Joint Multi-Label Multi-Instance Learning for Image Classification

*Zheng-Jun Zha[†] Xian-Sheng Hua[‡] Tao Mei[‡] Jingdong Wang[‡] Guo-Jun Qi[†] Zengfu Wang[†]
[†]Department of Automation [‡]Internet Media Group
[†] University of Science and Technology of China [‡]Microsoft Research Asia

Abstract

In real world, an image is usually associated with multiple labels which are characterized by different regions in the image. Thus image classification is naturally posed as both a multi-label learning and multi-instance learning problem. Different from existing research which has considered these two problems separately, we propose an integrated multi-label multi-instance learning (MLMIL) approach based on hidden conditional random fields (HCRFs), which simultaneously captures both the connections between semantic labels and regions, and the correlations among the labels in a single formulation. We apply this MLMIL framework to image classification and report superior performance compared to key existing approaches over the MSR Cambridge (MSRC) and Corel data sets.

1. Introduction

With the proliferation of digital photography, image understanding becomes increasingly important. Image semantic understanding is typically formulated as a *multi-class* or *multi-label* learning problem. In multi-class setting [18], each image will be categorized into one and only one of a set of predefined categories. In other words, only one label will be assigned on each image in this setting. In multi-label setting [1] [13] [16] [9], which is more challenging but much closer to real world applications, each image will be assigned with one or multiple labels from a predefined label set, such as "sky," "mountain," and "water," illustrated in Figure 1. This paper is about *multi-label learning* (MLL) for image classification.

Multi-label classification can be solved by transferring it into a set of independent two-class (binary) classification problems [1], while more sophisticated solutions also leverage the correlations of the labels (either after modeling each individual label [9] or modeling the labels and the correlations among labels simultaneously [13] [16]). However, all these approaches regard an image as one indiscrete entity and neglect the fact that mostly each individual label of the image is actually more closely related to one or more regions instead of the entire image. In other words, the multiple semantic meanings (labels) of an image arise from different components (regions) in it. As illustrated in Figure 1, the three labels "sky," "mountain," and "water" are characterized by three different regions, respectively, rather than the entire image.

Modeling the relations between labels and regions (instead of the entire image) will reduce the noises in the corresponding feature space, and hence the learned semantic models will be more accurate. To address this issue, many researchers formulate image classification as a *multi-instance learning* (MIL) task. In MIL, an image is viewed as a *bag*, which contains a number of *instances* corresponding to the regions in the image [3] [20] [7] [19]. If any of these instances is related to a label, the image will be associated with the label. However, these methods mainly focus on single-label scenario and multi-label problems need to be implemented label-by-label independently. That is to say, the label correlations are not taken into account in these MIL-based classification methods. However, researchers have proved that exploiting label correlations will significantly improve the performance of image classification [13] [16].

To address the above issues of existing MLL and MIL approaches, in this paper, we formulate image classification as a joint *multi-label multi-instance learning* (MLMIL) problem. Different from existing research which has not simultaneously considered the multi-label and multi-instance problems, we model them in an integrated framework by capturing both the connections between semantic labels and regions, as well as the correlations among the labels in a single formulation. Moreover, the proposed framework is also able to capture other dependencies among the regions, such as the spatial relations. Figure 1 illustrates the comparison of MLL, MIL and MLMIL in terms of the modeled relations.

There is an initial attempt to address this problem [22]. However, as to be detailed, in that work this problem is

[*] This work was performed when Zheng-Jun Zha and Guo-Jun Qi were visiting Microsoft Research Asia as research interns.

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

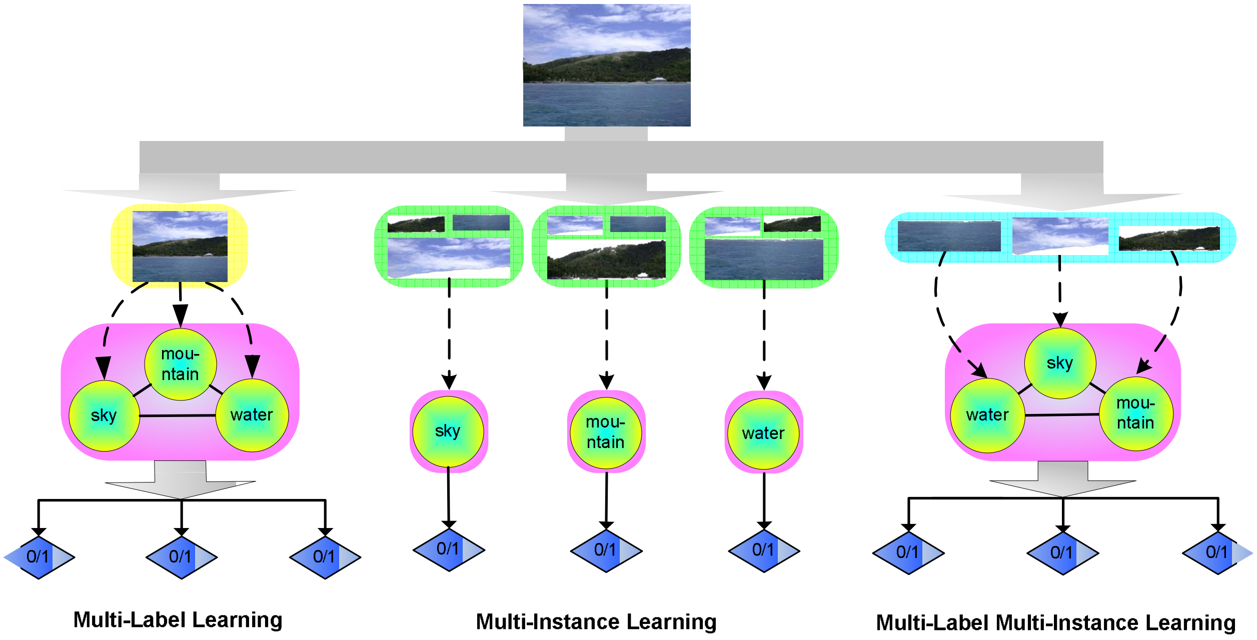

Figure 1. Comparison of three paradigms of image classification approaches. From leftmost to the rightmost, they are *multi-label learning* (MLL), *multi-instance learning* (MIL), and the proposed joint *multi-label multi-instance learning* (MLMIL) framework. MLL captures the correlations of the labels, while MIL models the connection between labels and regions. The proposed MLMIL framework models both relations simultaneously.

transferred into typical multi-instance learning or multi-label learning problem, in which label-label correlation is actually not modeled. Therefore, the connections between instances and labels, and the correlations among labels are not sufficiently leveraged to improve the classification performance.

To summarize, the proposed joint multi-label multi-instance learning framework has the following key advantages compared to the existing methods:

- Compared to the MLL framework, the MLMIL method captures the intrinsic causation of each individual label and directly models the latent semantic meaning of regions. .
- In contrast with the MIL methods which model individual labels independently, MLMIL simultaneously models both the individual labels and their interactions.
- Moreover, the MLMIL framework is flexible to capture the various dependencies among the regions.

The rest of this paper is organized as follows. We review related work in Section 2. Section 3 gives detail description of the proposed joint multi-label multi-instance learning framework. Experimental results on both *MSR Cambridge* (MSRC) and Corel data sets are reported in Section 4, followed by concluding remarks in Section 5.

2. Related Work

Related research on image classification can be summarized along three paradigms: *multi-label learning* (MLL), *multi-instance learning* (MIL), and *multi-label multi-instance learning* (MLMIL).

2.1. Multi-Label Learning

An image is typically described by multiple semantic labels (Figure 1); therefore real world image classification is generally formulated as a multi-label learning problem. The typical solution of multi-label classification is to translate the multi-label learning task into a set of single-label classification problems. For example, Boutell *et al.* [1] solved the multi-label scene classification problem by building individual classifier for each label. The labels of a new sample are determined by the outputs of these individual classifiers.

The above solution treats the labels in isolation and ignores the correlations among the labels. However, these labels are usually interacting with each other naturally. For example, “mountain” and “sky” tend to appear simultaneously, while “sky” typically does not appear with “indoor”. To exploit these correlations, some researchers have proposed fusion-based methods [9]. Godbole *et al.* [9] proposed to leverage the correlations by adding a contextual fusion step based on the outputs of the individual classifiers.

More sophisticated MLL approaches model labels and correlations between labels simultaneously [13] [16]. Kang *et al.* [13] developed a *Correlation Label Propagation* (CLP) approach to explicitly capture the interactions between labels. Rather than treating labels independently, CLP simultaneously co-propagates multiple labels from training examples to testing examples. More recently, Qi *et al.* [16] proposed a unified *Correlative Multi-Label* (CML) *Support Vector Machine* (SVM) to simultaneously classify labels and model their correlations in a new feature space which encodes both the label models and their interactions together.

The first-paradigm approaches treat an image as an indiscrete unit and do not capture the semantic meanings of the regions which actually contribute to the corresponding labels. In addition, these approaches cannot model the dependencies among the regions which are also helpful for improving the classification performance.

2.2. Multi-Instance Learning

Multi-instance learning (MIL) based image classification takes the relations between labels and regions into account [7] [21] [3] [20]. In this paradigm, an image is regarded as a bag consisting of multiple instances (i.e., regions). MIL allows of only labeling images at the image level, instead of labeling at region level, when building classifiers. For a specific semantic label, a bag is labeled positive if at least one instance has the corresponding semantic meaning; otherwise, it is negative. Thus an essential question of MIL is: *which instances indeed contribute to the semantic meaning of the bag-level labels?* Different perspectives on this question lead to different MIL approaches. For example, Gartner *et al*. [7] assumed that all the instances in the bag are related to the labels. In EM-DD [21], only one instance per bag is regarded determining the bag label. This instance is estimated using *Expectation Maximization* (EM) [4] style approach. Chen *et al.* proposed DD-SVM [3] and *Multiple-Instance Learning via Embedded Instance Selection* (MILES) [20]. DD-SVM [3] assumes the semantic labels are related to a set of prototypes, which are selected from the local maximum of *Diverse Density* (DD) function. In MILES [20], bags are embedded into a feature space defined by instances. The semantically representative instances are determined during learning the bag classifier.

Although these MIL approaches have been proved effective, they limit in dealing with single label problems, though it can also be applied in multi-label problems (by treating it as a set of independent single-label problems). That is to say, label correlations are not taken into account in these MIL-based methods.

2.3. Multi-Label Multi-Instance Learning

As aforementioned, an image can be described by multiple semantic labels and these labels are often highly related to respective regions rather than the entire image.

Therefore, a more rational and natural strategy is to model image classification as a *multi-label multi-instance learning* (MLMIL) problem. To the best of our knowledge, this problem has seldom been explored. An initial attempt was made by Zhou *et al.* [22]. They proposed MIML-BOOST and MIML-SVM to solve the multi-label multi-instance problem. In MIML-BOOST, they translated the MLMIL task into typical MIL problem. Specifically, each MLMI sample was transformed into a set of MI samples each of which corresponds to a single label. MIBOOSTING [19] was then adopted to solve the MIL problem by further translating it into a set of typical supervised learning tasks (i.e., single label single instance problem). In MIML-SVM, the MLMIL task was transformed into typical MLL problem. Specifically, K medoids were generated firstly. Each MLMI sample was then mapping into a K-dimensional feature vector by computing its Hausdoff distance to these K medoids. After that, this ML problem was addressed by adopting MLSVM [1], which decomposes the MLL task into a set of single label classification problems. Both MIML-BOOST and MIMI-SVM did not take label correlations into account.

The to-be-detailed joint *multi-label multi-instance learning* (MLMIL) framework addresses the drawbacks of these existing methods. By simultaneously modeling the relations between the labels and regions, as well as the correlations among the multiple labels, the proposed MLMIL method solves the multi-label and multi-instance problems in an integrated manner.

3. Joint Multi-Label Multi-Instance Learning

Let $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label space, respectively. The training dataset is denoted by $\{(\mathbf{x}^1,\mathbf{y}^1),(\mathbf{x}^2,\mathbf{y}^2),\cdots(\mathbf{x}^N,\mathbf{y}^N)\}$ where $\mathbf{x}^i \in \mathcal{X}$ indicates a bag (image) of instances (regions) and $\mathbf{x}^i = \{\mathbf{x}^i_1,\mathbf{x}^i_2,\cdots,\mathbf{x}^i_{R_i}\}$. $\mathbf{x}^i_j$ denotes the feature vector of the jth instance. $\mathbf{y}^i$ is a K dimensional label vector $[y^i_1,y^i_2,\cdots,y^i_K]^T$ and $y^i_k \in \{+1,-1\}$. Each entry y^i_k indicates the membership associating $\mathbf{x}^i$ with the kth label. The task is to learn a classification function $f:\mathcal{X}\rightarrow\mathcal{Y}$ from the training dataset. However, the relation between $\mathbf{y}^i$ and each instance $\mathbf{x}^i_j$ is not explicitly indicated in the training data. Therefore, we introduce an intermediate hidden variable $\mathbf{h}^i_j$ for $\mathbf{x}^i_j$, where $\mathbf{h}^i_j$ is a binary K dimensional vector indicating the label vector of each instance. Such hidden variables explicitly capture the semantic meanings of the instances and the connection between the instances and the bag labels. As demonstrated in [17], *Hidden Conditional Random Fields* (HCRFs) is able to capture such model structure. Accordingly, we model the multi-label multi-instance learning problem based on HCRFs.

3.1. Formulation

For any image, the posterior distribution of the label vector $\mathbf{y}$ given the observation $\mathbf{x}$ can be obtained by integrating out the latent variables $\mathbf{h}$. We formulate the MLMIL problem as

$$P(\mathbf{y}|\mathbf{x};\theta) = \sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) = \frac{\mathbf{1}}{\mathbf{Z}(\mathbf{x})} \sum_{\mathbf{h}} \exp\{\mathbf{\Phi}(\mathbf{y},\mathbf{h},\mathbf{x};\theta)\} \ , \tag{1}$$

where $Z(\mathbf{x}) = \sum_{\mathbf{y}} \sum_{\mathbf{h}} \exp(\Phi(\mathbf{y},\mathbf{h},\mathbf{x};\theta))$ is a partition function. $\Phi(\mathbf{y},\mathbf{h},\mathbf{x};\theta)$ is a scale-valued potential function parameterized by θ. For the sake of simplicity, we drop θ from the formula in the following context.

This probabilistic model makes it feasible to incorporate the connection between labels and regions, the spatial relations between the region labels, and the correlations among the labels in a single unified formulation. To encode all this information, we decompose the overall potential function Φ into four component potential functions according to the relations between those variables

$$\Phi(\mathbf{y},\mathbf{h},\mathbf{x}) = \Phi_a(\mathbf{h},\mathbf{x}) + \Phi_s(\mathbf{h},\mathbf{x}) + \Phi_{hy}(\mathbf{h},\mathbf{y}) + \Phi_{yy}(\mathbf{y}) \ , \tag{2}$$

where Φ_a models the association between hidden labels and the corresponding instances, Φ_s aims to model the spatial dependence among the hidden labels, Φ_{hy} associates the hidden variables to bag labels, and Φ_{yy} models the correlation of bag labels. In the following, we give the details of these four potentials.

Association between a region and its label: This association potential function is designed for modeling the latent labels of the regions. We define a local association potential$\phi(\mathbf{h}_j,\mathbf{x}_j)$ to capture the appearance of each region, which is dependent on the corresponding region rather than the entire image. It can be modeled by a local soft classifier. To reduce the computational complexity, we assume that each region is related to at most one label which is reasonable in most real world cases. Based on the posterior probability $P(\mathbf{h}_j|\mathbf{x}_j;\lambda)$ from the classifier, the association potential is given by

$$\Phi_a(\mathbf{h},\mathbf{x}) = \sum_j \phi(\mathbf{h}_j,\mathbf{x}_j) = \sum_j \log P(\mathbf{h}_j|\mathbf{x}_j;\lambda); 1 \le j \le R \ , \tag{3}$$

where λ are the parameters of the classifier and R is the number of regions. In this paper, we learn the local soft classifier using *Support Vector Machine* (SVM) [2] [15].

Spatial relation between region labels: This interaction potential function is for modeling the spatial dependence between region labels. Intuitively, the semantic labels are spatially related. Some labels often occur in the neighboring regions, such as "mountain" and "sky." Such a spatial relation can be exploited to improve classification performance. We define the following interaction potential to capture the the neighborhood relationship of each pair of labels.

$$\begin{aligned} \Phi_s(\mathbf{h},\mathbf{x}) &= \sum_{m,n} \alpha_{m,n} f_{m,n}(\mathbf{h},\mathbf{x}) \\ &= \sum_{m,n} \alpha_{m,n} \sum_{i,j} \delta \llbracket h_{i,m} = 1 \rrbracket \, \delta \llbracket h_{j,n} = 1 \rrbracket \, \delta \llbracket \mathbf{x}_i \sim \mathbf{x}_j \rrbracket \ , \\ & m,n \in \{1,2,\cdots,K\}, 1 \le i < j \le R \end{aligned} \tag{4}$$

where $\delta \llbracket \cdot \rrbracket$ is an indicator function that takes on value 1 if the predicate is true and 0 otherwise. m and n are the label indices, while i and j are the region indices. $h_{i,m}$ denotes the mth entry of $\mathbf{h}_i$. $\mathbf{x}_i \sim \mathbf{x}_j$ indicates that region i is adjacent to region j. $\alpha_{m,n}$ is the weighting parameter.

Coherence between region and image labels: This coherence potential function models the coherence between region labels and image labels. According to the bag-instance setting, for a specific label, an image is labeled positive if at least one region has the corresponding semantic meaning; otherwise, it is negative. To impose the consistency between $\mathbf{h}$ and $\mathbf{y}$, we adopt the commonly used Ising model [14] to formulate $\Phi_{hy}(\mathbf{h},\mathbf{y})$ as $\gamma \mathbf{v}^T \mathbf{y}$, which penalizes the inconsistency between $\mathbf{v}$ and $\mathbf{y}$ by cost γ. $\mathbf{v}$ is a K-dimensional label vector, where the ith entry is defined as

$$v_i = \begin{cases} +1, & \text{if} \quad \exists_{1 \le r \le R} \, h_{r,i} = 1, \\ -1, & \text{if} \quad \forall_{1 \le r \le R} \, h_{r,i} \ne 1. \end{cases} \tag{5}$$

Correlations of image labels: This correlation potential function is designed for modeling the label correlations. In real world, the semantic labels do not exist in isolation. Instead, they appear correlatively and naturally interact with each other at the semantic level. For example, "sheep" and "grass" often appear simultaneously, while "fire" and "water" commonly do not co-occur. These correlations can serve as a useful hint to improve the classification performance. We define the following potential to exploit such correlations.

$$\begin{aligned} \Phi_{yy}(\mathbf{y}) &= \sum_{k,l} \sum_{p,q} \mu_{k,l,p,q} f_{k,l,p,q}(\mathbf{y}) \\ &= \sum_{k,l} \sum_{p,q} \mu_{k,l,p,q} \delta \llbracket y_k = p \rrbracket \, \delta \llbracket y_l = q \rrbracket \ , \\ & p,q \in \{+1,-1\}, 1 \le k,l \le K \end{aligned} \tag{6}$$

where k and l are the label indices, p and q are the binary labels (positive and negative label). $\mu_{k,l,p,q}$ is the weighting parameter.

The potential $\Phi_{yy}(\mathbf{y})$ serves to capture the relations between all the possible pairs of labels. Note that both the positive and negative relations are captured with this potential. For example, the label "sheep" and "grass" is a positive label pair while "fire" and "water" is a negative label pair. Furthermore, we can also model high-order correlations. However, the cost of employing such statistics may surpass the benefits they can bring since it will require more training samples to estimate more parameters.

3.2. Learning

Let $\tilde{\Gamma}$ and Γ denote the empirical and model distribution, respectively. The parameters in MLMIL model are estimated under the criterion of the penalized maximum likelihood with respect to the conditional distribution

$$\begin{aligned} L(\theta) &=< \log P(\mathbf{y}|\mathbf{x};\theta) >_{\tilde{\mathbf{\Gamma}}} - \tfrac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}} \left\|\theta\right\|^{\mathbf{2}} \\ &=< \log \textstyle\sum_{\mathbf{h}} P(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) >_{\tilde{\Gamma}} - \tfrac{1}{2\sigma^2} \left\|\theta\right\|^2 \end{aligned} \quad , \qquad (7)$$

where $< \cdot >_P$ denotes the expectation with respect to distribution P. The first term in Equation (7) is the log-likelihood of the training data. The second term is a penalization factor to improve the model's generalization ability. It is the log of a Gaussian prior with variance σ^2, i.e, $p(\theta) \sim \exp(-\frac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}} \left\|\theta\right\|^{\mathbf{2}})$. However, it is difficult to optimize $L(\theta)$ directly. Instead, we use the *Expectation Maximization* (EM) algorithm [4] to solve this optimization problem as follows.

E-Step: Given the current tth step parameter estimation $\theta^{(\mathbf{t})}$, the Q-function (i.e., the expectation of $L(\theta)$ under the current parameter estimates) can be written as

$$Q(\theta, \theta^{(\mathbf{t})}) =< \mathbf{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})}} \log \mathbf{P}(\mathbf{y},\mathbf{h}|\mathbf{x};\theta) >_{\tilde{\mathbf{\Gamma}}} - \tfrac{\mathbf{1}}{\mathbf{2}\sigma^{\mathbf{2}}} \left\|\theta\right\|^{\mathbf{2}} \quad , \qquad (8)$$

where $\mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})}}$ is the expectation operator given the concurrent estimated conditional probability $P(\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})})$.

M-Step: A new parameter vector $\theta^{(\mathbf{t+1})}$ is updated by maximizing the Q-function:

$$\theta^{(\mathbf{t+1})} = \arg\max_{\theta} \mathrm{Q}(\theta, \theta^{(\mathbf{t})}) \quad . \qquad (9)$$

The derivatives of Q-function with respect to its parameters are

$$\begin{aligned} &\tfrac{\partial Q}{\partial \gamma} =< \mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})}} \mathbf{v}^T\mathbf{y} >_{\tilde{\Gamma}} - < \mathbf{v}^T\mathbf{y} >_{\Gamma} - \tfrac{1}{\sigma^2}\gamma \\ &\tfrac{\partial Q}{\partial \alpha_{m,n}} =< \mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})}} f_{m,n}(\mathbf{h},\mathbf{x}) >_{\tilde{\mathbf{\Gamma}}} \\ &- < f_{m,n}(\mathbf{h},\mathbf{x}) >_{\mathbf{\Gamma}} - \tfrac{\mathbf{1}}{\sigma^{\mathbf{2}}}\alpha_{\mathbf{m},\mathbf{n}} \\ &\tfrac{\partial Q}{\partial \mu_{k,l,p,q}} =< \mathrm{E}_{\mathbf{h}|\mathbf{y},\mathbf{x};\theta^{(\mathbf{t})}} f_{k,l,p,q}(\mathbf{y}) >_{\tilde{\mathbf{\Gamma}}} \\ &- < f_{k,l,p,q}(\mathbf{y}) >_{\mathbf{\Gamma}} - \tfrac{\mathbf{1}}{\sigma^{\mathbf{2}}}\mu_{\mathbf{k},\mathbf{l},\mathbf{p},\mathbf{q}} \end{aligned} \quad . \qquad (10)$$

Given the above derivatives, we can use a gradient-based algorithm to maximize $Q(\theta, \theta^{(\mathbf{t})})$. However, this procedure requires computing the expectation under the model distribution, which is NP-hard due to the partition function. To overcome this difficulty, various approximate inference algorithms can be used. One possible solution is a sampling-based method such as *Markov chain Monte Carlo* (MCMC) [8]. However, sampling-based methods may take a large number of iterations to converge. Here we resort to *contrastive divergence* (CD) algorithm [12], which only needs to take a few steps in the Markov chain to approximate the gradients. This property of CD can lead to huge savings particularly when the inference algorithm will be repeatedly invoked during the model training. Note that, in each the iteration of EM algorithm, we train the local classifier before the other components by adopting the standard quadratic optimization. This sequential solution is more efficient than joint training all the components [11].

3.3. Inference

Given a new image $\mathbf{x}$, the inference procedure is to find the optimal label configuration $\mathbf{y}$. A widely-used criteria for inferring labels from the posterior distribution is *Maximum Posterior Marginal* (MPM) [14] [11], which is adopted in this paper. The computation of MPM requires marginalization over a large number of variables, which is generally NP-hard. To tackle this difficulty, we adopt a frequently-used approximate inference method, Gibbs sampling [8], because of its fast convergence. A reasonable initial point for the sampling can be obtained by considering the outputs of the local classifiers. Using a similar approach, we also estimate the region label $\mathbf{h}$.

4. Experiments

We constructed extensive experiments to compare the innovative framework against other four representative methods from the three paradigms: (1) a state-of-the-art multi-instance learning approach MILES [20], which has been reported to outperform many other competitive MIL approaches for image classification; (2) a representative multi-label learning approach CML [16] which is also a competitive method due to it captures the label correlations; and two multi-label multi-instance learning approaches reported in [22], i.e., (3) MIML-SVM and (4) MIML-BOOST, which translate the multi-label multi-instance learning task into typical multi-label learning and multi-instance learning problem, respectively. The comparison is conducted over two data sets, *i.e.*, *Microsoft Research Cambridge* (MSRC) and Corel data set.

4.1. Evaluation on MSRC data set

MSRC data set contains 591 images with 23 classes. Around 80% images are associated with more than one label and there are around three labels per image on average. These labels often arise from respective regions in the images. Figure 2 illustrates some sample images in this data set. MSRC data set also provides pixel level ground truth, where each pixel is labeled as one of 23 classes or "void." We treat "horse" and "mountain" as "void" since they have few positive samples. Thus there are 21 labels in total. Note that we only use the image-level ground truth to train the models.

We have performed 5-fold cross validation over MSRC data. Specifically, the images were randomly splitted into

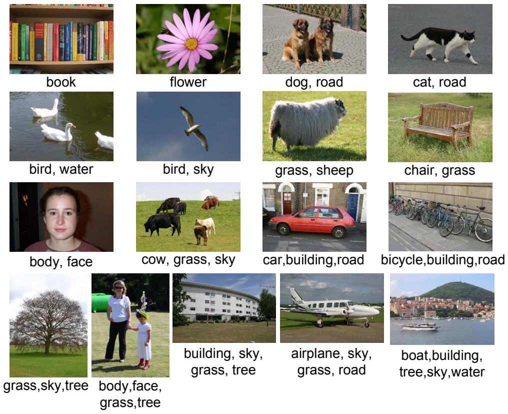

Figure 2. Sample images from MSRC data set.

five parts with equal size and an additional constraint that there should be at least five positive images of each class per partition. We selected each of the five parts as testing set, and the others as training set. The average performance over five iterations is reported for evaluation.

In our implementation, all the images were firstly segmented using JSEG [5]. A set of low-level features was extracted from each region to represent an instance, including region size, color correlogram, color moment, wavelet texture and shape descriptors [3]. We constructed the bag feature for CML by concatenating a fixed number of instances [22]. For the sake of fair comparison, we also fixed the region number for MIML-BOOST, MIML-SVM, MILES, and the proposed MLMIL method. All the algorithmic parameters in all five approaches were determined by a twofold cross-validation process on training set. The reported performance were from the best set of parameters in the five approaches.

There are various measurements for evaluating the classification performance, including ROC curve, precision-recall curve, and so on. The most widely accepted measurement is AUC (area under ROC curve) [10], which is adopted in this paper. Specifically, AUC describes the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample. Table 1 gives the comparison results of different models in terms of average AUC over the 21 labels, while Figure 4 illustrates the detailed results for individual labels. From the experimental results, the following observations can be obtained:

- MLMIL achieves the best overall performance and obtains around 8.8%, 10.3%, 11.5% and 17.8% improvement compared to CML, MILES, MIML-SVM, and MIML-BOOST, respectively.
- MLMIL performs the best on 19 of all the 21 labels. By exploiting the label correlations, MLMIL outperforms the approaches which treat semantic labels separately and neglect their interactions. The significant pair-wise label correlations are illustrated in Figure 3 (please be noted that this correlation matrix is not used explicitly in our MLMIL approach, instead, we encode it in potential function (7)). On the other hand, the labels are usually characterized by several regions rather than the entire image. As illustrated in Figure 5, "road" is related to several regions in the three images. These regions appear similarly, while the three images have various appearances. By connecting labels with regions, MLMIL reduces the noises in the corresponding feature space. Therefore, MLMIL can perform better than CML which regards an image as one indiscrete entity. The experimental results confirmed this observation. MLMIL outperforms CML over all the 21 labels.
- MLMIL degrades slightly on two labels: "book" and "sign" compared to MILES. The main reason is that each of these two labels has weak interactions with other labels. As a result, the presence/absence of these two labels cannot benefit from those of the others.

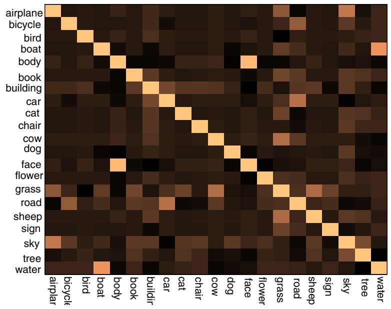

Figure 3. The pair-wise label correlations measured by the normalized mutual information between each pair of the 21 labels in the MSRC data set.

Approach	Avg. AUC
CML [16]	0.829
MILES [20]	0.818
MIML-BOOST [22]	0.766
MIML-SVM [22]	0.809
MLMIL	**0.902**

Table 1. The image level average AUC for MSRC data set by different approaches.

Approach	Avg. AUC
MILES [20]	0.736
MIML-BOOST [22]	0.652
MLMIL	**0.863**

Table 2. The region level average AUC for MSRC data set by the three approaches.

Figure 5. Example images contain regions related to “road.” These regions appear similarly, while the entire images have various appearances.

Figure 6. Sample images from Corel 1000 data set.

Approach	Avg. AUC
CML [16]	0.851
MILES [20]	0.828
MIML-BOOST [22]	0.796
MIML-SVM [22]	0.830
MLMIL	**0.913**

Table 3. The image level Average AUC on 1000 image Corel data set by different approaches.

- MLMIL can label not only an image but also label the regions within this image. The experimental results proved MLMIL has a competitive performance at region level. As shown in Table 2, MLMIL achieves the best region level performance compared to MILES and MIML-BOOST which can also generate the region labels.

In summary, MLMIL consistently achieved the best performance on diverse 21 labels among the five methods.

4.2. Evaluation on Corel 1000 data set

The second experiment was carried out on Corel data set from [6]. There are 50 Stock Photo CDs in this data set. Each CD includes 100 images on the same topic. All the images have been manually annotated with 1 $\sim$ 5 labels and there are 374 labels. We conducted the experiments on around 1000 images for the ten object/scene classes: “bird,” “boat,” “building,” “flower,” “grass,” “mountain,” “people,” “sky,” “tree,” and “water.” Figure 6 shows some example images. In our implementation, we followed the same setup in Section 4.1. Table 3 and Table 4 show the comparisons of the performances on image and region level, respectively. We can observe that the proposed MLMIL method achieves the best performance both at image and region level. We provide the detailed AUC value for individual labels in Figure 7. MLMIL performs the best on all the 10 labels.

Approach	Avg. AUC
MILES [20]	0.752
MIML-BOOST [22]	0.647
MLMIL	**0.872**

Table 4. The region level average AUC on 1000 image Corel data set by the three approaches.

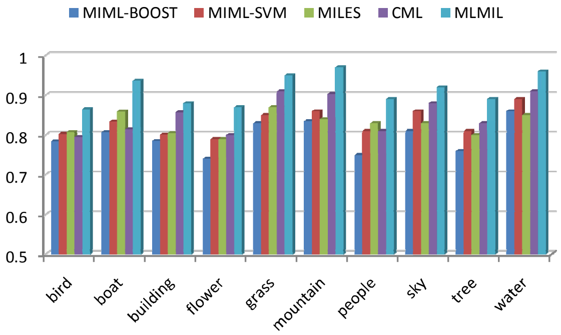

Figure 7. AUC value of 10 labels on Corel data set by MIML-BOOST, MIML-SVM, MILES, CML, and MLMIL.

5. Conclusion

In this paper, we proposed a joint multi-label multi-instance learning framework for image classification. Our MLMIL framework can model both the relation between semantic labels and regions, and the correlations among the labels in an integrated manner. Also, MLMIL is flexible to capture various dependencies between regions, such as the spatial configuration of the region labels. We evaluate the performance of MLMIL on two challenging data sets: MSRC data set and 1000 image Corel data set. The experiments validate MLMIL achieves the high classification performance on both image and region level, and is robust to different data sets.

References

[1] M. Boutell, J. Luo, X. Shen, and C. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, September 2004. 1, 2, 3

[2] C. C. Chang and C. J. Lin. *LIBSVM: a library for support vector machines*, 2001. 4

[3] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004. 1, 3, 6

[4] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society, Series B*, 39(1), 1977. 3, 5

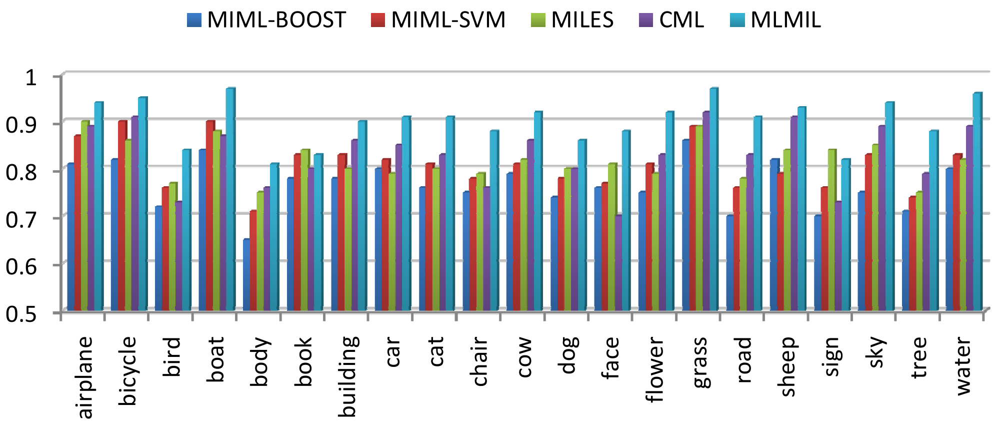

Figure 4. AUC value of 21 labels on MSRC data set by MIML-BOOST, MIML-SVM, MILES, CML, and MLMIL.

[5] Y. Deng and b. s. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(8):800–810, 2001. 6

[6] P. Duygulu, K. Barnard, J. d. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of European Conference on Computer Vision*, pages IV:97–112, 2002. 7

[7] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola. Multi-instance kernels. In *Proc. of International Conference on Machine Learning*, pages 179–186, 2002. 1, 3

[8] S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.*, 6(6):721–741, Nov. 1984. 5

[9] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2004. 1, 2

[10] J. A. Hanley and B. J. Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, April 1982. 6

[11] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *Proc. of IEEE Conference on Computer Vision & Pattern Recognition*, pages 695–702, 2004. 5

[12] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Comput.*, 14(8):1771–1800, 2002. 5

[13] F. Kang, R. Jin, and R. Sukthankar. Correlated label propagation with application to multi-label learning. In *Proc. of IEEE International Conference on Computer Vision & Pattern Recognition*, pages 1719–1726, 2006. 1, 3

[14] S. Kumar and M. Hebert. Discriminative random fields: A discriminative framework for contextual interaction in classification. In *Proc. of the International Conference on Computer Vision (ICCV)*, volume 2, pages 1150–1157, 2003. 4, 5

[15] C. H. Lee, R. Greiner, and M. Schmidt. Support vector random fields for spatial classification. In *Proc. of European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 121–132, 2005. 4

[16] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM International Conference on Multimedia*, pages 17–26. ACM, 2007. 1, 3, 5, 6, 7

[17] A. Quattoni, S. B. Wang, L.-P. Morency, M. Collins, and T. Darrell. Hidden conditional random fields. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(10):1848–1852, 2007. 3

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *Proc. of European Conference on Computer Vision*, pages I: 1–15, 2006. 1

[19] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 272–281, 2004. 1, 3

[20] J. B. Yixin Chen and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):1931–1947, 2006. 1, 3, 5, 6, 7

[21] Q. Zhang and S. A. Goldman. Em-dd: An improved multiple-instance learning technique. In *Advances in neural information processing systems*, pages 1073–1080, 2001. 3

[22] Z.-H. Zhou and M.-L. Zhang. Multi-instance multi-label learning with application to scene classification. *Advances in Neural Information Processing Systems 19*, pages 1609–1616, 2007. 1, 3, 5, 6, 7