Semi-supervised Relational Topic Model for Weakly Annotated Image Recognition in Social Media

Zhenxing Niu
Xidian University
zhenxingniu@gmail.com

Gang Hua
Stevens Institute of Technology
ghua@stevens.edu

Xinbo Gao
Xidian University
xbgao@ieee.org

Qi Tian
University of Texas at San Antonio
qitian@cs.utsa.edu

Abstract

In this paper, we address the problem of recognizing images with weakly annotated text tags. Most previous work either cannot be applied to the scenarios where the tags are loosely related to the images; or simply take a pre-fusion at the feature level or a post-fusion at the decision level to combine the visual and textual content. Instead, we first encode the text tags as the relations among the images, and then propose a semi-supervised relational topic model (ss-RTM) to explicitly model the image content and their relations. In such way, we can efficiently leverage the loosely related tags, and build an intermediate level representation for a collection of weakly annotated images. The intermediate level representation can be regarded as a mid-level fusion of the visual and textual content, which is able to explicitly model their intrinsic relationships. Moreover, image category labels are also modeled in the ss-RTM, and recognition can be conducted without training an additional discriminative classifier. Our extensive experiments on social multimedia datasets (images+tags) demonstrated the advantages of the proposed model.

1. Introduction

With the ever popularity of social networks (*e.g*. Facebook) and content-sharing websites (*e.g*. Flickr), images are often accompanied by text tags. Although these text tags are noisy in nature, due to the fact that they are annotated by a large group of heterogeneous users, it is commonly acknowledged that they may still provide beneficial information for image recognition. The question is then how such noisy tags could be leveraged to benefit image recognition?

To treat the visual content and the text tags as two different modalities, many methods have been proposed to combine them for better image recognition. Some methods focused on modeling the joint distribution of image content and the associated keywords [1, 7, 2, 3]. In [1, 7], the process of building the relationship between the visual features and the keywords was analogous to a language translation. Further, Corr-LDA was proposed to extend this approach through a hierarchical probabilistic mixture model for image annotation [2]. Recently, image recognition and annotation are simultaneously considered in [18] by modeling the joint distribution of image content, annotation keywords, and class labels.

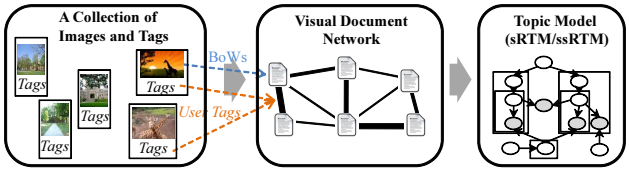


Figure 1. Image recognition with network structured topic models. Firstly, a Visual Document Network (VDN) is constructed with image relations defined by their associated text tags; and then the VDN are modeled with the proposed s-RTM or ss-RTM.

However, these methods assumed that there were explicit correspondences between keywords and image regions, and focused on image annotation rather than image recognition. Thus, they can only be applied to the case where all the keywords have a visual interpretation rather than realistic scenarios where the images are weakly annotated, *i.e*.,the tags are loosely related to images [11]. For example, a photo for the 'Lincoln Memorial' could have a tag 'National Mall' since the Lincoln Memorial is located at the National Mall street, but there does not exist a correspondence between the photo content and such tag.

Alternatively, some discriminative methods were proposed to fuse the visual and textual features for image recognition. For example, textual features were concatenated with visual features to train an SVM classifier for the recognition of touristic landmarks in [12]. In [19], two separate classifiers were built, one from the textual features, and the other one from the visual features. Then a third classifier was trained to combine the confidence values of these two different classifiers for the final prediction. Guillaumin

et al. [10] proposed to use a semi-supervised learning algorithm to explore both labeled and unlabeled images, where visual and textual features were combined under the Multiple Kernel Learning (MKL) framework.

However, these methods either take a pre-fusion at the feature level (*i.e.* the visual and textual features are concatenated together) or a post-fusion at the decision level (*i.e.* the classification scores from the two different modalities are combined) to combine the visual and textual content. Thus, their intrinsic relationships are neglected by these methods.

To address the shortcomings of previous work, we propose an approach to efficiently leverage loosely related tags, and explicitly model the intrinsic relationships between the visual and textual content as combining them for image recognition. As shown in Figure 1, in our formulation each image is represented as a visual document by using a bag-of-words representation, meanwhile text tags are leveraged to define the relations between each pair of images. As a result, a visual document network (VDN) is constructed where the nodes indicate images and links indicate image relations.

After that, we build an intermediate level representation in a joint latent space by modeling the VDN with a network structured topic model, which jointly models the image content and their relations. Such an intermediate level representation can be regarded as a mid-level fusion between the visual and textual content. In particular, images are represented as topic mixtures in a latent space by analyzing their visual content, meanwhile two images with shared common tags are encouraged to have similar representations. Therefore, the intrinsic relationships between the visual and textual content are explicitly modeled as building the image representation.

Recently, some network structured topic models such as the Relational Topic Model (RTM) [6] have been proposed to model a document network. However, the original RTM model is an unsupervised model, and an additional discriminative classifiers (*e.g.*, SVM classifiers) is required to conduct the final recognition tasks.

To effectively model the VDN and leverage the discriminative labels, we first extend the RTM to a supervised model, namely supervised RTM (s-RTM), where the image category labels are incorporated into the process of topic modeling. Therefore, image content, their relations and their category labels could be jointly modeled. Furthermore, to effectively exploit the relations between training and testing images, a transductive learning model, *i.e.* a semi-supervised RTM (ss-RTM) is proposed to jointly model the training and testing images, as shown in Figure 3.

Although many loosely related tags do not directly correspond to the image content, two images usually have a certain relationship (*e.g.*, contain a same object) if they share common properties (*e.g.*, text tags) [15]. For example, two

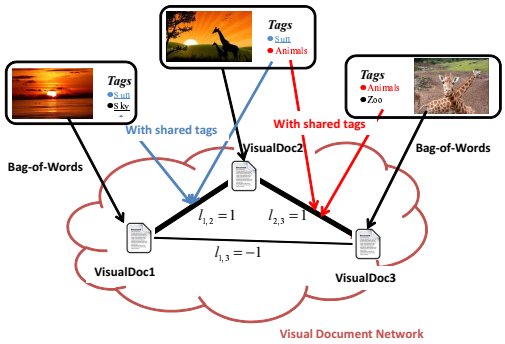


Figure 2. Construction of visual document network, where each node indicates an image and each link indicates a pair-wise image relation. The positive (*i.e.*, with shared tags) and negative (*i.e.*, without shared tags) relations are denoted as bold and thin links respectively.

photos tagged with a common tag (*e.g.*, 'National Mall') are more likely to contain a same object (*e.g.*, 'Lincoln Memorial'). Therefore, we establish the pair-wise image relations by using text tags rather than modeling the direct correspondences between image content and tags.

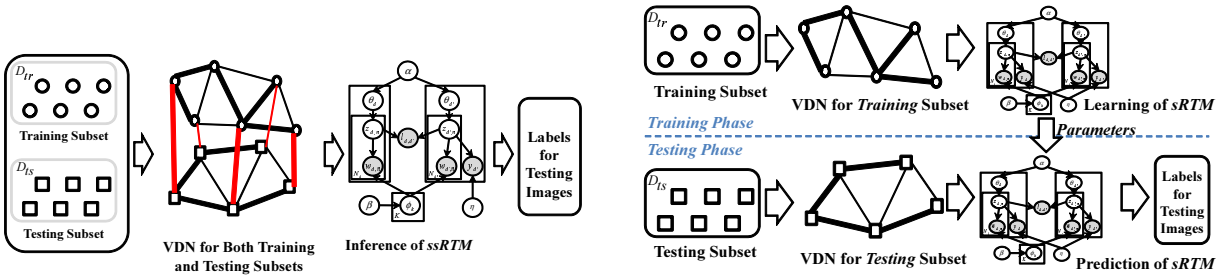
In particular, two kinds of image relations are defined by using the text tags, *i.e.*, if two images share the common tags (both of them are annotated with one or more common tags), we define that they have a *positive* relation which indicates they are more likely to be from the same image category; otherwise we define that they have a *negative* relation, as shown in Figure 2. Thus, our approach is not constrained by the assumption that tags should exactly correspond to image regions, and hence can efficiently leverage loosely related tags. Another advantage of encoding tags as image relations is that, it can tolerate incorrect tags since the probability for two image share a common incorrect tag is relatively small.

In summary, our contributions are two-folds. First, we propose to exploit visual and textual content to form an intermediate level representation for visual recognition through principled probabilistic modeling. Second, two network structured topic models, *i.e.* s-RTM and ss-RTM are proposed which simultaneously model the image content, their relations, and their category labels.

2. Related Work

Topic models were originally proposed for document understanding [4], and have been successfully adapted to image understanding and recognition. Fei-Fei [8] and Bosch [5] exploited LDA and pLSA for scene recognition respectively. Niu [16] presented a context aware topic model for scene category recognition. Wang *et al.* [18] extended a supervised topic model sLDA [3] for simultaneous image classification and annotation. However, these methods simply neglect the relations among images.

To the best of our knowledge, our approach is the first



(a) Image recognition with ss-RTM (b) Image recognition with s-RTM

Figure 3. The comparison of image recognition between ss-RTM and s-RTM. The image relations within the training and within the testing subsets are explored separately in s-RTM, whereas the relations within and between (denoted as red links in Figure 3(a)) the training and testing subsets are both explored in ss-RTM.

one that explicitly models the image relations by using a network structured topic model. Recently, some multi-label joint learning methods were proposed to explicitly model the relations among image categories [17]. In contrast, our approach focuses on modeling the relations among image instances which come from different image categories. So it is a fine-grained joint learning method compared with these multi-label learning methods.

There are some other methods leveraged multi-modality information for annotation [14, 13, 20]. However, these work focus on building a graph with images and tags, and propagating tags for image annotation rather than image category recognition.

From the perspective of learning a joint latent space for the visual and textual modalities, another method [11] is related to our approach. However, it focuses on an image retrieval scenario where a narrative text is loosely related to an image and where only a few image-text pairs are available. In contrast, for our task the image-text pairs are easily collected but a long narrative text is unavailable.

3. Overview of Our Approach

As shown in Figure 1, the proposed method for weakly annotated image recognition has two stages. Specifically, for each image $d, d \in D$, its local features (*e.g.*, Dense-SIFT) are extracted, and it is represented as a visual document $\{w_{d,n}, n = 1, 2, ..., N_d\}$ by a bag-of-words representation. This visual document is treated as a node in the VDN. Meanwhile, the image relations are encoded as links in VDN. Similar to the original RTM model, the relation between image d and d' is also described by a binary variable $l_{d,d'}$. The $l_{d,d'}$ is defined to indicate whether d and d' come from a same image category, and is modeled according to whether d and d' share common tags. Specifically, if d and d' share common user tags, we define they have a *positive* relation $l_{d,d'} = 1$; otherwise we define they have a *negative* relation $l_{d,d'} = -1$, as shown in Figure 2.

It is noticed that we only encode the relations for a subset of image pairs in our model for two reasons: 1) due to the noisy or missing tags, some image relations cannot be correctly modeled. Thus, only some reliable image relations should be selected to construct the VDN; 2) if all image relations are selected, the VDN is a full-connected network and it will significantly increase the computational cost of topic modeling.

Next, regards the stage of topic modeling, we have two options, *i.e.*, ss-RTM and s-RTM, which depends on whether the training and testing images are available at the same time. If they are, we can construct one VDN with both the training and testing images, and model the VDN with a ss-RTM, where the labels for testing images can be directly predicted by conducting the inference on the ss-RTM, as shown in Figure 3(a). Otherwise, we first construct a VDN with the training images, and learn a s-RTM model; when the testing images are available, another VDN is constructed with the testing images, and their category labels are predicted with the learnt s-RTM, as shown in Figure 3(b).

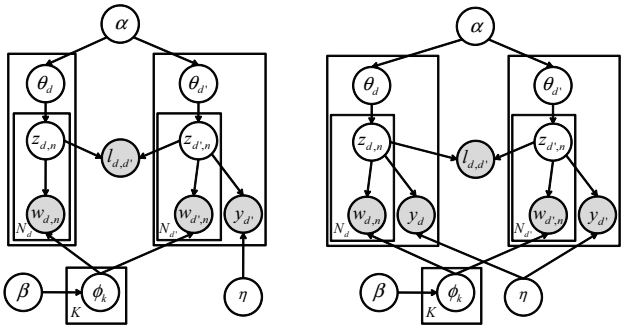
Obviously, more image relations (*i.e.*, relations between the training and testing images) will be modeled in the ss-RTM, thus the ss-RTM should achieve better recognition performance over the s-RTM, which will be demonstrated in the Section 6. The details for the ss-RTM and s-RTM will be presented in the Section 4 and 5 respectively.

4. Semi-supervised RTM (ss-RTM)

Similar to the RTM [6], the ss-RTM is a generative probabilistic model to model a network of visual documents. Different from the RTM, the category labels of images are considered in the ss-RTM.

As we know, the image category labels were considered and efficiently modeled in the sLDA [3]. Inspired by it, the image category labels are incorporated into the ss-RTM in a similar way. Specifically, the task of image category recognition is formulated as a binary classification problem in a One-vs-Other fashion in this paper. So, we build one ss-RTM for each image category to distinguish it from other categories, and use a binary random variable y_d to describe the image category label, where $y_d = 1$ indicates the image d comes from this category and $y_d = -1$ indicates it comes from one of other categories.

For each image category, its training set D_{tr} and test-



(a) ss-RTM (b) s-RTM

Figure 4. The graphical models for ss-RTM and s-RTM. They only illustrate the graphical model for a single pair of images. For the ss-RTM, it models an image network where only some image labels are known, *e.g.*, the d' indicates a training image with label $y_{d'}$ and the d indicates a testing image without label.

ing set D_{ts} constitute its entire image set $D = D_{ts} + D_{tr}$. The image set D, the set of image relations $L = \{l_{d,d'}|d \neq d', d \in D, d' \in D\}$, and the category labels of training images $\{y_d\}_{d \in D_{tr}}$ are modeled by a ss-RTM. The generative process is as follow:

1. For each topic k:
   (a) Draw topic distribution over the visual vocabulary $\phi_k \sim \text{Dir}(\beta)$.
2. For each visual document $d \in D$:
   (a) Draw topic proportions $\theta_d \sim \text{Dir}(\alpha)$.
   (b) For each visual word $w_{d,n}$:
      i. Select a topic $z_{d,n} \sim \text{Multi}(\theta_d)$.
      ii. Draw a visual word $w_{d,n} \sim \text{Multi}(\phi_{z_{d,n}})$.
   (c) If y_d is observed (*i.e.*, $d \in D_{tr}$), draw image category label y_d
   $$y_d|\bar{z}_d, \eta \sim \rho(y_d|\bar{z}_d, \eta).$$
3. For each observed link $l_{d,d'} \in L$:
   (a) Draw a link indicator $l_{d,d'}$:
   $$l_{d,d'}|\bar{z}_d, \bar{z}_{d'} \sim \psi(l_{d,d'}|\bar{z}_d, \bar{z}_{d'}).$$

where the parameters for the ss-RTM model are K distributions over the visual vocabulary $\phi_{1:K}$, a K-dimensional Dirichlet parameter α, a function ψ that provides link probabilities, and a function ρ that provides label probabilities.

Figure 4(a) illustrates the graphical model for this process for a single pair of documents, where the d' indicates a training image with label $y_{d'}$ and the d indicates a testing image without label.

We model the image category labels as that in the sLDA [3]. Since the category label is model with a binary random variable y_d, the distribution over the y_d is described with a logistic function in this paper, *i.e.*,

$$\rho(y_d = 1|\bar{z}_d, \eta) = \frac{1}{1 + exp(-\eta^T \bar{z}_d)} \quad (1)$$

where η is the parameter of the logistic function, and $\bar{z}_d = (1/N)\sum_{n=1}^{N} z_n$.

To model image relations, we define a specific *link probability function* $\psi(l_{d,d'}|\bar{z}_d, \bar{z}_{d'})$ to describe the image relation $l_{d,d'}$ defined on top of image representations (*i.e.*, $\bar{z}_d, \bar{z}_{d'}$). In the original RTM, two specific link probability functions (*i.e.*, the *logistic regression function* and the *exponential mean function*) were used to model the document relations (*i.e.* document citations). These two functions have the capacity to describe a complex relationship between the document citations and document representations.

Instead of using these two functions, in our model, we assume that two images are more likely to have similar representations if they have a positive relation (*i.e.* share common tags) . Therefore, we define the *histogram intersection* between vectors $\bar{z}_d$ and $\bar{z}_{d'}$ to measure the similarity of representations between d and d', *i.e.*, $s_{d,d'} = \sum_{k=1}^{K} min(\bar{z}_{d,k}, \bar{z}_{d',k})$. And we define the link probability function as,

$$\psi(l_{d,d'}|\bar{z}_d, \bar{z}_{d'}) = \begin{cases} s_{d,d'} & , \quad l_{d,d'} = 1 \\ 1 - s_{d,d'} & , \quad l_{d,d'} = -1 \end{cases} \quad (2)$$

This function can not only properly describe image relation defined on top of the image representations, but also simplify the learning algorithm due to that no parameters to be estimated.

According to the generative process of ss-RTM, the joint distribution of visual words $\boldsymbol{w}$, image relations $\boldsymbol{l}$, the label of training images $\boldsymbol{y}$, topic mixture $\boldsymbol{\theta}$, topic distribution $\boldsymbol{\phi}$, and a set of topics $\boldsymbol{z}$ is given by

$$\begin{aligned} &p(\boldsymbol{w}, \boldsymbol{l}, \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{z}|\alpha, \beta, \eta, \omega) = \\ &\prod_{d \in D} p(\theta_d|\alpha) \prod_{n \in Nd} p(z_{dn}|\theta_d) p(w_{dn}|\phi_{z_{dn}}) \prod_{k \in K} p(\phi_k|\beta) \\ &\prod_{d \in D_{tr}} \rho(y_d|\bar{z}_d, \eta) \prod_{l \in L} (\psi(l_{d,d'}|\bar{z}_d, \bar{z}_{d'}))^{\omega}, \end{aligned} \quad (3)$$

where the first three, the fourth and the fifth item indicate the generation of image visual content, image category label, and image relations respectively.

At last, another parameter ω is introduced in our model to further tune the fifth item, as shown in Eq 3. As we know, for a certain image category, the visual content and text tags usually have different discriminative ability for image recognition. For example, since the object 'bird' usually covers a small region in an image, only a little of discriminative visual features can be extracted from that region. Thus, for the 'bird' category, visual content do not have enough discriminative ability. But for the 'sunset' category, the situation is opposite since it is about a scene that describes what a whole image looks like.

Since the text tags have been encoded as the image relations, we need a parameter ω to trade off the effect between

Table 1. Inference of ss-RTM

	Input: $\{w_d\}_{d\in D}, \{y_d\}_{d\in D_{tr}}$ and $\{l\}_{l\in L}$ Output: $\{y_d\}_{d\in D_{ts}}$
1.	Inference of *sLDA* with $\{w_d\}_{d\in D}, \{y_d\}_{d\in D_{tr}}$ and initialize $\eta^0 = \eta^{sLDA}$;
2.	Loop $i = 1, 2, ..., N$: (2.1) Given η^{i-1}, do Gibbs sampling as Eq 4 and get $\{z_d^i\}_{d\in D}$; (2.2) Given $\{z_d^i, y_d\}_{d\in D_{tr}}$, do logistic regression as Eq 1 and estimate η^i;
3.	Given $\eta^* = \eta^N$ and $\{z_d^N\}_{d\in D_{ts}}$, predict the category labels $\{y_d\}_{d\in D_{ts}}$ as Eq 1 .

the visual content and image relations for the image recognition. From Eq 3, we can assign a large value to ω if the image relations (*i.e.*, text tags) are more discriminative for recognition, and vice versa.

4.1. Inference

Given the topic assignments of testing images $\{z_d\}_{d\in D_{ts}}$ and the parameter of logistic function η, the category labels of testing images can be predicted with Equation 1. Thus, the problem of image recognition with the ss-RTM is formulated as the inference of topic assignments $\{z_d\}_{d\in D}$ and the estimation of parameter η.

If the parameters of the ss-RTM are known, the topic assignments can be inferred by the Gibbs sampling [9]. However, since the topic assignments $\{z_d\}_{d\in D}$ and the parameter η are both unknown, we carry out an algorithm to estimate them iteratively, *i.e.*, fix one and estimate another iteratively, as shown in Table 1.

In particular, the parameter η is firstly initialized by the inference of the sLDA model [3]. Then, $\{z_d\}_{d\in D}$ and η are estimated iteratively. Specifically, because only the category labels of training images are known, the η is estimated with $\{z_d^i, y_d\}_{d\in D_{tr}}$. After several iterations upon convergence, the optimal value of parameter η^* can be estimated, and the corresponding topic assignments $\{z_d\}_{d\in D}$ are inferred, too. In the end, the category labels of testing images are predicted according to Equation 1.

According to the generative process of the ss-RTM, the Gibbs sampling equation for the ss-RTM can be derived in a similar way as the RTM, *i.e.*,

$$p(z_{d,n} = k|z^{-dn}, \boldsymbol{w}, \boldsymbol{l}, \boldsymbol{y}) = (\alpha + m_{d,k}^{-dn})\frac{n_{k,w_{d,n}}^{-dn} + \beta}{\sum_w n_{k,w}^{-d,n} + K\beta}$$
$$\frac{\rho(y_d|z_d, \eta)}{\rho(y_d|z_d^{-dn}, \eta)} \prod_{l\in L} \left(\frac{\psi(l_{d,d'}|z_d, z_{d'})}{\psi(l_{d,d'}|z_d^{-dn}, z_{d'})} \right)^{\omega}, \quad (4)$$

where $m_{d,k}^{-dn}$ stands for the number of times that topic k is assigned to visual document d except for $z_{d,n}$, and $n_{k,w}^{-dn}$ indicates the number of times that topic k is assigned to w except for $z_{d,n}$.

Table 2. Learning of s-RTM

	Input: $\{w_d, y_d\}_{d\in D_{tr}}$, and $\{l\}_{l\in L_{tr}}$ Output: $\phi_{1:K}^*$ and η^*
1.	Inference of *sLDA* with $\{w_d, y_d\}_{d\in D_{tr}}$ and initialize $\eta^0 = \eta^{sLDA}$;
2.	Loop $i = 1, 2, ..., N$: (2.1) Given η^{i-1}, do Gibbs sampling as Eq 8 and get $\{z_d^i\}_{d\in D_{tr}}$; (2.2) Given $\{z_d^i, y_d\}_{d\in D_{tr}}$, do logistic regression as Eq 1 and estimate η^i;
3.	Given $\{z_d^N\}_{d\in D_{tr}}$, evaluate $\phi_{1:K}^*$ as Eq 7 and output $\eta^* = \eta^N$

Furthermore, according to Equation 1, the last term in Equation 4 can be computed as

$$\frac{\rho(y_d|z_d, \eta)}{\rho(y_d|z_d^{-dn}, \eta)} = \frac{1}{1 + exp(-(y_d)(\eta^T \bar{z}_d^{-dn} + \frac{\eta_k}{N_d}))} \quad (5)$$

5. Supervised RTM (s-RTM)

As aforementioned, when the testing images cannot be obtained with the training images at the same time, we should conduct the image recognition with the s-RTM model, which consists of two phases: at training phase, we firstly construct a VDN from the training images, learn a s-RTM for the VDN, and estimate its model parameters (*i.e.*, ϕ_k, η); at testing phase, another VDN is constructed with the testing images, and the category labels for testing images are predicted with the learnt s-RTM.

5.1. Learning

Given the training data, the learning of s-RTM can be formulated as

$$\{\phi_{1:K}^*, \eta^*\} = \underset{\phi_{1:K},\eta}{\operatorname{argmax}}\, p(\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{l}|\phi_{1:K}, \eta), \quad (6)$$

where $\boldsymbol{w}$ and $\boldsymbol{y}$ indicate the visual documents and category labels for the training images, and $\boldsymbol{l}$ indicate the relations among training images.

Since there are two kinds of parameters to be estimated, *i.e.*, topic-word distribution $\phi_{1:K}$ and the parameter η of the logistic function. We still carry out an algorithm to estimate them iteratively, as shown in Table 2.

Generally, we can directly estimate $\phi_{1:K}$ with the topic assignments $\{z_d\}_{d\in D_{tr}}$ as

$$\phi_k = \frac{n_{w,k} + \beta}{\sum_k n_{w,k} + W\beta}, \quad (7)$$

where $n_{k,w}$ indicates the number of times that topic k is assigned to w. So, we only need to infer $\{z_d\}_{d\in D_{tr}}$ at the iteration step and estimate $\phi_{1:K}^*$ at last.

Particularly, firstly the η is again initialized by the inference of the sLDA model; then the topic assignments

$\{z_d\}_{d\in D}$ and the η are estimated iteratively; after several iterations upon convergence, the optimal value $\phi^*_{1:K}$ and η^* are obtained at last.

Since the images relations are specific to the training images L_{tr} for the s-RTM learning, the Gibbs sampling equation for the s-RTM is modified as

$$p(z_{d,n}=k|z^{-dn},\boldsymbol{w},\boldsymbol{l},\boldsymbol{y}) = (\alpha + m^{-dn}_{d,k})\frac{n^{-dn}_{k,w_{d,n}}+\beta}{\sum_w n^{-d,n}_{k,w}+K\beta}$$
$$\frac{\rho(y_d|z_d,\eta)}{\rho(y_d|z^{-dn}_d,\eta)}\prod_{l\in L_{tr}}\left(\frac{\psi(l_{d,d'}|z_d,z_{d'})}{\psi(l_{d,d'}|z^{-dn}_d,z_{d'})}\right)^{\omega} \quad (8)$$

5.2. Prediction

At the testing phase, we focus on applying the learnt s-RTM for the prediction of category labels of testing images. Specifically, we firstly do Gibbs sampling on the RTM with the testing data (*i.e.*, $\{w_d\}_{d\in D_{ts}}$ and $\{l\}_{l\in L_{ts}}$), where all category labels are unknown, and the parameter $\phi_{1:K}$ is fixed as the learnt $\phi^*_{1:K}$. After that, the topic assignments of testing images $\{z_d\}_{d\in D_{ts}}$ are inferred. With the inferred $\{z_d\}_{d\in D_{ts}}$ and the learnt η^*, the category label of a testing image can be predicted according to Equation 1.

6. Evaluation

To evaluate the performance of the proposed methods, we conduct some experiments on two social media datasets, which are described below:

- **NUS-WIDE**: It contains 269,648 images which are crawled from Flickr website. The crawled images are linked to 1,000 different user tags, which are annotated by users registered in Flickr. Beyond these images and user tags, 81 concepts are defined in the dataset.
- **MIRFLICKR-25k**: It contains 25,000 images which are also crawled from Flickr website. In the collection there are 1,386 tags which occur in at least 20 images. And 23 potential labels are defined in the dataset.

6.1. Experimental setting

The 81 concepts and 23 potential labels are regarded as different image categories for the evaluation of image category recognition in this paper. For each image category, we generate its training and testing subsets in an One-vs-Other fashion, *i.e.*, we randomly select N images from its category as positive samples, and randomly select N images from other categories as negative samples.

The number of image relations increases rapidly when more images are considered. Even through our algorithm only considers a portion of image relations, it still takes a long time to conduct image recognition with all images in the dataset. So, we set $N = 500$ in this paper to evaluate our approach. Furthermore, to evaluate the stability of our algorithm, we repeat the process 50 times independently. Thus, 50 independent training and testing subsets are generated for each image category. The algorithms are evaluated on each subset, and the average performance on the 50 subsets is regarded as the final performance of the algorithm.

In addition, to estimate the optimal value of parameter ω for each category, we use a 5-fold cross validation on their training subsets.

For each image, we densely extract SIFT features from 10×10 image patches. These SIFT features are quantized to form a visual codebook of size 500. For tags, following [10], the 457 most frequently used tags are leveraged to form a tag codebook.

Table 3. The performances of all the competing algorithms on the two datasets, and the performance is evaluated in terms of AP.

Methods	NUS-WIDE	MIRFLICKR-25k
BoW+SVM	$70.8\% \pm 0.2\%$	$72.9\% \pm 0.3\%$
Tag+SVM	$74.4\% \pm 0.3\%$	$73.8\% \pm 0.4\%$
BoW+Tag+SVM	$75.1\% \pm 0.2\%$	$74.2\% \pm 0.3\%$
BoW+Tag+MKL	$76.2\% \pm 0.2\%$	$77.4\% \pm 0.3\%$
LDA+SVM	$72.3\% \pm 0.1\%$	$73.1\% \pm 0.2\%$
sLDA	$72.8\% \pm 0.1\%$	$73.8\% \pm 0.2\%$
RTM+SVM	$74.1\% \pm 0.2\%$	$75.1\% \pm 0.3\%$
s-RTM	$80.2\% \pm 0.2\%$	$78.3\% \pm 0.4\%$
ss-RTM	$\mathbf{84.1}\% \pm 0.2\%$	$\mathbf{81.1}\% \pm 0.4\%$

6.2. Image recognition

We compare our method with two categories of image recognition methods. The first category is discriminative methods [10]. Specifically, local features are extracted from those images and each image is represented as a BoWs vector. Meanwhile, each image can also be represented as a tag vector by using its associated tags. Thus the image recognition can be conducted in four ways: using an SVM classifier with only the BoWs or Tag vector (denoted as 'BoWs+SVM' or 'Tag+SVM' respectively); using an SVM classifier with the vector generated by concatenating BoWs and Tag vectors together, *i.e.*, in a pre-fusion way (denoted as 'BoWs+Tag+SVM'); and using Multiple Kernel Learning (MKL) to fuse both BoWs and Tag vectors, *i.e.*, in a post-fusion way (denoted as 'BoWs+Tag+MKL').

The other category is based on topic models, which includes LDA [4], RTM [6], and sLDA [3]. Specifically, LDA and RTM are unsupervised topic models and a binary linear SVM classifier is used to conduct the final recognition based on their representation vectors. The sLDA is a supervised method and is directly employed for image recognition.

Table 3 illustrates the performances of all the competing algorithms on the two datasets. Obviously, both 'BoWs+Tag+SVM' and 'BoWs+Tag+MKL' consider the

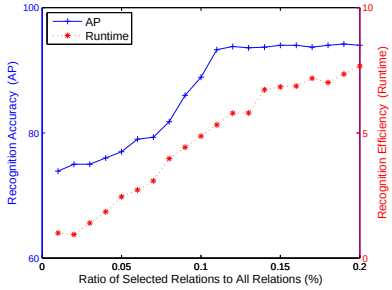


Figure 5. The recognition accuracy and efficiency for the selection of image relations. The x-axis indicates the ratio for the selected relations to all relations (from 1% to 20%), the recognition accuracy is measured in terms of AP, and the recognition efficiency is measured in terms of the relative runtime of the recognition algorithm (*i.e.*, the runtime for the ratio 1% is defined as unit '1').

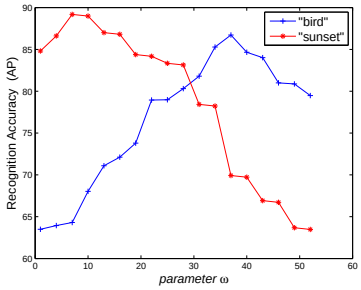


Figure 6. Tradeoff between image content and their relations. Obviously, the optimal value of ω for image category 'bird' is relatively large. in contrast, it is relatively small for image category 'sunset'.

image features and tags at the same time, and hence achieve better recognition performance over 'BoWs+SVM' and 'Tag+SVM'. On the other hand, the RTM, s-RTM, and ss-RTM jointly consider the image content and their relations, and achieve better performance over 'LDA' and 'sLDA'. Furthermore, since image label is incorporated into the process of topic modeling, the s-RTM and ss-RTM not only achieve better performance over the RTM, but also can be used to conduct recognition without an SVM classifier.

Obviously, the 'BoWs+Tag+MKL' performs better over the 'BoWs+Tag+SVM', which indicates the post-fusion based method, *i.e.*, multiple kernel learning can better weight and leverage the visual and textual content compared with the pre-fusion based method. Furthermore, we build an intermediate level representation, and it can be regarded as a mid-level fusion of the visual and textual modality, which performs better than both the pre-fusion and post-fusion based methods.

More importantly, because the performance of ss-RTM is improved 3% or 4% on the two datasets compared with the s-RTM, we can see that the relations between the training and testing images are more helpful for image recognition. The detailed comparisons among all these methods in terms of AP over each individual category on the two datasets are illustrated in Figure 7 and 8 respectively.

6.3. Discussion

6.3.1 The selection of image relations

To reduce the impact of noisy tags and the computational cost of topic modeling, we only select some reliable image relations to model.

Obviously, the number of shared tags usually indicates the reliability of an image relation. In other word, the larger the number of shard tags is, the more reliable the relation is. So, we propose a relation selection scheme based on the number of shared tags in this paper. Specifically, for each image d in VDN, the reliability of relations $s_{d,d'}$ between it and another image d' is described by the number of shared tags between them, and they are sorted in an ascending order. Thus, if we want to select M relations at last, the top $M/2$ relations in the sorted list are selected as positive relations. And the $M/2$ random relations with zero shared tags (*i.e.*, $s_{d,d'} = 0$) are selected as negative relations.

In Figure 5, take the image category 'airport' in the NUS-WIDE dataset as an example. Obviously, the accuracy increases rapidly when the percentage of selected relations is relatively small, and will quickly reach a saturation point. In contrast, the runtime of the recognition algorithm increases linearly. So there is a upper bound for recognition performance, which cannot be improved by simply increasing the number of relations modeled. As a result, we just need to select a portion of reliable image relations to ensure the recognition accuracy and simultaneously reduce the computational cost.

Even through, it takes a long time for our inference algorithm to conduct image recognition with all images in the dataset. We will exploit some strategies to improve the scalability of our method in future work.

6.3.2 Tradeoff between image content and relations

In the proposed models, both visual and textual content are incorporated into topic modeling. However, for a certain image category, they usually have different discriminative ability for image recognition. Obviously, for the image category about an small object such as 'bird', text tags are more discriminative; for the image category about a scene such as 'sunset', visual content are more discriminative.

As aforementioned, a parameter ω is introduced to trade off their impact on image recognition. In particular, we should assign a large value to ω if the text tags are more discriminative, and vice versa. As shown in Figure 6, the optimal value of ω for image category 'bird' is larger than that for image category 'sunset'. And the optimal value of parameter ω is estimated by a 5-fold cross validation.

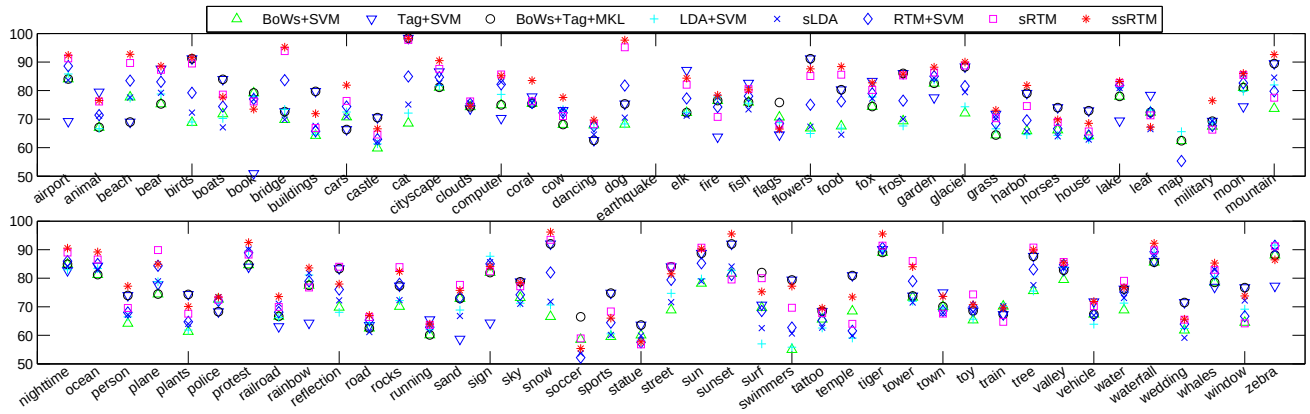


Figure 7. Comparison of different algorithms over 81 concepts on NUS-WIDE dataset in terms of AP.

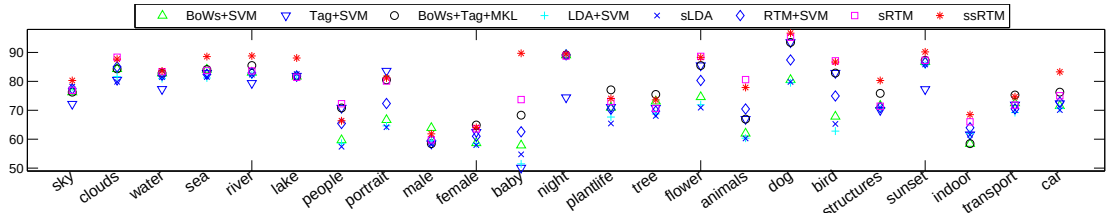


Figure 8. Comparison of different algorithms over 23 concepts on MIRFLICKR-25k dataset in terms of AP.

7. Conclusion and Future Work

In this paper, visual content and text tags are leveraged together for image recognition in social media. By encoding the text tags as the image relations, the loosely related tags can be efficiently leveraged. By building an intermediate representation with ss-RTM, the visual and textual content can be fused at the mid-level, where their intrinsic relationships are explicitly modeled. Moreover, image category labels are also modeled in the ss-RTM, and recognition can be conducted without an additional discriminative classifier. Our experiments clearly demonstrate the advantages of our approach. Our future work will focus on improving the scalability of our method.

8. Acknowledgements

This research was supported partly by the NSFC (Grant Nos. 61125204, 61172146 and 61201294), the Fundamental Research Funds for the Central Universities (Grant Nos. K5051202048 and BDZ021403), the Program for Changjiang Scholars and Innovative Research Team in University of China (No.IRT13088) and the Shaanxi Innovative Research Team for Key Science and Technology (No.2012KCT-02). Dr. Gang Hua is partly supported by US NSF Grant IIS 1350763, NSFC Grant 61228303, GH's start-up funds from SIT, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs American. Dr. Qi Tian is partly supported by ARO grant W911NF-12-1-0057, Faculty Research Awards by NEC Labs American, 2012 UTSA START-R Research Award, and NSFC Grant 61128007 respectively.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. Freitas, D. Blei, and M. Jordan. Matching words and pictures. In *JMLR*, 2003.

[2] D. Blei and M. Jordan. Modeling annotated data. *SIGIR*, 2003.

[3] D. M. Blei and J. D. McAuliffe. Supervised topic models. *NIPS*, 2007.

[4] D. M. Blei, A. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via plsa. In *ECCV*, 2006.

[6] J. Chang and D. M. Blei. Relational topic models for document networks. In *AISTATS*, 2009.

[7] P. Duygulu, K. Barnard, J. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *PNAS*, 2004.

[10] M. Guillaumin, J. Verbeek, and C. Schmid. Multimodal semi-supervised learning for image classification. In *CVPR*, 2010.

[11] Y. Jia, M. Salzmann, and T. Darrell. Learning cross-modality similarity for multinomial data. In *ICCV*, 2011.

[12] Y. Li, D. Crandall, and D. Huttenlocher. Landmark classication in large-scale image collections. In *ICCV*, 2009.

[13] H. Ma, J. Zhu, M. Lyu, and I. King. Bridging the semantic gap between image contents and tags. In *IEEE Trans. Multimedia.*, 2010.

[14] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV*, 2008.

[15] J. McAuley and J. Leskovec. Image labeling on a network: using social-network metadata for image classification. In *ECCV*, 2012.

[16] Z. Niu, G. Hua, X. Gao, and Q.Tian. Context-aware topic model for scene recognition. In *CVPR*, 2012.

[17] Z. Shi, T. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013.

[18] C. Wang, D. Blei, and L. Fei-Fei. Simultaneous image classification and annotation. In *CVPR*, 2009.

[19] G. Wang, D. Hoiem, and D. Forsyth. Building text features for object image classification. In *CVPR*, 2009.

[20] L. Zheng, S. Wang, Z. Liu, and Q.Tian. Packing and padding: Coupled multi-index for accurate image retrieval. In *CVPR*, 2014.