|_F^2 + \|\mathbf{Y}-\mathbf{Y}_2\|_F^2) \\ &= \|\mathbf{Y}_1\|_* + \lambda_1\|\mathbf{F}-\mathbf{Y}_2\|_1 + \frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 \\ &+\frac{\eta}{2}(\|\mathbf{Y}-\mathbf{Y}_1+\frac{1}{\eta}\mathbf{Z}_1\|_F^2 + \|\mathbf{Y}-\mathbf{Y}_2+\frac{1}{\eta}\mathbf{Z}_2\|_F^2) \\ &-\frac{1}{2\eta}(\|\mathbf{Z}_1\|_F^2 + \|\mathbf{Z}_2\|_F^2) \end{aligned} \quad (14)$$

According to the IALM method, the objective function converges with a sequence of closed form updating steps. The variable $\mathbf{Y}$, $\mathbf{Y}_1$ or $\mathbf{Y}_2$ is updated with other variables fixed. The detailed updating rules are presented as follows.

$$\begin{aligned}\mathbf{Y} &= \arg\min_{\mathbf{Y}} \frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 + <\mathbf{Z}_1,\mathbf{Y}-\mathbf{Y}_1> \\ &+<\mathbf{Z}_2,\mathbf{Y}-\mathbf{Y}_2> + \frac{\eta}{2}(\|\mathbf{Y}-\mathbf{Y}_1\|_F^2 + \|\mathbf{Y}-\mathbf{Y}_2\|_F^2) \\ &= \frac{1}{\lambda_2+2\eta}(\lambda_2\mathbf{W}g(\mathbf{X}) - \mathbf{Z}_1 - \mathbf{Z}_2 + \eta\mathbf{Y}_1 + \eta\mathbf{Y}_2) \end{aligned} \quad (15)$$

$$\begin{aligned}\mathbf{Y}_1 &= \arg\min_{\mathbf{Y}_1} \|\mathbf{Y}_1\|_* + \frac{\eta}{2}\|\mathbf{Y}-\mathbf{Y}_1+\frac{1}{\eta}\mathbf{Z}_1\|_F^2 \\ &= \Gamma_{\frac{1}{\eta}}(\mathbf{Y}+\frac{1}{\eta}\mathbf{Z}_1) \end{aligned} \quad (16)$$

$$\begin{aligned}\mathbf{Y}_2 &= \arg\min_{\mathbf{Y}_2} \frac{\lambda_2}{\eta}\|\mathbf{Y}_2-\mathbf{Y}\|_1 + \frac{1}{2}\|\mathbf{Y}-\mathbf{Y}_2+\frac{1}{\eta}\mathbf{Z}_2\|_F^2 \\ &= soft(\frac{1}{\eta}\mathbf{Z}_2, \frac{\lambda_2}{\eta}) + \mathbf{Y} \end{aligned} \quad (17)$$

Here Γ is singular value soft-thresholding operator and soft is soft-thresholding operator. In detail, the form of analytic solution for soft is as follows.

$$soft(A_{ij},\gamma) = \text{sign}(A_{ij})\max(|A_{ij}|-\gamma, 0) \quad (18)$$

Then, we have the definition of Γ.

$$\Gamma_\gamma(\mathbf{A}) = \mathbf{U}soft(\Lambda,\gamma)\mathbf{V}^T \quad (19)$$

in which $\mathbf{A} = \mathbf{U}\Lambda\mathbf{V}^T$ is the SVD of $\mathbf{A}$.

In the following, for ease of presentation, we use $\mathcal{O}$ to denote the objective function with respect to $\mathbf{W}$,$\mathbf{W}_l$ and $\mathbf{b}_l$.

$$\mathcal{O} = \frac{\lambda_2}{2}\|\mathbf{Y}-\mathbf{W}g(\mathbf{X})\|_F^2 + \frac{\lambda_3}{2}(\|\mathbf{W}\|_F^2 + \sum_{l=1}^{L}(\|\mathbf{W}_l\|_F^2 + \|\mathbf{b}_l\|_2^2)) \quad (20)$$

With the learned $\mathbf{Y}$, we have the following updating rules for $\mathbf{W}$,$\mathbf{W}_l$ and $\mathbf{b}_l$ by simple inference.

$$\mathbf{W} = \mathbf{Y}g(\mathbf{X})^T(g(\mathbf{X})g(\mathbf{X})^T + \frac{\lambda_3}{\lambda_2}\mathbf{I})^{-1} \quad (21)$$

$$\mathbf{W}_l = \mathbf{W}_l - \mu\frac{\partial\mathcal{O}}{\partial\mathbf{W}_l} \quad (22)$$

$$\mathbf{b}_l = \mathbf{b}_l - \mu\frac{\partial\mathcal{O}}{\partial\mathbf{b}_l} \quad (23)$$

$$\frac{\partial\mathcal{O}}{\partial\mathbf{W}_l} = \lambda_2\sum_{i=1}^{n}\nabla_l(\mathbf{x}_i)h_{l-1}^T(\mathbf{x}_i) + \lambda_3\mathbf{W}_l \quad (24)$$

$$\frac{\partial\mathcal{O}}{\partial\mathbf{b}_l} = \lambda_2\sum_{i=1}^{n}\nabla_l(\mathbf{x}_i) + \lambda_3\mathbf{b}_l \quad (25)$$

$$\nabla_l(\mathbf{x}_i) = (\mathbf{W}_{l+1}^T\nabla_{l+1}(\mathbf{x}_i)) \circ s'(\mathbf{W}_l h_{l-1}(\mathbf{x}_i) + \mathbf{b}_l) \quad (26)$$

$$\nabla_L(\mathbf{x}_i) = (\mathbf{W}^T(\mathbf{W}g(\mathbf{x}_i) - \mathbf{Y})) \circ s'(\mathbf{W}_L h_{L-1}(\mathbf{x}_i) + \mathbf{b}_L) \quad (27)$$

The parameter μ is the learning rate for the gradient descent algorithm. We summarize the proposed optimization algorithm in Algorithm 1.

Algorithm 1 The Proposed WDNL Method

Input:

Visual features $\mathbf{X}$ and the observed tagging matrix $\mathbf{F}$.

$\rho = 1.1$; $\eta = 0.1$;
Initialize $\mathbf{W}_l$ and $\mathbf{b}_l$, $l = 1, \cdots, L$;
repeat
  Update $\mathbf{Y}$ according to Eq. 15;
  Update $\mathbf{Y}_1$ according to Eq. 16;
  Update $\mathbf{Y}_2$ according to Eq. 17;
  $\mathbf{Z}_1 \leftarrow \mathbf{Z}_1 + \eta(\mathbf{Y} - \mathbf{Y}_1)$;
  $\mathbf{Z}_2 \leftarrow \mathbf{Z}_2 + \eta(\mathbf{Y} - \mathbf{Y}_2)$;
  $\eta = \rho\eta$;
  Set $h_0(\mathbf{x}_i) = \mathbf{x}_i$;
  for $l = 1, \cdots, L$
    Calculate $h_l(\mathbf{x}_i)$ by forward propagation;
  end
  Update $\mathbf{W}$ according to Eq. 21;
  for $l = L, \cdots, 1$
    Update $\mathbf{W}_l$ and $\mathbf{b}_l$ by back propagation according to Eq. 22 and Eq. 23;
  end
until Convergence criterion satisfied

Output:

The model parameters $\mathbf{Y}$, $\mathbf{W}$, $\mathbf{W}_l$ and $\mathbf{b}_l$, $l = 1, \cdots, L$.

Experiments

In this section, we present extensive experiments to validate the effectiveness of the proposed method. We first present the experimental setup. Then the experimental results and analysis are presented.

Experimental Setup

Social images can cover almost all the concepts people always use, which makes researchers to build social image datasets for experimental purpose. In this work, we conduct our experiments on two publicly-available social datasets: MIRFlickr (Huiskes and Lew 2008) and NUS-WIDE (Tang et al. 2016b). MIRFlickr contains 25,000 images collected from Flickr. A vocabulary of 457 tags is used, which contains the ground-truth annotation of 18 concepts. NUS-WIDE has about $270,000$ images from Flickr associated with 3,137 unique tags. 81 concepts are labeled by human for all the images as ground truth for evaluation. In our experiments, data are partitioned into two groups: the learning data and the testing data. The learning data is used for model estimation and evaluate the performance of noisy tagged data while the testing data is utilized to test the performance of new data. We randomly select 5000 and $10,000$ samples for the MIRFlickr and NUS-WIDE datasets respectively as learning data and the remaining samples are used as testing data. During the partition process, each label is guaranteed to be associated with at least one images. To alleviate the instability introduced by the randomly selected training data, we independently repeat experiments 5 times to generate different learning and testing data, and report the average results. We adopt the grid-search strategy to tune the hyper-parameters following previous works. For the deep architecture, we set the number of layers to 3, i.e., $L = 3$.

Evaluation Metric: The results on the noisy tagged learning data and the testing data are both reported. Area Under Cures (AUC) is utilized as evaluation metric in our experiments. Both the microaveraging and macroaveraging measures are adopted to measure both the global performance across multiple concepts and the average performance of all the concepts (Li et al. 2015). We also compute Mean Average Precision (MAP) over concepts by averaging average precision over all concepts to measure the performance for image retrieval.

Visual Features: For the MIRFlickr dataset, two types of global descriptors (Gist features and color histograms with 16 bins in each color channel for LAB and HSV representations) and one type of local feature (SIFT feature) are utilized to describe the visual content. One image is described by one 3,659-D vector. For the NUS-WIDE dataset, 1,134-D visual features provided by the dataset are used to describe the image visual content, including 64-D color histogram (LAB), 144-D color auto-correlation (HSV), 73-D edge direction histogram, 128-D wavelet texture and 225-D block-wise color moments (LAB) and 500-D bag of words based on SIFT descriptions. It is worth noting that binary vectors are adopted to represent the observed relevance of tags to an image.

Compared Algorithms: Extensive comparisons are presented with several related work, including state-of-the-art methods. The compared methods are listed as follows: LR (Candès et al. 2011), LRES (Zhu, Yan, and Ma 2010), CCA (Sun, Ji, and Ye 2011), C2MR (Qi et al. 2012), TCMR (Feng et al. 2014) and CCA-CNN (Murthy, Maji, and Manmatha 2015). It is worth noting that for CCA-CNN, we utilize the binary tag-image vector to represent one tag instead of the word embedding vector in (Murthy, Maji, and Manmatha 2015), and a 4096-D feature vector is used to represent the visual content of images extracted by AlexNet (Krizhevsky, Sutskever, and Hinton 2012).

Results for Social Image Tag Refinement

We first conduct experiments to verify the performance of the proposed method for the task of social image tag refinment. The compared experimental results in terms of the mean MicroAUC and mean MacroAUC are presented in Table 1.

From the compared results, we can observe that the proposed method obtains the best performance. It demonstrates the advantages of the proposed method. The compared results between CCA and CCA-CNN (Murthy, Maji, and Manmatha 2015) show that the representations learned by the deep learning models can lead to better performance. Comparing the proposed WDNL model with other methods, we can see that it is benefical to link the low-level visual information and the high-level semantic information through a deep architecture. It is reasonable to believe that the results can be improved if we extract features directly from the raw pixels and use the learned semantic information to tune the deep architecture, which will be explored in our future work.

Table 1: Experimental results (mean microauc ± standard deviation and mean macroauc ± standard deviation) on the MIRFlickr and NUS-WIDE datasets for image tag refinement. The best results are highlighted in bold.

Method	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
LR	0.614 ± 0.005	0.591 ± 0.007	0.695 ± 0.005	0.684 ± 0.006
LRES	0.637 ± 0.006	0.621 ± 0.006	0.755 ± 0.005	0.738 ± 0.004
C2MR	0.640 ± 0.007	0.620 ± 0.005	0.765 ± 0.002	0.650 ± 0.002
CCA	0.583 ± 0.004	0.572 ± 0.004	0.630 ± 0.003	0.728 ± 0.005
TCMR	0.631 ± 0.001	0.623 ± 0.003	0.770 ± 0.004	0.719 ± 0.008
CCA-CNN	0.644 ± 0.006	0.631 ± 0.004	0.675 ± 0.005	0.743 ± 0.007
DMF	0.639 ± 0.003	0.628 ± 0.002	0.751 ± 0.005	0.739 ± 0.004
WDNL	$\mathbf{0.685 \pm 0.003}$	$\mathbf{0.671 \pm 0.003}$	$\mathbf{0.789 \pm 0.006}$	$\mathbf{0.762 \pm 0.006}$

Table 2: Experimental results (mean microauc ± standard deviation and mean macroauc ± standard deviation) on the MIRFlickr and NUS-WIDE datasets for image tag assignment. The best results are highlighted in bold.

Method	MIRFlickr		NUS-WIDE	
	MicroAUC	MacroAUC	MicroAUC	MacroAUC
LR	0.598 ± 0.004	0.577 ± 0.005	0.678 ± 0.004	0.576 ± 0.007
LRES	0.627 ± 0.002	0.618 ± 0.003	0.713 ± 0.005	0.639 ± 0.003
CCA	0.575 ± 0.006	0.560 ± 0.008	0.598 ± 0.005	0.624 ± 0.006
C2MR	0.634 ± 0.004	0.624 ± 0.006	0.725 ± 0.007	0.638 ± 0.005
TCMR	0.627 ± 0.003	0.589 ± 0.005	0.735 ± 0.004	0.643 ± 0.006
CCA-CNN	0.642 ± 0.005	0.627 ± 0.002	0.617 ± 0.004	0.641 ± 0.003
DMF	0.635 ± 0.002	0.623 ± 0.003	0.737 ± 0.007	0.632 ± 0.004
WDNL	$\mathbf{0.665 \pm 0.004}$	$\mathbf{0.652 \pm 0.005}$	$\mathbf{0.758 \pm 0.004}$	$\mathbf{0.671 \pm 0.007}$

Results for Social Image Tag Assignment

With the learned prediction matrix $\mathbf{W}$ and the parameters of the deep architecture $\mathbf{W}_l$, $\mathbf{b}_l$ $(1 \leq l \leq L)$, we can compute the relevance scores between any image and tags, that is, we can assign the relevant semantic tags to any image. Now in this section, experiments are conducted on the MIRFlickr and NUS-WIDE datasets to verify the effectiveness of the proposed method for tag assignment. For LR and LRES, we use the KNN voting strategy (Li, Snoek, and Worring 2009) to find the relevant tags of images. The corresponding quantitative results are shown in Table 2.

From the compared results, it can be seen that the proposed method achieves the best performance on both the MIRFlickr and NUS-WIDE datasets for the task of image tag assignment. By seamlessly incorporating the visual information and the semantic information, the proposed method achieves significant improvement over other methods for assigning tags to new images. That is, the motivation of this work is empirically verified.

In addition, to better empirically evaluate the effectiveness of the proposed method, experiments on these two datasets for the task of tag-based image retrieval are carried out. The performance is measured in terms of MAP. Figure 2 and Figure 3 illustrate the corresponding quantitative results on the MIRFlickr and NUS-WIDE datasets, respectively. From the experimental results, it can be observed that the proposed method achieves the best results on both datasets for image retrieval. The advantages of the proposed method are verified again.

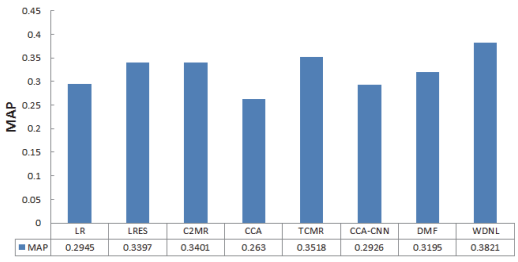

Figure 2: Retrieval results on MIRFlickr in terms of MAP.

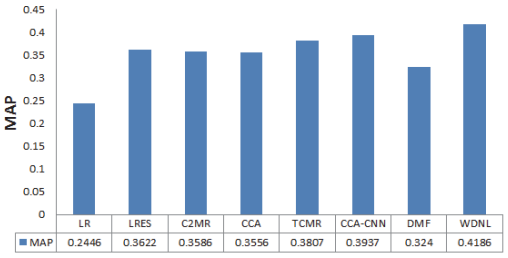

Figure 3: Retrieval results on NUS-WIDE in terms of MAP.

Conclusion

In this paper, we propose a weakly supervised social image tag refinement and tag assignment method via the deep non-negative low-rank model. The visual features and the high-

level tags are connected by the deep architecture. The tag refinement and the learning of parameters are jointly implemented, which makes the proposed method have good scalability. Extensive experiments are conducted on two widely used datasets and the experimental results show the advantages of the proposed method for tag refinement and assignment. In future, we will focus on uncovering the latent structures of data and incorporating it into the proposed model in this work. How to extract representations from raw pixels based on the proposed model is also our future work.

Acknowledgments

This work was partially supported by the 973 Program of China (Project No. 2014CB347600), the National Natural Science Foundation of China (Grant No. 61672304, 61672285, 61522203 and 61402228) and National Youth Top-notch Talent Support Program in China.

References

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *JMLR* 3:1107–1135.

Bengio, Y. 2009. Learning deep architectures for ai. *Foundations and Trends in Machine Learning,* 2(1):1–127.

Candès, E. J.; Li, X.; Ma, Y.; and Wright, J. 2011. Robust principal component analysis? *Journal of the ACM* 58(3):1–37.

Chen, M.; Zheng, A.; and Weinberger, K. Q. 2013. Fast image tagging. In *ICML*.

Feng, Z.; Feng, S.; Jin, R.; and Jain, A. K. 2014. Image tag completion by noisy matrix recovery. In *ECCV*, 424–438.

Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; and Ioffe, S. 2013. Deep convolutional ranking for multilabel image annotation. *CoRR* abs/1312.4894.

Huiskes, M., and Lew, M. 2008. The mir flickr retrieval evaluation. In *ACM CIMR*.

Johnson, J.; Ballan, L.; and Fei-Fei, L. 2015. Love thy neighbors: Image annotation by exploiting image metadata. In *ICCV*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.

Li, Z., and Tang, J. 2015a. Unsupervised feature selection via nonnegative spectral analysis and redundancy control. *IEEE TIP* 24(12):5343–5355.

Li, Z., and Tang, J. 2015b. Weakly supervised deep metric learning for community-contributed image retrieval. *IEEE TMM* 17(11):1989–1999.

Li, Z., and Tang, J. 2016. Weakly-supervised deep matrix factorization for social image understanding. *IEEE TIP*. Accepted.

Li, Z.; Liu, J.; Zhu, X.; Liu, T.; and Lu, H. 2010. Image annotation using multi-correlation probabilistic matrix factorization. In *ACM MM*.

Li, Z.; Liu, J.; Tang, J.; and Lu, H. 2015. Robust structured subspace learning for data representation. *IEEE TPAMI* 37(10):2085–2098.

Li, X.; Snoek, C.; and Worring, M. 2009. Learning social tag relevance by neighbor voting. *IEEE TMM* 11(7):1310–1322.

Lin, Z.; Chen, M.; and Ma, Y. 2009. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. Technical Report UILU-ENG-09-2215, University Illinois at Urbana-Champaign, UIUC Technical Report.

Makadia, A.; Pavlovic, V.; and Kumar, S. 2010. Baselines for image annotation. *IJCV* 90(1):88–105.

Murthy, V. N.; Maji, S.; and Manmatha, R. 2015. Automatic image annotation using deep learning representations. In *ACM ICMR*.

Qi, G.-J.; Aggarwal, C.; Tian, Q.; Ji, H.; and Huang, T. 2012. Exploring context and content links in social media: A latent space method. *IEEE TPAMI* 34(5):850–862.

Sun, L.; Ji, S.; and Ye, J. 2011. Canonical correlation analysis for multilabel classification: A least-squares formulation, extensions and analysis. *IEEE TPAMI* 33(1):2194–200.

Tang, J.; Hong, R.; Yan, S.; Chua, T.-S.; Qi, G.-J.; and Jain, R. 2011. Image annotation by knn-sparse graph-based label propagation over noisily tagged web images. *ACM TIST* 2(2):14: 1–15.

Tang, J.; Shu, X.; Qi, G.-J.; Li, Z.; Wang, M.; Yan, S.; and Jain, R. 2016a. Tri-clustered tensor completion for social-aware image tag refinement. *IEEE TPAMI*. Accepted.

Tang, J.; Shu, X.; Li, Z.; Qi, G.-J.; and Wang, J. 2016b. Generalized deep transfer networks for knowledge propagation in heterogeneous domains. *ACM TOMM*. Accepted.

Wang, X.-J.; Zhang, L.; Jing, F.; and Ma, W.-Y. 2010. Annosearch: Image auto-annotation by search. In *CVPR*.

Wong, R., and Leung, C. 2008. Automatic semantic annotation of real-world web image. *IEEE TPAMI* 30(11):1933–1944.

Wu, L.; Jin, R.; and Jain, A. 2013. Tag completion for image retrieval. *IEEE TPAMI* 35(3):716–727.

Yang, L.; Jing, L.; and Ng, M. K. 2015. Robust and non-negative collective matrix factorization for text-to-image transfer learning. *IEEE TIP* 24(12):4701–4714.

Zhao, R., and Grosky, W. 2002. Narrowing the semantic gap-improved text-based web document retrieval using visual features. *IEEE TMM* 4(2):189–200.

Zhu, G.; Yan, S.; and Ma, Y. 2010. Image tag refinement towards low-rank, content-tag prior and error sparsity. In *ACM MM*.

Znaidia, A.; Borgne, H. L.; and Hudelot, C. 2013. Tag completion based on belief theory and neighbor voting. In *ACM ICMR*.