Multi-Modal Image Annotation with Multi-Instance Multi-Label LDA *

Cam-Tu Nguyen, De-Chuan Zhan, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
{nguyenct, zhandc, zhouzh}@lamda.nju.edu.cn

Abstract

This paper studies the problem of image annotation in a multi-modal setting where both visual and textual information are available. We propose Multi-modal Multi-instance Multi-label Latent Dirichlet Allocation (M3LDA), where the model consists of a visual-label part, a textual-label part and a label-topic part. The basic idea is that the topic decided by the visual information and the topic decided by the textual information should be consistent, leading to the correct label assignment. Particularly, M3LDA is able to annotate image regions, thus provides a promising way to understand the relation between input patterns and output semantics. Experiments on Corel5K and ImageCLEF validate the effectiveness of the proposed method.

1 Introduction

In image annotation and retrieval, one image often has multiple labels owing to its complicated semantics, whereas different image regions often provide different hints for the labels. Therefore, multi-instance multi-label (MIML) learning [Zhou and Zhang, 2007; Zhou *et al.*, 2012] provides a natural formulation, where each example (image) is represented by a bag of instances each corresponding to one region, and the example is associated with multiple labels simultaneously. Labels of the training examples are known, however, labels of instances are unknown. Formally, let $\mathcal{X}$ denote the instance (or feature) space and $\mathcal{Y}$ the set of class labels. Given a training dataset $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$, where $X_i \subseteq \mathcal{X}$ is a set of instances $\{\mathbf{x}_{i1}, \mathbf{x}_{i2}, \ldots, \mathbf{x}_{i,M_i}\}$, $\mathbf{x}_{ij} \in \mathcal{X}$ $(j = 1, \ldots, M_i)$, and $Y_i \subseteq \mathcal{Y}$ is a set of labels $\{y_{i1}, y_{i2}, \ldots, y_{i,L_i}\}$, $y_{ik} \in \mathcal{Y}$ $(k = 1, \ldots, L_i)$, M_i and L_i denotes the number of instances in X_i and the number of labels in Y_i, respectively, the goal is to predict Y for unseen example X.

Although the MIML framework captures the information that an image is composed of a set of entities rather than a single entity, in its original form, MIML does not consider information from other modalities. In many real tasks such as web image annotation, in addition to the visual information in images, the surrounding texts or user tags for images are also helpful. Particularly, with the rapid evolvement of social networks and online services such as Facebook or Flickr, more and more human collaborative tags are accumulated. By exploiting the visual and textual information together, better performances can be expected.

In this paper, we extend the standard MIML framework to a multi-modal setting, and propose M3LDA, Multi-modal Multi-instance Multi-label Latent Dirichlet Allocation. In this setting, both visual and textual information are exploited. M3LDA is inspired from Latent Dirichlet Allocation [Blei *et al.*, 2003], and the model consists of a visual-label part, a tag-label part and a label-topic part. The visual-label and textual-label parts are devoted to the mappings from the visual and textual spaces to the label space; the label-topic part helps capture label relationships, i.e., a topic groups highly related labels to form a reasonable visual and tag appearance. Unlike previous settings that exploit topic models with multi-modal data [Blei and Jordan, 2003; Wan *et al.*, 2010], in this paper, we make the distinction between labels and textual modality. Here, we regard a subset of human vocabulary as annotation terms, and refer to this set as the *label set*. A larger vocabulary used in the surrounding texts or tags is referred to as the *tag set*, which is usually quite large and contains a lot of irrelevant terms.

The basic idea behind M3LDA is that the topic decided by the visual information and the topic decided by the tag information should be consistent, leading to the correct label assignment. Moreover, M3LDA is able to give annotations to image regions, providing a promising way to understand the relation between input patterns and output semantics. Although our model works with two modalities, it is not difficult to extend to more modalities or views. Experiments show the advantages of M3LDA over state-of-the-art approaches.

The rest of the paper is organized as follows. Section 2 briefly introduces related work. Section 3 presents the M3LDA approach. Section 4 reports on experiments. Finally, Section 5 concludes the paper.

2 Related Work

During the past few years, many multi-instance multi-label learning approaches have been developed. To name a few, the

*This research was supported by 973 Program (2010CB327903), NSFC (61105043, 61273301), JiangsuSF (BK2011566) and Huawei Fund (YBCB2012085).

MIMLSVM and MIMLBoost approaches work by degenerating a MIML task to a simplified supervised learning task using single-instance multi-label or multi-instance single-label approaches as bridges [Zhou and Zhang, 2007]; the DBA approach formulates MIML as a probabilistic generative model [Yang *et al.*, 2009]; the RankingLoss approach optimizes the label ranking loss for bag and instance annotation [Briggs *et al.*, 2012] etc. Topic models have been applied to multi-modal learning [Blei and Jordan, 2003; Wan *et al.*, 2010; Jain *et al.*, 2007; Putthividhy *et al.*, 2010], multi-label learning [Rosen-Zvi *et al.*, 2004; Rubin *et al.*, 2012], and multi-instance learning [Yang *et al.*, 2009]; however, MIML in multi-modal setting has not been investigated before.

Specifically for image annotation, the most successful methods are based on propagating labels from nearest neighbors in the training data [Lavrenko *et al.*, 2003; Guillaumin *et al.*, 2009] or combining multiple binary classifiers from multiple modalities [Nowak *et al.*, 2011]. The methods, however, degenerate the task of image annotation from a MIML problem into a single-instance multi-label (SIML) problem; thus, it is hard to obtain region annotation. Moreover, although the propagation approaches are simple, the testing (annotation) time increases linearly with the size of the training dataset.

From the view of multi-label learning, our method is a kind of high-order dependency approach [Tsoumakas *et al.*, 2010; Zhang and Zhang, 2010], where the assignment of one label is influenced by the assignment of subgroups of labels. From the view of multi-modal learning [Atrey *et al.*, 2010], our method is a kind of hybrid fusion approach, where the fusion of multi-modalities is made in both decision level (labels) and feature level (visual/textual). Furthermore, our model is able to integrate contextual information into fusion process, where the context is defined in terms of topics and prior knowledge.

3 The Proposed Approach

3.1 The Generative Model

This section formalizes the problem of multi-modal image annotation and retrieval in the MIML framework. Let $\mathcal{Y} = \{y_1, y_2, \ldots, y_L\}$ denote a set of L labels, and $\mathcal{T} = \{t_1, t_2, \ldots, t_T\}$ denote a set of T user tags. Let $\mathcal{D} = \{([X_1, T_1], Y_1), \ldots, ([X_N, T_N], Y_N\}$ denote a training dataset of N examples where $X_n = \{\mathbf{x}_{n1}, \mathbf{x}_{n2}, \ldots, \mathbf{x}_{nM_n}\}$ is called a bag of M_n instances, $T_n = \{t_{n1}, t_{n2}, \ldots, t_{nG_n}\}$ is a set of G_n user tags, and $Y_n = \{y_{n1}, y_{n2}, \ldots, y_{nL_n}\}$ is a set of L_n labels from $\mathcal{Y}$. The goal is to generate a learner to annotate a new image (and its regions) based on its instances X and user tags T (if available).

Following [Zhou and Zhang, 2007], we build a set of prototypes $\mathcal{C} = \{c_1, c_2, ..., c_C\}$ by clustering in visual feature space. Like [Yang *et al.*, 2009] and without loss of generality, we assume that each instance is represented by a bag of prototypes. In other words, $\mathbf{x}_i$ is a vector of size C where $\mathbf{x}_{i,c}$ counts the number that prototype c appears in $\mathbf{x}_i$.

The generative model of M3LDA is illustrated in Figure 1 and Algorithm 1. During training, for each image X_n, we set ξ_{ni} ($\forall i \in [1 \ldots L]$) to 0 to constrain the topics of instances/tags to the labels in Y_n. During testing, ξ_i can be either a non-zero constant, or a confident value of a light binary classifier for y_i; while η controls the influence of topics and priors on annotation. K is a predefined number of topics, where one topic $\phi_k^{(y)}$ groups correlated labels. Each image has a label distribution (θ') that is determined by three components: the label distribution over instances $\phi^{(c)}$, the label distribution over tags $\phi^{(t)}$, and the topic distribution (θ). By allowing the topic distribution θ affect the label distribution θ', the correlations of labels can be exploited for annotation. Here, $\mathbf{z}$, $\mathbf{v}$, and $\mathbf{g}$ are variables that assign topics to labels, labels to instances, and labels to tags; α, $\beta^{(y)}$, $\beta^{(c)}$ and $\beta^{(t)}$ are hyper-parameters for θ, $\phi^{(y)}$, $\phi^{(c)}$ and $\phi^{(t)}$, respectively.

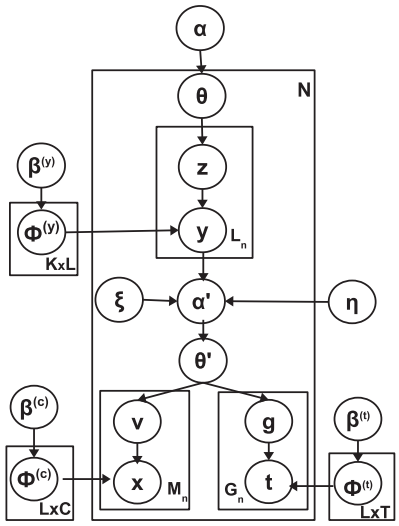

Figure 1: The M3LDA model

M3LDA assigns multiple labels (via $\mathbf{v}, \mathbf{g}$) to tags and instances. During the assignment process, the correlations between labels are taken into account owing to the influence of "label topics" (θ). Here, the topics of visual instances and tags (θ') are restricted by the set of labels, thus we define an one-to-one correspondence between the labels and the topics of visual instances/tags, allowing us to perform region annotation (assigning labels to instances). The instance-topic-label and tag-topic-label components, therefore, are referred to as the instance-label and the tag-label, respectively; also θ' is called the label distribution to avoid the confusion with the topic distributions of labels (θ). M3LDA is the multi-modal setting, where the label distribution (θ') is decided by both visual content and tags. Therefore, the commonness from the two modalities is gathered to obtain "more confident" labels. Here, the label set plays as a bridge between the visual content and the rich, dynamic set of human vocabulary (tags). Although M3LDA is designed for image annotation, it is general enough to be applied to other multi-modal tasks.

Note that M3LDA allows the incorporation of the label priors to each testing image via ξ_i. As a result, our model is easier to be tuned in practice. For example, one can incorporate the results from "face recognition" systems to trigger the labels such as "persons", "female", "male". Alternatively, we can first perform scene classification using global features,

Algorithm 1: Generative Process for M3LDA

for *each image* X_n **do**
- Sample a topic distribution of labels $\theta \sim$ Dirichlet(α); θ is a K-dimensional Dirichlet distribution parameterized by α.
- **for** *each label* y_{ni} *in* Y_n *of image* X_n **do**
  - Sample a topic assignment $z_i \sim$ Multinominal(θ).
  - Sample a label from $p(y|z_i, \phi^{(y)}) =$ Multinominal($\phi^{(y)}_{z_i}$) from the topic z_i.
- Compute the label priors for X_n: α'_n=$\{\eta \times N_1^{(Ln)}/L_n + \xi_{n1}, \ldots, \eta \times N_L^{(Ln)}/L_n + \xi_{nL}\}$ where $N_i^{(Ln)}$ is the number of y_i in Y_n; ξ_{ni}=0, η>0 during training; and $\xi_{ni} > 0, \eta$>0 during testing.
- Sample a label distribution $\theta' \sim$ Dirichlet($.|\alpha'_n$); θ' is a L-dimensional Dirichlet distribution parameterized by α'_n.
- **for** *each instance* $\mathbf{x_{ni}}$ *of* X_n **do**
  - Sample a label assignment $v_i \sim$ Multinominal(θ').
  - Sample an instance from $p(\mathbf{x}|v_i, \phi^{(c)})$=$\prod_{c=1}^{C} \left(\phi^{(c)}_{c,v_i}\right)^{\mathbf{x_c}}$; $\phi^{(c)}_{.,v_i}$ is a C-dimensional Multinominal for the label v_i.
- **for** *each tag* t_{ni} *in* T_n *of image* X_n **do**
  - Sample a label assignment $g_i \sim$ Multinomial(θ').
  - Sample a tag from $p(t|g_i, \phi^{(t)}) =$ Multinomial($\phi^{(t)}_{g_i}$) from the label g_i.

and associate labels related to the scene with higher values of ξ_i. In this work, we follow two ways for setting ξ: 1) ξ_i is set to a constant value (0.1); and 2) ξ_i is proportional to output values of "light" binary classifiers.

Unlike previous LDA-based multi-modal methods [Blei and Jordan, 2003; Jain *et al.*, 2007; Wan *et al.*, 2010; Putthividhy *et al.*, 2010], M3LDA provides a full multi-instance, multi-label multi-modal solution. Moreover, it is easy to extend M3LDA to include more modalities, while this is not a trivial task in the previous models.

When tags are not available, the model still works by ignoring the tag-label part. The version of M3LDA in one modality is indeed a MIML model (MIML-LDA), which is quite similar to Dependence-LDA [Rubin *et al.*, 2012]. The difference between MIML-LDA and Dependence-LDA lies in the fact that MIML-LDA follows MIML setting that allows $\mathbf{x}$ to be a bag of prototypes. Additionally, prior knowledge is incorporated to MIML-LDA for individual example instead of a common prior for the whole set. Moreover, we consider the problem in image annotation instead of text classification.

3.2 Training Process

For training and testing with M3LDA, we follow Gibbs Sampling approach because it is intuitive to interpret the interactions in M3LDA. However, it is also not difficult to derive a variational method for M3LDA by relaxing the dependency on α', θ' and obtaining three LDA models, for which the variational method in [Blei *et al.*, 2003] can be applied.

During training, the observed components (α' and Y) block the flow from the label-topic part (the higher part) to the instance/tag parts (the lower parts) in Figure 1. Consequently, the estimation of the label-topic part and instance/tag parts can be conducted independently. Similar to LDA [Blei *et al.*, 2003], we can design a collapsed Gibbs Sampling method for effective model inference. As the label-topic part is estimated independently, it is exactly like a LDA model [Griffiths and Steyvers, 2004]. In the following, we show the sampling equations for label-instance and label-tag parts. From the generative model of M3LDA, we have:

$$p(X, \mathbf{t}, \mathbf{v}, \mathbf{g}) = p(\mathbf{v}, \mathbf{g})p(X|\mathbf{v})p(\mathbf{t}|\mathbf{g}). \quad (1)$$

For collapsed Gibbs Sampling, we first integrate out (collapse) distributions $\theta', \phi^{(c)}, \phi^{(t)}$ to compute $p(X, \mathbf{t}, \mathbf{v}, \mathbf{g})$. Let $N_{yn}^{(CLn)}$ (resp. $N_{yn}^{(TLn)}$) denote the number of times that label y is assigned to visual (resp. textual) modality of the n-th example in training dataset, and $N_{cy}^{(CL)}$ (resp. $N_{ty}^{(TL)}$) denote the number of times that prototype c (resp. tag t) is assigned to label y; we then attain the Gibbs sampling equation for updating label assignment for $\mathbf{x}_{ni}$ as follows:

$$\begin{aligned} & P(v_{ni} = y|\mathbf{x}_{ni}, X_{-i}, \mathbf{v}_{-i}, \mathbf{t}, \mathbf{g}, \alpha'_n, \beta^{(c)}) \\ \propto & \frac{N_{yn,-i}^{(CLn)} + N_{yn}^{(TLn)} + \alpha'_{ny}}{N_{.n,-i}^{(CLn)} + N_{.n}^{(TLn)} + \sum_{y'} \alpha'_{ny'}} \\ \times & \frac{\prod_{c=1}^{C} \prod_{l=1}^{\mathbf{x}_{nic}} (N_{cy,-i}^{(CL)} + \mathbf{x}_{nic} - l + \beta^{(c)})}{\prod_{m=1}^{N_i^{(n)}} (N_{.y,-i}^{(CL)} + N_i^{(n)} - m + C\beta^{(c)})}. \end{aligned} \quad (2)$$

where X_{-i} is obtained from X by excluding the instance $\mathbf{x}_{ni}$ which contains $N_i^{(n)}$ prototypes in total, and $\mathbf{x}_{nic}$ prototype c. When we consider only one modality and each instance $\mathbf{x}_{ni}$ contains only one prototype, Eq.(2) has the same form as standard LDA. Intuitively, the first term of Eq.(2) measures the rate of assigning label y to the whole image and tags, and the second term corresponds to how likely the current instance $\mathbf{x}_{ni}$ is associated with y. This equation also shows that the tags and priors in α'_n play important roles in the assignments of labels to instances (region annotation). Similarly, we derive the label assignment for tags as follows:

$$\begin{aligned} & P(g_{ni} = y|t_{ni} = t, \mathbf{t}_{-\mathbf{i}}, \mathbf{g}_{-i}, X, \mathbf{v}, \alpha'_n, \beta^{(t)}) \\ \propto & \frac{N_{yn,-i}^{(TLn)} + N_{yn}^{(CLn)} + \alpha'_{ny}}{N_{.n,-i}^{(TLn)} + N_{.n}^{(CLn)} + \sum \alpha'_{ny'}} \frac{N_{ty,-i}^{(TL)} + \beta^{(t)}}{N_{.y,-i}^{(TL)} + T\beta^{(t)}}. \end{aligned} \quad (3)$$

After the model is sufficiently burned in, we can obtain the posterior distribution of labels θ'_n, which can be used for annotating image X_n:

$$\theta'_{ny} = \frac{N_{yn}^{(TLn)} + N_{yn}^{(CLn)} + \alpha'_{ny}}{N_{.n}^{(TLn)} + N_{.n}^{(CLn)} + \sum_{y'} \alpha'_{ny'}}. \quad (4)$$

The updates of $\phi^{(c)}, \phi^{(t)}, \phi^{(y)}$ are like in Gibbs Sampling for LDA [Griffiths and Steyvers, 2004].

3.3 Testing Process

Given $\phi^{(c)}, \phi^{(t)}, \phi^{(y)}$ obtained from the training process, the fast inference method in [Rubin *et al.*, 2012] can be applied for a new image $X_{\tilde{n}}$. First, we derive the probability of assigning labels over instances as follows:

$$P(v_{\tilde{n}i} = y | \mathbf{x}_{\tilde{n}i}, X_{\tilde{n},-i}, \mathbf{v}_{\tilde{n},-i}, \mathbf{t}_{\tilde{n}}, \mathbf{g}_{\tilde{n}}, \alpha'_{\tilde{n}})$$
$$\propto \frac{N_{y\tilde{n},-i}^{(CL\tilde{n})} + N_{y\tilde{n}}^{(TL\tilde{n})} + \alpha'_{\tilde{n}y}}{N_{.\tilde{n},-i}^{(CL\tilde{n})} + N_{.\tilde{n}}^{(TL\tilde{n})} + \sum_{y'} \alpha'_{\tilde{n}y'}} \prod_{c=1}^{C} \left(\phi_{cy}^{(c)}\right)^{\mathbf{x}_{\tilde{n}ic}}. \quad (5)$$

Similarly, we can obtain the probability of label assignments for tags:

$$P(g_{\tilde{n}i} = y | t_{\tilde{n}i} = t, \mathbf{t}_{\tilde{n},-i}, \mathbf{g}_{\tilde{n},-i}, X_{\tilde{n}}, \mathbf{v}_{\tilde{n}}, \alpha'_{\tilde{n}})$$
$$\propto \frac{N_{y\tilde{n},-i}^{(TL\tilde{n})} + N_{y\tilde{n}}^{(CL\tilde{n})} + \alpha'_{\tilde{n}y}}{N_{.\tilde{n},-i}^{(TL\tilde{n})} + N_{.\tilde{n}}^{(CL\tilde{n})} + \sum_{y'} \alpha'_{\tilde{n}y'}} \times \phi_{ty}^{(t)}. \quad (6)$$

The fast inference means that we directly obtain $Y_{\tilde{n}}$ by joining the sets $\mathbf{g}_{\tilde{n}}$ and $\mathbf{v}_{\tilde{n}}$. Once we have $Y_{\tilde{n}}$, the estimates of topic assignments for sampled labels and topic distribution $\theta_{\tilde{n}}$ for $X_{\tilde{n}}$ are performed like in LDA. Given $\phi^{(y)}$ and $\theta_{\tilde{n}}$, we can approximate $\alpha'_{\tilde{n}} = \eta(\phi^{(y)} \times \theta_{\tilde{n}}) + \xi$ [Rubin *et al.*, 2012].

The testing process is summarized in a *5-step* procedure: **1)** Update $\mathbf{v}_{\tilde{n}}$ using Eq.(5); **2)** Update $\mathbf{g}_{\tilde{n}}$ using Eq.(6); **3)** Update $Y_{\tilde{n}}$ by directly joining $\mathbf{v}_{\tilde{n}}$ and $\mathbf{g}_{\tilde{n}}$; **4)** Sample topic assignment $\mathbf{z}_{\tilde{n}}$ based on $Y_{\tilde{n}}$ and update topic distribution $\theta_{\tilde{n}}$ like in LDA; and **5)** Update $\alpha'_{\tilde{n}} = \eta(\phi^{(y)} \times \theta_{\tilde{n}}) + \xi$. The sampling is repeated until it is burned in.

After inference, we can obtain label distribution $\theta'_{\tilde{n}}$ from Eq.(4), which is sorted to obtain labels with highest values for annotation. Additionally, the label assignment $\mathbf{v}_{\tilde{n}}$ can be used for region annotation. Eqs.(4-6) show that our fusion is conducted both in the feature level (Eqs.(5-6)) where two modalities can influence each others, and decision level (Eq.(4)). Moreover, contexts can be exploited in the fusion process in terms of topics (θ) and the priors (ξ).

It is worth mentioning that when sampling for one modality, we keep the other modality fixed. Thus, it is not difficult to extend our model to other modalities by including the label assignment counting in the new modality similar to $N^{(CLn)}$ and $N^{(TLn)}$ in Eqs.(2-6).

3.4 Implementation Details

In the model, we assume that two modalities are equivalent. However, in practice, we may like to weight some modality more than the other. In M3LDA, we can achieve this objective via repeatedly sampling instances/tags where more important modality will be sampled more.

On the other hand, Eqs.(2-6) also show an easy way to incorporate weights for different modalities (tags, instances) by setting different weights for instance-label assignment counting $N^{(CLn)}$ and tag-label assignment counting $N^{(TLn)}$. In the implementation, we follow this approach. First, we normalize terms $N^{(CLn)}$, $N^{(TLn)}$ with the length of the corresponding modalities M_n and G_n, respectively. We then introduce the weighting mixture λ ($\lambda \in [0, 1]$). More specifically, $N^{(CLn)} + N^{(TLn)}$ in Eqs.(2-6) will be replaced with $\left[(1-\lambda)N^{(CLn)}/M_n + \lambda N^{(TLn)}/G_n\right](M_n + G_n)$. Thus, in addition to the parameters of standard LDA models, we have three parameters to incorporate these components into M3LDA, i.e., η, ξ, and λ. We will study the influence of these parameters in Section 4.

4 Experiments

We evaluate our proposed method on Corel5K and ImageCLEF using the example-pivot and label-pivot evaluation protocols [Rubin *et al.*, 2012]. For example-pivot protocol, we calculate AP (average precision) over the ranking list of labels for each image, then average over all images. For label-pivot protocol, like [Guillaumin *et al.*, 2009] we obtain a fixed annotation length of N labels per image, where N is 5 or 10 depending on the dataset, then calculate Precision (P) and Recall (R). We also calculate mAP (mean Average Precision) by averaging APs over the whole label set.

4.1 Corel5K

The Corel5k benchmark [Duygulu *et al.*, 2002] contains 5,000 images that are pre-divided into a training set of 4,500 images and a test set of 500 images. Each image is annotated with 1 to 5 labels, and one image has 3.22 labels on average. We test on 260 labels which appear on both training and testing datasets. We use blob features as described in [Duygulu *et al.*, 2002]. As Corel5K does not contain tag information, M3LDA becomes to MIML-LDA.

We compare M3LDA with two MIML models: RankLoss [Briggs *et al.*, 2012], DBA [Yang *et al.*, 2009], and two annotation models that allow region annotation: TM [Duygulu *et al.*, 2002] and Corr-LDA [Blei *et al.*, 2003]. To reduce computational complexity and obtain a fair comparison, we modify Corr-LDA so that it works with bags of prototypes like DBA and M3LDA. More specifically, the Gaussian distribution in Corr-LDA is replaced by a multinomial distribution. We set $K = 50$ for both Corr-LDA and M3LDA; $\eta = 20$ for M3LDA. Beside M3LDA in which we set $\xi_i = 0.1, \forall i$, we also obtain M3LDA with priors by training binary classifiers using LIBSVM [Chang and Lin, 2011] with Gist features [Torralba *et al.*, 2010]. For a test image I, the probability estimates of SVM ($h(I, y_i)$) is incorporated into M3LDA by setting $\xi_i = \eta \times h(I, y_i) + 0.1$. M3LDA with this setting of ξ is called M3LDA+. Also, the results of SVM-Gist are evaluated only for a comparison with M3LDA+. A direct comparison of SVM-Gist with other models, however, is not fair as they work in different feature space. Note that SVM-Gist is not able to perform region annotation. Considering that the number of labels per image is less than 5,

Table 1: Performance comparison on Corel5K, where the best performances are highlighted with boldface.

Method	Label Pivot			Example Pivot
	MAP	P	R	AP
Corr-LDA	0.067	0.058	0.085	**0.294**
DBA	0.073	0.066	0.095	0.150
TM	N/A	0.040	0.060	N/A
RankLoss	0.050	0.031	0.049	0.260
M3LDA	**0.084**	**0.079**	**0.117**	0.284
SVM-Gist	0.135	**0.170**	0.150	0.403
M3LDA+	**0.148**	0.134	**0.179**	**0.410**

Table 2: Performance comparison on ImageCLEF. Here, Corr, M3L-v, M3L-t and M3L are short for Corr-LDA, M3LDA-visual, M3LDA-tag and M3LDA, respectively.

	Label Pivot			Ex. Pivot
	MAP	P	R	AP
Corr	.141±.002	.155±.009	.189±.003	**.306±.005**
DBA	.120±.003	.092±.011	.174±.005	.152±.007
SVM	.125±.003	.142±.020	.057±.004	.192±.005
M3L-v	.143±.003	.161±.016	.224±.021	.241±.001
M3L-t	.150±.005	.178±.008	.209±.006	.248±.007
M3L	**.185±.007**	**.197±.008**	**.276±.016**	.245±.009

we perform label-pivot evaluation with $N = 5$ like [Guillaumin *et al.*, 2009]. The results are summarized in Table 1, where the results of TM are reported from previous studies using the same feature space [Duygulu *et al.*, 2002; Lavrenko *et al.*, 2003].

Table 1 reveals that TM and RankLoss obtain lower P and R compared with the other methods. Although RankLoss obtains high AP with the example-pivot protocol, it does not work well with label-pivot protocol. This suggests that RankLoss suffers from label imbalance, i.e, this method favors a small number of frequent labels such as "sky", "water". Corr-LDA is able to obtain better results in example-pivot protocol than DBA possibly owing to the fact that Corr-LDA considers label relationship but DBA does not. On the other hand, DBA is better than Corr-LDA in label-pivot protocol in both P and R, possibly owing to the fact that DBA directly casts labels as topics of instances. M3LDA is comparable to Corr-LDA in the example-pivot protocol, and outperforms Corr-LDA and DBA in label-pivot protocol. M3LDA+ obtains significant improvement partly due to the fact that M3LDA+ can gather information from Gist prior and the visual instances.

4.2 ImageCLEF

The ImageCLEF 2011 challenge [Müller *et al.*, 2010; Nowak *et al.*, 2011] contains 18,000 Flickr images of size 300×500 with user tags and 99 visual concepts (labels). For scene-related annotation, we exclude the emotional labels (e.g. "cute"), and the technical terms (e.g. "partly_blur") from the label set to obtain a set of 78 labels. The obtained dataset has around 10 labels per image, thus we evaluate label-pivot protocol with the fixed annotation length of $N = 10$.

For tag modality, we select non-numerical tags that appear more than 20 times, resulting in a set of 806 tags. For visual modality, we apply OpponentSIFT with dense sampling using ColorDescriptor toolbox [van de Sande *et al.*, 2010]. We then build a codeword of 1,000 entries by clustering feature vectors. Each bag of instances is then extracted by sliding a window of 20×20 along an image, thus an instance corresponds to a 20×20 patch. Note that we have tried to employ image segmentation, but the annotation results are not better than our patch-based presentation. This might be due to the fact that overly relying on non-perfect segmentation is less informative than using a lot of small patches.

We compare our method with Corr-LDA, DBA, and SVM (default parameters in LibSVM). Here, by accumulating all the instances of each bag, we can degenerate the problem of multi-instance learning to single-instance learning and apply SVM in one-vs-all like [Nowak *et al.*, 2011]. Note again that we are not able to obtain region annotation with SVM. We also evaluate M3LDA-visual and M3LDA-tag; the degenerated versions of M3LDA where only visual or textual modality is used, respectively. For parameter settings, K is set to 100 for both Corr-LDA and M3LDA. For M3LDA models, η is set to 250 considering that the number of instances and tags for each image is larger than 400. The mixture weight λ for combining modalities in M3LDA is set to 0.8. The influence of these parameters on testing will be studied in Section 4.3. The experiments are conducted on 30 random selections of 2,000 images from ImageCLEF, in which 1,000 images are used for training and 1,000 images are used for testing. The average and standard deviations are also reported in Table 2.

Table 2 shows that DBA and SVM perform worse than Corr-LDA and M3LDA models in all measures. This might be due to that the strong dependency among labels (female, male, etc.) in ImageCLEF makes DBA, SVM that considers labels independently, perform inferior. Compared with Corr-LDA, M3LDA approaches significantly improve the performance in mAP, P, and R (two tailed t-tests at 95% significance level) but perform worse in AP. This result is consistent with what we observed in Corel5K. It might imply that Corr-LDA is able to capture label dependency better. It is, however, more difficult to extend Corr-LDA because we may need to retrain the whole model if we want to refine the topic model or to add new modality. Also, it is possible to replace the label-topic part of M3LDA with more advanced topic models for better label correlations while fixing the trained visual-label and tag-label parts. This is because the label-topic part is independent with the rest of M3LDA during training.

Among M3LDA approaches, M3LDA-tag performs better than M3LDA-visual in mAP, P, and AP but worse than M3LDA-visual in R. When we combine two modalities, M3LDA obtains significant improvements in all measures except for AP, which is comparable to M3LDA-tag.

4.3 Parameter Influence

Figure 2 shows performance of M3LDA on ImageCLEF when we change η and λ. The experiments for each setting of η and λ are conducted 30 times; the averages and standard deviations are shown in the figure. Here, $\eta = 0$ means topics play no role in annotation. It is observable that without topics, the performance is always worse than when we consider

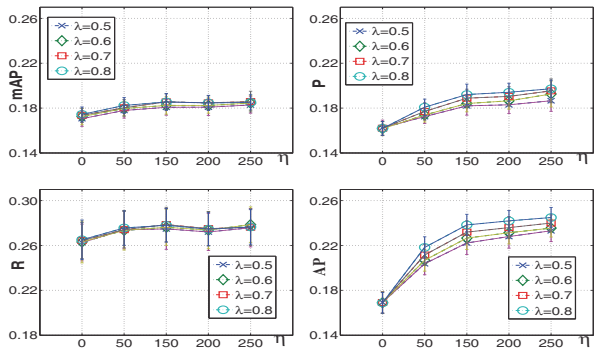

Figure 2: Parameter Influence on M3LDA.

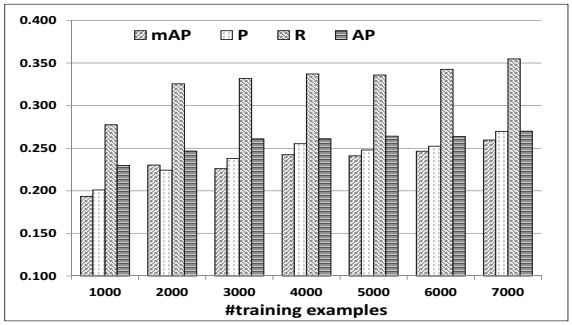

Figure 3: M3LDA with different amount of training data.

topics ($\eta > 0$). Moreover, when η is larger, the performance is getting better. There is, however, not much difference in performance when η is large enough ($\eta \in \{200, 250\}$).

Figure 2 also indicates that our method is quite insensitive to the value of λ. This is reasonable because the APs of the M3LDA-tag and M3LDA-visual in Table 2 are comparable. In practice, if the performance of the two modalities are very different, the mixture can be tuned on a validation dataset.

We study the impact of training set size on the performance by fixing 1,000 images as testing and varying the training size from 1,000 to 7,000. For each image during training, we subsample a set of 200 instances. Figure 3 reveals that when increasing the number of training examples, all the evaluation measures increase. For 1,000 training examples, each Gibbs sampling iteration (of M3LDA written in Java) takes 9.4 seconds on a computer of 3.30GHz, 4GB memory. This duration rises linearly on the number of training examples. For testing, the annotation time is around 1(s) per image.

4.4 Annotation Results

Figure 4 shows examples of annotation results on ImageCLEF. Here, M3LDA-visual and M3LDA-tag with $\eta = 0$ are considered to eliminate the impact of topics, thus allow us to study the performance of individual modalities. In the first image, because tags are uninformative for labeling, M3LDA-tag gives incorrect annotation. M3LDA-visual, on the other hand, highly supports the "mountain view" topic. When we combine two modalities with ($\eta > 0$), the labels relevant to "mountain view" get higher output values. In the second image, both modalities highly contribute to a consistent topic (street view); and M3LDA is able to predict some labels (e.g. "outdoor", "citylife") that even do not appear on top ranks of both M3LDA-tag and M3LDA-visual, possibly owing to the information reinforced between the two modalities.

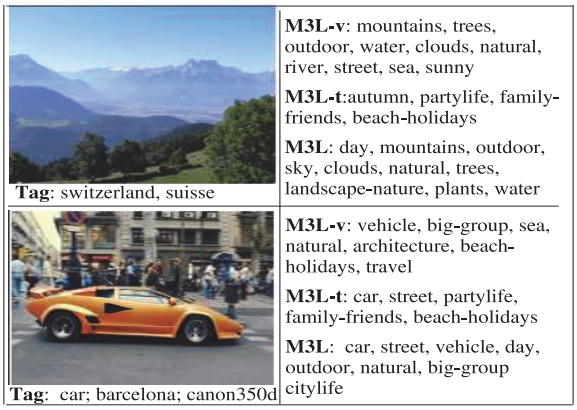

Figure 4: Example Annotation Results on ImageCLEF, where M3L-v, M3L-t and M3L are short for M3LDA-visual; M3LDA-tag and M3LDA.

Figure 5: Region Annotation with M3LDA+ on Corel5K.

5 Conclusion

In this paper, we propose the M3LDA approach which leverages the advantages of MIML learning and probabilistic generative model to exploit multi-modal information for image annotation. Annotation label assignment is obtained by maintaining the consistency between the topics by the visual information and tag information. Experiments show the advantages of M3LDA over state-of-the-art algorithms. It is worth noting that, in addition to the annotations for whole images, M3LDA is able to give annotations to image regions.

Although M3LDA is designed for image annotation, it is possible to be applied to other multi-modal tasks. As for the exploitation of multi-modal information, multi-view learning [Blum and Mitchell, 1998] provides a promising way. It is interesting to combine M3LDA with multi-view learning techniques in the future.

References

[Atrey *et al.*, 2010] P. K. Atrey, M. A. Hossain, A. El-Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 16(6):345–379, 2010.

[Blei and Jordan, 2003] D. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134, Toronto, Canada, 2003.

[Blei *et al.*, 2003] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Blum and Mitchell, 1998] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, WI, 1998.

[Briggs *et al.*, 2012] F. Briggs, F.Z. Xiaoli, and R. Raich. Rank-loss support instance machines for miml instance annotation. In *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 534–542, Beijing, China, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for SVMs. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Duygulu *et al.*, 2002] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European Conference on Computer Vision*, pages 97–112, London, UK, 2002.

[Griffiths and Steyvers, 2004] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[Guillaumin *et al.*, 2009] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the 12th IEEE International Conference on Computer Vision*, pages 309–316, Kyoto, Japan, 2009.

[Jain *et al.*, 2007] V. Jain, E. L.-Miller, and A. McCallum. People-LDA: Anchoring topics to people using face recognition. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, October 2007.

[Lavrenko *et al.*, 2003] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems 16*, pages 553–560. MIT Press, 2003.

[Müller *et al.*, 2010] H. Müller, P. Clough, T. Deselaers, and B. Caputo. *ImageCLEF: Experimental Evaluation of Visual Information Retrieval*. Springer Publishing Company, Incorporated, Berlin, German, 2010.

[Nowak *et al.*, 2011] S. Nowak, K. Nagel, and J. Liebetrau. The CLEF 2011 photo annotation and concept-based retrieval tasks. In *Proccedings of the 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, Amsterdam, The Netherlands, 2011.

[Putthividhy *et al.*, 2010] D. Putthividhy, H.T. Attias, and S.S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *Proceedings of the 23rd IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3415, San Francisco, CA, 2010.

[Rosen-Zvi *et al.*, 2004] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proccedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, Banff, Canada, 2004.

[Rubin *et al.*, 2012] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012.

[Torralba *et al.*, 2010] A. Torralba, K. P. Murphy, and W. T. Freeman. Using the forest to see the trees: Exploiting context for visual object detection and localization. *Communications of the ACM*, 53(3):107–114, 2010.

[Tsoumakas *et al.*, 2010] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer, 2010.

[van de Sande *et al.*, 2010] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.

[Wan *et al.*, 2010] K-W Wan, A-H Tan, J-H Lim, and L-T Chia. Faceted topic retrieval of news video using joint topic modeling of visual features and speech transcripts. In *Proceedings of the 2010 IEEE International Conference Multimedia and Expo*, pages 843–848, Singapore, 2010.

[Yang *et al.*, 2009] S.-H. Yang, H. Zha, and B.-G. Hu. Dirichlet-Bernoulli alignment: A generative model for multi-class multi-label multi-instance corpora. In *Advances in Neural Information Processing Systems 22*, pages 2143–2150. MIT Press, 2009.

[Zhang and Zhang, 2010] M-L Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 999–1008, Washington, DC, 2010.

[Zhou and Zhang, 2007] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. In *Advances in Neural Information Processing Systems 19*, pages 1609–1616. MIT Press, 2007.

[Zhou *et al.*, 2012] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, and Y.-F. Li. Multi-instance multi-label learning. *Artificial Intelligence*, 176(1):2291–2320, 2012.