A PLSA-based Semantic Bag Generator with Application to Natural Scene Classification under Multi-Instance Multi-Label Learning Framework*

Shuangping Huang, Lianwen Jin

School of Electronic and Information Engineering, South China University of Technology, Guangzhou, 510640, P. R. China
{huangshuangping, lianwen.jin}@gmail.com

Abstract

Classifying natural scenes into semantic categories has always been a challenging task. So far, many works in this field are primarily intended for single label classification, where each scene example is represented as a single instance vector. The multi-instance multi-label (MIML) learning framework proposed by Z. H. Zhou et al [1] provides a new solution to the problem of scene classification in a different way. In this paper, we propose a novel scene classification method based on pLSA-based semantic bag generator and MIML learning framework. Under the framework of MIML learning, we introduce the mechanism that transfers an image into a set of instances through the pLSA-based bag generator. Experiments show that our approach achieves better classification performance comparing with the previous work.

1. Introduction

Classifying natural scene images into semantic categories has attracted more and more interests from researchers in both academy and industry in recent years [1~6]. It is a challenging task due to the variability, ambiguity, and the wide range of illumination and scale conditions that may apply [3].

Recently some successful works are reported in literatures [1~6] to address the challenges of scene classification. Among them, Z. H. Zhou [1] proposed multi-instance multi-label (MIML) learning as a promising framework, which provides a new solution for multi-label natural scene classification. Zhou’s approach mainly consists of two parts, multi-instance bag generator and MIML learning algorithm. Bag generator is a mechanism that can translate an image into a set of instances referred to as image bag. MIML learning algorithm is mainly for solving problem of the image bag’s classification, which takes into consideration the ambiguity of image representation and labeling.

1.1. MIML learning algorithm

An image usually contains multiple regions and each sub-region can be represented by an instance. In the mean time an image may also belong to multiple classes simultaneously. This is typical of what we see in real world. It is referred to as MIML phenomenon, which means there is explicit ambiguity with both input representation and output labeling. Figure 1 shows a typical example of MIML phenomenon. The image in Figure 1 can be viewed as mountain, lake and trees simultaneously.

Figure 1: A multi-label example from dataset [1]

MIML is a good framework for learning with ambiguous data and it offers a good many-to-many mapping mechanism between multiple instances and multiple labels. The framework is illustrated in Figure 2. Early research on multi-label learning (MLL) [7]

*This paper is sponsored in part by NSFC (no U0735004, 60772116)
and multi-instance learning (MIL) [8] can be regarded as degenerated versions of MIMIL.

Z. H. Zhou et al first proposed two MIML algorithms, MIMLSVM and MIMLBOOST in application to natural scene classification [1]. Among them, MIMLSVM algorithm uses MLL as bridge, which first carries out representation transformation on original MIMIL data set in order to obtain a single-instance multi-label data set and the MLSVM [9] algorithm is used for multi-label learning. MIL further decomposes the multi-label learning problem into multiple independent binary classifications, which is in fact traditional single-instance single-label learning problem (SISL). MIMLBOOST algorithm uses MIL as bridge, which first carries out category-wise decomposition on MIMIL examples to obtain multi-instance data set and then uses MIBOOSTING [10] to implement MIL learning on these multi-instance examples. The two MIML solutions are illustrated in Figure 3. Details can be found in [1].

SBN bag generator was proposed by Maron et al [2] based on color statistics and color distribution. In SBN approach, an instance is the mean color of a $2 \times 2$ blob and the color difference with its four neighboring blobs. In order to reduce the number of instances in an image bag, each image is initially sub-sampled to $8 \times 8$ after being smoothed using a Gaussian filter. This way each example is represented as a bag of nine instances.


In this paper, we propose a novel pLSA-based semantic bag generator under MIML learning framework to carry out multi-label classification of natural scenes. In our approach each scene image is firstly divided into multiple sub-regions of same size, which corresponds to a bag instance. BoVW (bag of visual words) representations of sub-images are then computed independently. Thirdly the pLSA model is used to obtain an intermediate representation of each instance, which is referred to as Z-vector [3]. In this way each picture is represented by a bag of Z-vectors. Finally all the Z-vector bags act as input to the MIML learning system. Experiments using the multi-Instance multi-Label learning algorithms MIMLSVM and MIMLBOOST [1] are conducted on the same natural scenes database as in [1]. For convenience, this scene dataset is referred to as ZM dataset here. By introduction of pLSA-based bag generator into MIML learning framework, we have achieved superior results to those from MIML based method [1].

2. PLSA-based semantic instance representation and bag generator

2.1. PLSA model

The pLSA [13] model was originally developed for topic discovery in text corpus, where each document is represented by its word frequency. It was later applied to computer vision domains as represented by the frequency of “visual words” in 2005 [14~15]. In this section, we will briefly explain how pLSA model is used in image or instance representation.

Suppose we have a set of image patches denoted as $D = \{d_1, d_2, ..., d_N\}$. Each $d_i (i = 1, 2, ..., N)$ will first be represented by means of BoVW. Before that a visual vocabulary should be formed using the method as proposed in [3]. Each element in the visual vocabulary is a “visual word” denoted as $w_i (i = 1, 2, ..., M)$. The
visual words’ frequency in \( d_i \) is represented by a histogram vector denoted as \( (n(w_1,d_i), n(w_2,d_i), ..., n(w_M,d_i)) \), where \( n(w_j,d_i) \) denotes how often the word \( w_j \) occurred in an image patch \( d_i \). Documents’ BoVW vectors assemble the word-document co-occurrence table denoted as \( N_w = n(w_j,d_i) \). Word-document co-occurrence matrix will act as the input of pLSA modeling.

During the building of the model, pLSA associates each observation with an unobserved class variable denoted as \( z_i (l = 1,2,...,L) \). Here an observation is the occurrence of a word in a particular document. A joint probability \( P(w,d) \) is modeled as following on the word-document data table:

\[
p(w,d) = p(d) \sum_{z \in Z} P(w|z)P(z|d)
\]

In equation (1), \( p(d) \) is the probability of observing a particular document, \( P(w|z) \) is topic specific distribution and \( P(z|d) \) is the document specific probability over the latent variable space. Furthermore, given equation (2), equation (1) can be simplified to equation (3):

\[
p(w,d) = p(d)P(w|d)
\]

\[
p(w|d) = \sum_{z \in Z} P(w|z)P(z|d)
\]

We use EM (Expectation Maximization) algorithm [16] to learn the unobservable probability distribution \( P(z|d) \) and \( P(w|z) \) from word-document co-occurrence data table. By observing \( P(w|z) \), we know which part of words belong to the same topic and thus can obtain a rough understanding of the meaning of each latent class. By observing \( P(z|d) \), the probability distribution over the latent variable space for an image or image patch can be estimated. Since a word may belong to more than one latent class, pLSA can also be viewed as a dimension reduction method. Detailed explanation of the model can be found in [13,15,17].

2.2. PLSA-based semantic bag generator

Our pLSA-based semantic bag generator is inspired by several previous papers [3, 1~2, 11~12, 18]. Among these works, some researchers illustrated the semantic trend of instance representation within an image bag [2, 11~12]. Anna Bosch et al [3, 18] suggested that pLSA has the ability to obtain a robust, low-dimensional instance representation and it automatically captures meaningful scene aspects. Z. H. Zhou et al [1] proposed using MIML learning algorithm to solve multi-instance multi-label natural scene classification where multi-instance representation is as important as a strong learning algorithm.

In our approach, each image is divided into several image patches of the same size. These image patches are considered as independent documents. All the image patches form a collection of documents. Each document is represented as a BoVW based on a visual vocabulary that was formed previously. Then we use the co-occurrence table as the input for pLSA modeling. The corresponding result is intermediate representation of each instance, which is referred to as \( z \)-vector denoted as \( p(z/d) \). In this way each picture is represented by a bag of \( Z \)-vectors \( p(z/d) \), which finally act as the input of MIML learning system to classify natural scenes. In more details, the procedure of our approach is given as following.

Firstly, the visual vocabulary has to be constructed which involves the following three steps: (i) automatically detect points of interest by means of dense grid, (ii) compute local descriptors over those points where the color SIFT [18] descriptor is applied, (iii) cluster all the SIFT vectors from training images using k-means method to form a specified number of words, which is called visual vocabulary.

Secondly, each image is divided into several image patches of same size. Each patch is regarded as an independent document and its BoVW representation can be computed by quantization of its descriptors into words. All the training and testing documents’ BoVW assemble co-occurrence table separately.

Thirdly, pLSA model is built. Two topic-relevant probability distributions \( p(w|z) \) and \( p(z|d_{\text{train}}) \) are learnt from training procedure to fit pLSA model to the entire set of training documents. Then the pLSA testing is carried out based on \( p(w|z) \) in order to obtain \( P(z|d_{\text{test}}) \). All the \( P(z|d) \) vectors act as instance representations.

Finally, \( z \)-vectors form the corresponding image bags that are input to the MIMLSVM or MIMLBOOST classifier to produce the classification result.

For clarity, we assume the following: each example is divided into four sub-images and the topic number adopted during pLSA training is 15, so pLSA-based bag generator will translate each image into a bag of four instances. Each instance is 15-dimension vector and corresponds to the probability distributions of specific document over latent variable space.
2.3. Hierarchy structure of our approach

The hierarchy structure of our pLSA-based semantic bag generator is shown in Figure 4, which consists of four layers. In the first (lowest) layer low-level feature such as SIFT is used to obtain structure information of image. This is an important basis for the final bag representation and it affects the performance of classification system. However, it may not be enough for robust classification problems if we only rely on delicate low-level features. This is especially the case for large number of image categories with large variety. The second layer is corresponding to image sub-regions, which is very important especially for multi-label object or scene image. Each sub-image consists of a number of low-level feature vectors referred to as document, where BoVW representation is applied. In the third layer, pLSA provides a higher level of semantic grouping of image patches by taking into consideration their co-occurrence relationships. In the last layer multiple \( p(z_i/d) \) vectors form instance bag, which is image representation referred to as image bag.

Finally, our classifying system is constructed by introducing this novel pLSA-based semantic bag generating mechanism into MIML framework. The system overview is illustrated in Figure 4.

3. Experiments

We evaluate our methodology on multi-label image dataset ZM [1] and compare classification performance with that of Z.H Zhou’s methods [1], where SBN bag generator was applied.

3.1. Experimental setup

The ZM dataset contains 2000 color natural scene images, which belong to the five classes of desert, mountains, sea, sunset, and trees. Over 22% of the images belong to multiple classes simultaneously. The size of images varies a lot, where the width ranges from 42 to 768 pixels and the height from 66 to 1029 pixels. Some of the images are extremely slim and some are extremely wide and flat. In addition, every scene category is characterized by high degree of diversity and potential ambiguities. The detailed description of ZM dataset can be found in [1].

For fair comparison, we used the same experimental setup as in [1], where a training set is created by randomly selecting 80% from the entire dataset and the remaining 20% is for testing.

We divide each image into four sub-regions since the maximum number of labels associating with one image...
is less than four. That means each image will be represented as a bag of four instances.

The dense grid point is extracted at a step of 10 pixels. At each grid point, color SIFT descriptors are computed over multiple circular support regions [18].

From experiments, we found parameters such as number of topics, number of concentric circular support regions etc., will affect classification performance to certain extent. We will investigate the impact of these parameters on the classification performance in more details as following.

3.2. Performance comparison with different parameters

The classification performance is evaluated according to five multi-label evaluation metrics, which includes HammingLoss, RankingLoss, OneError, Coverage, and Average_Precision [7]. The smaller HammingLoss, RankingLoss, OneError, Coverage are, the better the performance is. Meanwhile, the bigger the Average_Precision is, the better the performance is.

3.2.1. Performance against different latent topic number. Figure 5 shows results of changing the topic number with a step of 5. It can be seen that when the topic number is about 20, the system achieves the best performance.

3.2.2. Performance against different number of support regions.

Table 1 shows results of changing the number of support regions from 1 to 4. It can be seen that the number of support regions is an important factor that will affect the system performance. The best results are obtained with four support regions.

3.3. Comparison with previous work

According to the experiments shown in section 3.2, the best results overall are obtained with dense color SIFT, four circular supports for each interest point, and 20 topic number for pLSA training. Table 2 shows the comparison of results using our approach with Zhou’s method [1], which uses MIMLSVM learning algorithm. Table 3 shows the comparison of results using MIMLBOOST algorithm. From Table 2 and Table 3, one can see that our approach results in significantly better performance than that of Zhou’s [1]. Among the performance measures in the Table 2 and Table 3, Average_Precision increases by almost 9%, while the other four decrease by 22% to 34%. Comparing Table 2 to Table 3, we can find that MIMLSVM is slightly better than MIMLBOOST with 25 boosting rounds when using our approach.

4. Conclusions and future work

In this paper, we propose a novel semantic bag generator and introduce it into MIML framework for the classification of multi-label natural scenes. Using our approach, superior results have been obtained to that was reported previously in the literature [1]. Comparing with other bag generating methods, pLSA-based semantic bag generator has the following advantages: (i) It’s segmentation-free, (ii) it captures meaningful representation of bag instance by means of pLSA, (iii) it reduces the number of instances in a bag and dimension of instance feature vector, that will significantly reduce the running time during MIML classification.
Our method could be also extended to deal with object image classification, especially in cases where the object image belongs to multi-label simultaneously. That is one of the potential directions that merit our future study. One point to be sure is that we should try to seek more suitable interest point extraction and description method to obtain better low-level feature for object image. In addition, how to obtain a better visual vocabulary is another important issue to be studied in the future.

### Table 2: Performance comparison between method in [1] and our approach using MIMLSVM

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HammingLoss ↓</td>
<td>0.180±0.0017</td>
<td>0.135±0.0027</td>
</tr>
<tr>
<td>RankingLoss ↓</td>
<td>0.187±0.018</td>
<td>0.129±0.0044</td>
</tr>
<tr>
<td>OneError ↓</td>
<td>0.327±0.033</td>
<td>0.221±0.009</td>
</tr>
<tr>
<td>Coverage ↓</td>
<td>1.022±0.085</td>
<td>0.793±0.0175</td>
</tr>
<tr>
<td>Average Precision ↑</td>
<td>0.783±0.020</td>
<td>0.85±0.0054</td>
</tr>
</tbody>
</table>

### Table 3: Performance comparison between method in [1] and our approach using MIMLBOOST

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>HammingLoss ↓</td>
<td>0.189±0.009</td>
<td>0.137±0.004</td>
</tr>
<tr>
<td>RankingLoss ↓</td>
<td>0.181±0.026</td>
<td>0.138±0.001</td>
</tr>
<tr>
<td>OneError ↓</td>
<td>0.351±0.039</td>
<td>0.233±0.004</td>
</tr>
<tr>
<td>Coverage ↓</td>
<td>0.989±0.103</td>
<td>0.813±0.011</td>
</tr>
<tr>
<td>Average Precision ↑</td>
<td>0.777±0.025</td>
<td>0.847±0.001</td>
</tr>
</tbody>
</table>

### Reference