Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence

Adaptive Knowledge Transfer for Multiple Instance Learning in Image Classification

Qifan Wang, Lingyun Ruan and **Luo Si**
Computer Science Department, Purdue University
West Lafayette, IN 47907, US
wang868@purdue.edu, ruan1@purdue.edu, lsi@purdue.edu

Abstract

Multiple Instance Learning (MIL) is a popular learning technique in various vision tasks including image classification. However, most existing MIL methods do not consider the problem of insufficient examples in the given target category. In this case, it is difficult for traditional MIL methods to build an accurate classifier due to the lack of training examples. Motivated by the empirical success of transfer learning, this paper proposes a novel approach of Adaptive Knowledge Transfer for Multiple Instance Learning (AKT-MIL) in image classification. The new method transfers cross-category knowledge from source categories under multiple instance setting for boosting the learning process. A unified learning framework with a data-dependent mixture model is designed to adaptively combine the transferred knowledge from sources with a weak classifier built in the target domain. Based on this framework, an iterative coordinate descent method with Constraint Concave-Convex Programming (CCCP) is proposed as the optimization procedure. An extensive set of experimental results demonstrate that the proposed AKT-MIL approach substantially outperforms several state-of-the-art algorithms on two benchmark datasets, especially in the scenario when very few training examples are available in the target domain.

Introduction

With the explosive growth of data on the internet, a huge amount of images has been generated and thus automatic image classification has become increasingly important. Multiple Instance Learning (MIL) (Dietterich, Lathrop, and Lozano-Pérez 1997; Zhou, Sun, and Li 2009) is a popular technique in machine learning that addresses the classification problem of a bag of data instances. In MIL, each bag contains multiple data instances associated with input features. The purpose of MIL is to predict labels of bags based on all the instances in individual bags with the assumption that *a bag is labeled positive if at least one of the instances is positive, whereas a negative bag is only composed of negative instances*. For image classification, each image is treated as a bag and different regions inside the image are viewed as individual data instances.

One major advantage of MIL comes from the fact that in training process it only requires the label information of a bag instead of all individual instances in the bag. However, due to the label ambiguity issue of individual data instances in the MIL setting, traditional supervised classification methods for single instance learning may not be directly applied. Multiple instance learning methods have generated promising results in image classification. This is because the concept/object is usually contained in some certain region of the image, which is consistent with multiple instance setting. However, most existing MIL methods do not consider the problem when the number of training examples in the given target category is insufficient. In this case, it is difficult for traditional MIL methods to build accurate classifiers without sufficient training examples.

To address this problem, this paper proposes a novel approach of Adaptive Knowledge Transfer for Multiple Instance Learning (AKT-MIL) in image classification. The new method transfers cross-category knowledge from source categories in multiple instance setting for boosting the learning process in the target domain. The basic idea for the framework is that the modeling of the target category can become more effective and simpler with the extra information contained in the source categories. Our key observation is that semantic concepts contained in data instances do not exist independently, since many of them are closely correlated with each other. Fig.1 shows an example of several images from three correlated categories, 'sea', '$sand$' and 'sky', where instances in these categories appear concurrently. There exists certain correlation among these categories, and thus makes it possible to transfer knowledge across categories.

There are two main challenges in designing the knowledge transfer algorithm in multiple instance setting. Firstly, how to transfer knowledge across categories? We have no prior information about the correlations among different source categories and the target category. Moreover, we do not know which data instance in the bag represents the semantic concept in target category. To overcome this problem, we propose to use label propagation in multiple instance setting from different source categories to the target category by exploring the semantic correlation between categories. Secondly, when to transfer knowledge across categories? Transfer learning sometimes has detrimental effects when the knowledge propagation is noisy (Pan and Yang 2010; Kuzborskij, Orabona, and Caputo 2013). It

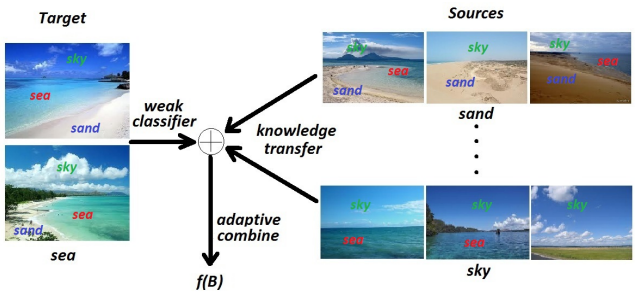

Figure 1: An overview of the proposed AKT-MIL approach.

could harm the modeling performance of the target category when transfer learning is used inappropriately. To avoid potential negative transfer, we propose a data-dependent mixture model that adaptively adjust the importance for transferring knowledge.

The main contributions of this paper are: (1) the proposed AKT-MIL approach intelligently transfers knowledge from source categories to a particular target category in multiple instance setting by exploring the semantic correlation among different categories. (2) an effective iterative coordinate descent method together with Constrained Concave-Convex Procedure (CCCP) is proposed in the learning process as the optimization algorithm. (3) we develop a data-dependent mixture model that adaptively adjusts the weights of transferred knowledge from source categories.

Related Work

Multiple Instance Learning

Image classification algorithms based on multiple instance learning (MIL) model the relationship between labels and regions (Hu, Li, and Yu 2008; Maron and Ratan 1998; Ray and Craven 2005; Zhang et al. 2011; Bunescu and Mooney 2007). One major challenge of MIL is the label ambiguity, i.e., which region contains the semantic concept of the target category is not known.

Existing MIL algorithms can be divided into two groups, generative models and discriminative models. Many generative algorithms predict bag labels by first inferring the hidden labels of individual instances in the bags. The Diverse Density (DD) (Maron and Lozano-Pérez 1997) method defines the DD value of data instances and uses a scaling and gradient search algorithm to find the prototype points in the instance space. The EM-DD method in (Zhang and Goldman 2001) combines the idea of Expectation-Maximization (EM) with DD to identify the most probable concept. Many discriminative methods directly predict bag labels in a large margin framework by using bag-level features. DD-SVM (Chen and Wang 2004) selects a set of instances using the DD function to train a SVM classifier based on the bag-level features. The MI-SVM (Andrews, Tsochantaridis, and Hofmann 2002) method formulates MIL as a mixed integer quadratic programming problem for learning instance and bag labels. In the work of MILES (Chen, Bi, and Wang 2006), bags are first embedded into a feature space defined by all the instances, and then a 1-norm SVM is built as the bag-level classifier. The MILBoost algorithm in (Viola, Platt, and Zhang 2005) translates MIL into the AdaBoost framework.

Recently, an instance selection MIL (IS-MIL) approach is proposed in (Fu and Robles-Kelly 2009), which selects one instance per positive bag to represent the concept. A standard SVM is applied to train the classifier based on the constructed bag-level features. The multi-label issue in multiple instance learning is explored in (Xue et al. 2011; Zha et al. 2008), which simultaneously captures both the connections between labels and regions and the correlations among the labels based on hidden conditional random fields. Most recently, a mixture model approach for MIL (Wang, Si, and Zhang 2012) has been proposed to handle the multi-target problem where positive instances may lie in different clusters in the feature space. The work in (Zhang et al. 2013) proposes to embed the global features with local features to learn more accurate classifier. A multi-view MIL method is proposed in (Zhang, He, and Lawrence 2013) to incorporate multiple features for boosting the learning performance.

Transfer Learning

In many real world applications, it is expensive or impossible to collect sufficient training examples for building accurate learning models. One possible way is to extract knowledge from other related source categories to enhance the learning process. This method is known as transfer learning. A comprehensive survey of transfer learning is summarized in (Pan and Yang 2010).

The work in (Tommasi, Orabona, and Caputo 2010) designs cross-domain adaptation by constraining the classification hyperplane in the target domain to be close to that in the source domain. A two phase transfer approach proposed in (Yao and Doretto 2010) identifies useful models from various sources to enhance the modeling of the target classifier. In work (Raykar et al. 2008), the authors address the multiple instance multiple task learning problem from a Bayesian perspective. Recently, the work in (Qi et al. 2011) proposes a cross-category knowledge transfer method that explores the knowledge in correlated categories. The cross-category classifiers are combined to integrate the knowledge from multiple source categories in an Adaboost framework. The only work we found using transfer learning in MIL setting is (Zhang and Si 2009), which directly uses instance level transfer learning to model the instance label. However, the semantic correlation among different categories is not modeled in their work. Moreover, this work does not handle the problem of when to do the transfer as we discussed before. Since transfer learning sometimes has detrimental effects, it is important to determine when the knowledge should be transferred.

Adaptive Knowledge Transfer Multiple Instance Learning

Problem Setting and Approach Overview

We first introduce some notation. There are a set of N training bags (image examples) in the target category. Let

us denote them as: $T = \{(\boldsymbol{B_i}, y_i)|i = 1, \ldots, N\}$, where $y_i \in \{+1, -1\}$ is the label of i^{th} bag. Let $\boldsymbol{B_i} = \{\boldsymbol{x_{ij}}|j = 1, \ldots, N_i\}$, where $\boldsymbol{x_{ij}}$ is the j^{th} instance in $\boldsymbol{B_i}$ and N_i is the number of instances in $\boldsymbol{B_i}$. Denote the sources as $S_l = \{(\boldsymbol{B_{l,i}}, y_{l,i})|i = 1, \ldots, N_l\}$ for $l = 1, \ldots, L$ over L source categories, where N_l is the number of bags in the l^{th} source and $y_{l,i} \in \{+1, -1\}$ is the label of source bag $\boldsymbol{B_{l,i}}$. We assume that the positive instance in source bag $\boldsymbol{B_{l,i}}$, can be estimated and denoted as $\boldsymbol{x^+_{l,i}}$[1]. This can be conducted by using any instance selection method such as (Fu and Robles-Kelly 2009; Wang, Si, and Zhang 2012) in an off-line manner.

The proposed AKT-MIL approach is a unified learning framework that consists of three components as shown in Fig.1: (1) A transfer function that propagates useful knowledge from source categories to the target category under MIL setting. (2) A weak classifier built in target category. (3) A combination method to adaptively integrate the transferred knowledge with the weak classifier. An iterative coordinate descent method with Constrained Concave-Convex Procedure (CCCP) is designed as the optimization algorithm for the unified learning framework. In the rest of this section, we first introduce the three components respectively and then give the optimization algorithm. Finally, some analysis on the convergence of the learning algorithm will be elaborated.

Knowledge Transfer

Existing MIL methods do not leverage knowledge from various source categories, although the modeling on target category could be much more effective and accurate by utilizing extra information from source categories. Therefore, how to transfer knowledge between categories becomes the key issue. In this work, we propose to directly transfer the label information from source categories to target category by exploring their semantic correlation under multiple instance setting. In particular, a discriminative classifier $f_S(\boldsymbol{B})$ is defined based on the transfer function $Tr(\boldsymbol{B}, \boldsymbol{B_{l,i}})$ to transfers the cross-category labeling information is constructed as follows:

$$f_S(\boldsymbol{B}) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_l}\sum_{i=1}^{N_l} y_{l,i} Tr(\boldsymbol{B}, \boldsymbol{B_{l,i}}) \tag{1}$$

where $y_{l,i}$ is the label of the i^{th} bag from l^{th} source, and $Tr(\boldsymbol{B}, \boldsymbol{B_{l,i}})$ is the transfer function that determines how the knowledge from a source bag $\boldsymbol{B_{l,i}}$ should be transferred to a target bag $\boldsymbol{B}$.

There are several important factors to define the transfer function. Firstly, the binary labels on the two categories can be different even for a same bag, depending on how they are collected or labeled. We need to model the semantic correlation between two different categories. For example, 'sea' and 'sky' may have high correlation, while the correlation between'sea' and '$tiger$' may be low or even negative. Therefore, we do not want to transfer knowledge from '$tiger$' to 'sea'. Secondly, it is important to measure the similarity between the two bags in order to propagate the label information. The reason is it is less likely that two totally different bags share much related knowledge, even though they are from two highly correlated categories. Therefore, the transfer function should also model the similarity between the two bags. Based on the above observations, we define the transfer function as:

$$Tr(\boldsymbol{B}, \boldsymbol{B_{l,i}}) = c_l S(\boldsymbol{B}, \boldsymbol{B_{l,i}}) \tag{2}$$

where c_l is coefficient representing the semantic correlation between the target category and the l^{th} source category. If two categories are highly correlated, the value of the corresponding c_l should also be high. $S(\boldsymbol{B}, \boldsymbol{B_{l,i}})$ measures the similarity between two bags/images. Previous transfer learning methods treat the whole image as one instance to calculate the similarity, which may introduce noise from the background regions. Therefore, instead of using the whole image $\boldsymbol{B_{l,i}}$, we use the positive instance $\boldsymbol{x^+_{l,i}}$ which contains the semantic concept of the source category to calculate the similarity between two bags as:

$$S(\boldsymbol{B}, \boldsymbol{B_{l,i}}) = s(\boldsymbol{B}, \boldsymbol{x^+_{l,i}}) = \max_{\boldsymbol{x_j} \in \boldsymbol{B}} \exp\left(-\frac{\|\boldsymbol{x_j} - \boldsymbol{x^+_{l,i}}\|^2}{\sigma^2}\right) \tag{3}$$

where σ^2 is the band width parameter. $s(\boldsymbol{B}, \boldsymbol{x^+_{l,i}})$ actually defines the similarity by choosing the closest instance in $\boldsymbol{B}$ to the positive instance $\boldsymbol{x^+_{l,i}}$. The idea is that the closest instance in bag $\boldsymbol{B}$ to the positive instance $\boldsymbol{x^+_{l,i}}$ carries the maximum amount of information, while the other instances in $\boldsymbol{B}$ might be irrelevant ones (backgrounds) to calculate the similarity. Intuitively, for example, assume $\boldsymbol{B}$ is an image from the target set 'sea' and $\boldsymbol{B_{l,i}}$ is an image from the source category 'sky'. The label $y_{l,i}$ should be transferred to image $\boldsymbol{B}$, if 'sea' and 'sky' are highly correlated and image $\boldsymbol{B}$ contains a region of 'sky' in it. In this case, it is very likely that the target concept 'sea' will also appear in image $\boldsymbol{B}$.

The major difference between our transfer function and those defined in previous transfer learning work (Qi et al. 2011; Wang et al. 2011) is that we transfer the knowledge from sources to target under multiple instance setting by choosing the positive instance inside each source bag to calculate the similarity, while previous methods treat the whole bag as one instance, which may introduce noisy knowledge propagation since only the positive instance of a bag contains the semantic concept.

Substituting Eqn.2 and Eqn.3 into Eqn.1, we have:

$$\begin{aligned} f_S(\boldsymbol{B}) &= \frac{1}{L}\sum_{l=1}^{L}\frac{1}{N_l}\sum_{i=1}^{N_l} y_{l,i} c_l \; s(\boldsymbol{B}, \boldsymbol{x^+_{l,i}}) \\ &= \sum_{l=1}^{L}\frac{\sum_{i=1}^{N_l} y_{l,i} \; s(\boldsymbol{B}, \boldsymbol{x^+_{l,i}})}{LN_l} c_l = \sum_{l=1}^{L} t_{B,l} c_l = \boldsymbol{c}^T \boldsymbol{t_B} \end{aligned} \tag{4}$$

here we use $t_{B,l}$ to denote $\frac{\sum_{i=1}^{N_l} y_{l,i} \; s(\boldsymbol{B}, \boldsymbol{x^+_{l,i}})}{LN_l}$. Note that $t_{B,l}$ can be pre-calculated and combined into a vector $\boldsymbol{t_B} = [t_{B,1}, t_{B,2}, \ldots, t_{B,L}]$. $\boldsymbol{c} = [c_1, c_2, \ldots, c_L]$ is the correlation vector.

[1] If $\boldsymbol{B_{l,i}}$ is a negative bag, then we use $\boldsymbol{x^+_{l,i}}$ to denote its most positive instance.

Weak Classifier

Traditional MIL methods build bag classifier solely based on the training examples from target category. In this paper, we build a weak classifier on the target category as:

$$f_T(\boldsymbol{B}) = \max_{\boldsymbol{x_j} \in \boldsymbol{B}} \boldsymbol{w}^T \boldsymbol{x_j} \tag{5}$$

here $\boldsymbol{w}$ is the linear classifier on instances and we assume that the bias has already been absorbed into feature vectors. The most positive instance in bag $\boldsymbol{B}$ is used to represent the bag, which is consistent with multiple instance setting. Note that we call this 'weak' classifier due to the situation of few training examples in target category. Although there are several alternatives to define the classifier for the target. The reason we choose this form is to reduce the complexity of the resulting optimization problem.

Adaptive Learning

Transfer learning sometimes has detrimental effects. When transfer learning is used inappropriately, it would harm the modeling performance of the target category. For example, a bag in the target category may contain instances with strong evidence indicating that this bag is obviously a positive bag, or the weak classifier in the target category may be well learned with sufficient training examples. In such cases, the weak classifier in the target category should be assigned more weight. On the other side, if there are very few training examples or the concept is too complex to be learned by the weak classifier, then more weight should be assigned to the transferred knowledge from source categories.

Therefore, to avoid negative effects in transfer, we propose a data-dependent mixture model that combines the transferred label knowledge from source categories with the weak classifier by adaptively adjusting their weights as follow:

$$f(\boldsymbol{B}) = (1 - \lambda_{\boldsymbol{B}}) f_S(\boldsymbol{B}) + \lambda_{\boldsymbol{B}} f_T(\boldsymbol{B}) \tag{6}$$

where $0 \leq \lambda_{\boldsymbol{B}} \leq 1$ is a data-dependent convex combination coefficient that balances the two terms. We propose to use a logistic function to model the combination coefficient as:

$$\lambda_{\boldsymbol{B}} = \frac{1}{1 + exp(-\boldsymbol{\theta}^T \boldsymbol{x_{j^*}})} \tag{7}$$

here $\boldsymbol{\theta}$ is the model parameter and $\boldsymbol{x_{j^*}} = \arg\max_{\boldsymbol{x_j} \in \boldsymbol{B}} \boldsymbol{w}^T \boldsymbol{x_j}$ is the most positive instance in bag $\boldsymbol{B}$. The data-dependent mixture model enables us to adaptively integrate the transferred knowledge with the weak classifier into a unified learning framework.

Objective and Optimization

Given the combined classifier $f(\boldsymbol{B})$ in Eqn.4, 5 and 6, we formulate the learning problem to minimize the following objective:

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{c},\boldsymbol{\theta},\xi} \alpha\|\boldsymbol{w}\|^2 + \beta\|\boldsymbol{c}\|^2 + \gamma\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^{N} \xi_i \\ &s.t.\ \forall i \in \{1, 2, \ldots, N\},\ \ \xi_i \geq 0 \\ &y_i \left((1 - \lambda_{\boldsymbol{B_i}}) \boldsymbol{c}^T \boldsymbol{t_{B_i}} + \lambda_{\boldsymbol{B_i}} \max_j \boldsymbol{w}^T \boldsymbol{x_{ij}} \right) \geq 1 - \xi_i \end{aligned} \tag{8}$$

here ξ_i is the classification loss on the i^{th} training bag in the target set T. We adopt hinge loss due to its superior performance in classification problem. α, β and γ are the trade-off parameters.

Directly minimizing Eqn.8 is intractable (Wang, Tao, and Di 2010), as many model parameters are coupled together and the formulation is a non-convex non-smooth optimization problem. An iterative coordinate descent method is employed to solve this problem. In particular, we optimize the objective function with respect to different parameters alternatively by the following two steps.

Step 1: Fix w and c, update θ. Given $\boldsymbol{w}$ and $\boldsymbol{c}$, the resulting optimization problem becomes:

$$\begin{aligned} &\min_{\boldsymbol{\theta},\xi} \gamma\|\boldsymbol{\theta}\|^2 + \sum_{i=1}^{N} \xi_i \\ &s.t.\ \forall i \in \{1, 2, \ldots, N\},\ \ \xi_i \geq 0 \\ &\frac{1}{1 + exp(-\boldsymbol{\theta}^T \boldsymbol{x_{ij^*}})} \geq a_i + b_i \xi_i \end{aligned} \tag{9}$$

where a_i and b_i are some constants which are computed using current $\boldsymbol{w}$ and $\boldsymbol{c}$. The above objective is still non-convex due to the logistic function in $\lambda_{\boldsymbol{B_i}}$. However, it is differentiable with respect to θ and thus can be solved efficiently using methods such as successive linear programming (Nocedal and Wright 2006).

Step 2: Fix θ, update w and c. Given θ, the objective function can be written as:

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{c},\xi} \alpha\|\boldsymbol{w}\|^2 + \beta\|\boldsymbol{c}\|^2 + \sum_{i=1}^{N} \xi_i \\ &s.t.\ \forall i \in \{1, 2, \ldots, N\},\ \ \xi_i \geq 0 \\ &y_i \left((1 - \lambda_i) \boldsymbol{c}^T \boldsymbol{t_{B_i}} + \lambda_i \max_j \boldsymbol{w}^T \boldsymbol{x_{ij}} \right) \geq 1 - \xi_i \end{aligned} \tag{10}$$

It is still a non-smooth optimization problem. But the form is less complicated than the problem in Eqn.8. There are multiple ways for solving the non-smooth optimization problems, such as Constrained Concave-Convex Procedure (CCCP) (Yuille and Rangarajan 2003) and bundle method (Bergeron et al. 2012). Due to the popularity of CCCP, we decompose this non-smooth problem into a series of smooth and convex sub-problems by CCCP. More precisely, CCCP iteratively computes $\boldsymbol{w^{(t)}}$ and $\boldsymbol{c^{(t)}}$ from $\boldsymbol{w^{(t-1)}}$ and $\boldsymbol{c^{(t-1)}}$ by replacing $\max_j \boldsymbol{w}^T \boldsymbol{x_{ij}}$ with its first order Taylor expansions at $\boldsymbol{w^{(t-1)}}$. For the t-th iteration of CCCP, the derived subproblem for solving problem in Eqn.10 is:

$$\begin{aligned} &\min_{\boldsymbol{w},\boldsymbol{c},\xi} \alpha\|\boldsymbol{w}\|^2 + \beta\|\boldsymbol{c}\|^2 + \sum_{i=1}^{N} \xi_i \\ &s.t.\ \forall i \in \{1, 2, \ldots, N\},\ \ \xi_i \geq 0 \\ &y_i \left((1 - \lambda_i) \boldsymbol{c}^T \boldsymbol{t_{B_i}} + \lambda_i \boldsymbol{w}^T \boldsymbol{x_{ij^*}} \right) \geq 1 - \xi_i \end{aligned} \tag{11}$$

where $j^* = \arg\max_j \boldsymbol{w^{(t-1)}}^T \boldsymbol{x_{ij}}$ represents the most positive instance in bag $\boldsymbol{B_i}$. The above subproblem is smooth and convex, which can be solved efficiently with a standard

Algorithm 1 Adaptive Knowledge Transfer Multiple Instance Learning (AKT-MIL)

Input: Target set $T = \{(\boldsymbol{B_i}, y_i)\}$, Sources sets $S_l = \{(\boldsymbol{B_{l,i}}, y_{l,i})\}$ and trade-off parameters.
Output: Category correlation $\boldsymbol{c}$, Adaptive transfer parameter $\boldsymbol{\theta}$ and Weak Classifier $\boldsymbol{w}$.

Initialize model parameters w and c.
Coordinate Descent
repeat
  Step 1: Update $\boldsymbol{\theta}$ by Eqn.9
  Step 2: Update $\boldsymbol{w}$ and $\boldsymbol{c}$ using CCCP, set $t = 0$
  repeat
    Replace $\max_j \boldsymbol{w}^T \boldsymbol{x_{ij}}$ with $\boldsymbol{x_{ij^*}}$ using
      j^*=arg $\max_j \boldsymbol{w^{(t-1)T}} \boldsymbol{x_{ij}}$
    Compute $\boldsymbol{w^{(t)}}$ and $\boldsymbol{c^{(t)}}$ by solving Eqn.11
    $t = t + 1$
  until CCCP converges
until Coordinate Descent converges

SVM. Through solving a series of subproblems derived from CCCP, the method is guaranteed to converge to a local optimal solution of problem in Eqn.10. The complete coordinate descent method together with CCCP for our AKT-MIL is shown in Algorithm 1.

Analysis

This section provides some analysis on the convergence of the learning algorithm. We first prove the convergence of the optimization algorithm. There are two loops in the optimization algorithm, an outer loop of coordinate descent method with an inner loop of CCCP to update $\boldsymbol{w}$ and $\boldsymbol{c}$. It has been shown that the coordinate descent method is guaranteed to converge to a local minimum. This is because the value of the objective function strictly decreases during the alternative updating of the parameters. For the CCCP iteration, if the selected instances j^* in the t-th iteration are the same as last iteration, then the algorithm will terminate since the optimization problem of Eqn.11 in t-th iteration is exactly the same as that in $(t-1)$-th iteration, which indicates that $\boldsymbol{w^{(t)}}$=$\boldsymbol{w^{(t-1)}}$ is the optimal solution. Otherwise, a different set of instances j^* are selected, which form a different problem and the optimal solution $\boldsymbol{w^{(t)}}$ will give a smaller objective value than $\boldsymbol{w^{(t-1)}}$. Therefore, the set of instances selected in each iteration are different until the optimization algorithm exists. Since there are only a finite number of possible sets of instances that can be selected at each iteration, the CCCP will terminate after a finite number of iterations. For the convergence speed, both successive linear programming for Eqn.9 and SVM for Eqn.11 converge very fast. The total computational cost of the learning algorithm depends on the number of iterations for coordinate descent and CCCP as well as the initial value. In our experiments, we have found that the coordinate descent method usually converges in less than 40 iterations and CCCP takes 20∼40 iterations to converge.

	*COREL*2000	*NUS-WIDE*
AKT-MIL	**0.782**	**0.746**
MILEAGE	0.774	0.702
MM-MIL	0.768	0.709
CCCL	0.744	0.703
DKT	0.707	0.671
MITL	0.593	0.606

Table 1: Average AUC results on two benchmarks by different algorithms.

Experimental Results

Datasets and Setting

The proposed AKT-MIL approach is evaluated with three configurations of experiments on two benchmark datasets, *COREL*2000 and *NUS-WIDE*. The *COREL*2000 (Chen, Bi, and Wang 2006) dataset includes 2000 images from 20 different categories, with 100 images in each category. This dataset contains various scenes and objects, e.g., '$building$', 'bus', '$elephant$' and '$tiger$'. The *NUS-WIDE* (Chua et al. 2009) dataset is created by *NUS* lab as a benchmark for evaluating image annotation and classification tasks. We use a subset of 6000 images from this benchmark. These 6000 images form 20 different categories, e.g., '$mountain$', '$beach$', 'sky' and 'sea', with 300 images in each category.

For MIL methods, each image is treated as a bag and segments of each image are instances. We extract a set of low-level features from each segment to represent an instance, including color histogram, color moment, region size, wavelet texture and shape (Wang, Si, and Zhang 2012). For transfer learning methods, the whole image is treated as one instance and the same set of features are extracted from the image. For each dataset, we randomly select 5 categories as the target categories, and the remaining 15 categories are used as the source categories. During each experiment, images in the target category are randomly partitioned into two halves to form the training and testing sets. Some parameters in our experiment are band width parameter σ^2, and the trade-off parameters α, β and γ. We use five-fold cross-validation for tuning the optimal values within the training set. We calculate the average result by repeating each experiment 10 times.

Evaluation on Different Algorithms

We first compare the proposed AKT-MIL approach with five different methods, including two multiple instance learning methods MM-MIL (Wang, Si, and Zhang 2012) and MILEAGE (Zhang et al. 2013) and three transfer learning methods CCCL (Qi et al. 2011), DKT (Wang et al. 2011) and MITL (Zhang and Si 2009) on both two benchmarks. For CCCL and MILEAGE, linear classifiers are used for fair comparison. For MM-MIL, the number of clusters is set to be 3, which shows good performance in their work. Various evaluation metrics can be used for comparing the performance. In our experiments, we use average area under the ROC curve (AUC) measure, which is a widely used metric in classification tasks.

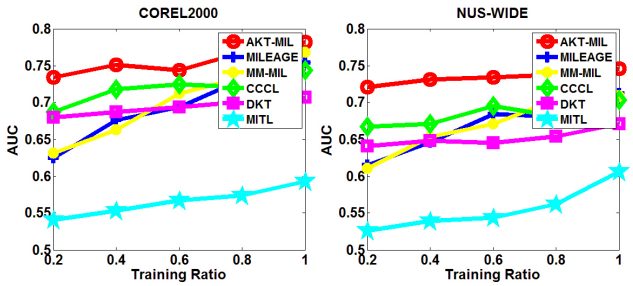

Figure 2: Average AUC results on two benchmarks by varying ration of training examples.

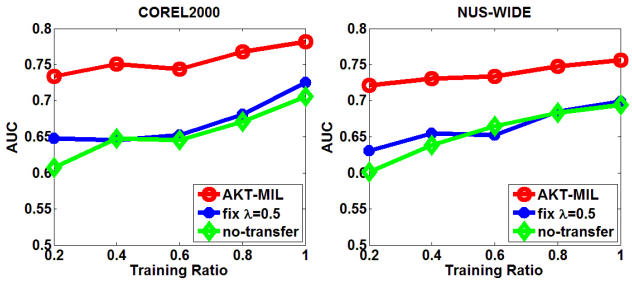

Figure 3: Average AUC results of different combination methods on two benchmarks.

The results of average AUC are reported in Table 1. It is clear that among all of these methods, AKT-MIL gives the best performance on both datasets. From the reported results, it can be seen that the proposed AKT-MIL substantially outperforms both compared MIL methods. Our hypothesis is that AKT-MIL benefits from the transferred knowledge in the source categories by exploring the semantic correlation among different categories, while traditional MIL methods do not leverage the knowledge contained in source categories. As seen in Table 1, our AKT-MIL also achieves higher AUC values than the transfer learning methods CCCL, DKT and MITL. The reason is that CCCL and DKT treat the whole image as one instance and do not take multiple instances into consideration, which could potentially introduce noisy knowledge transfer and thus limit their performance. Furthermore, MITL doesn't model the semantic correlation among different categories and has not considered the potential negative transfer effect.

Evaluation on Different Training Ratios

To evaluate the effectiveness of the proposed AKT-MIL approach with different number of training examples, we progressively increase the number of training examples in the target category by varying the training ratio from $\{0.2, 0.4, 0.6, 0.8, 1\}$ and compare our AKT-MIL approach with all the other baseline methods on both datasets described before. The results of average AUC are reported in Fig.2. From these results we can see that our AKT-MIL achieves the best performance among all compared methods on different training ratios. It can be observed from Fig.2 that the performance of transfer learning methods (i.e., CCCL, DKT, MITL and AKT-MIL) suffers less with small ratio of training data in target domain than the other non-transfer learning methods. The reason is that transfer learning methods leverage addition information contained in source categories and the cross-category knowledge could be considered as meaningful guidance for learning accurate classifier in the target domain, especially when there are very few training examples. However, our AKT-MIL consistently outperforms CCCL and DKT on different training ratios. We attribute this to the advantage of using multiple instance learning, since different instances in a bag carry different information and only one or few instances represent the semantic concept. Again, MITL does not perform well since it neither models the semantic correlation among different categories nor considers negative transfer effect.

Evaluation on Adaptive Learning

To evaluate the effectiveness of the proposed adaptive transfer scheme, we compare three different knowledge transfer strategies: (1) The proposed adaptive transfer learning algorithm with data-dependent knowledge transfer weight λ_B. (2) Fixing the knowledge transfer weight λ_B to 0.5, which means we treat the two classifiers from target and source domain equally. (3) Fixing the knowledge transfer weight λ_B to 1, which means we do not transfer any knowledge from source categories. The comparison results are shown in Fig.3, which demonstrates the advantage of the adaptive transfer scheme. As we can see in the figure, knowledge transfer algorithm with fixed weight sometime obtains even worse results than the algorithm without knowledge transfer. Our hypothesis is that the inappropriate knowledge transfer potentially hurts the performance of the transfer learning algorithm. Therefore, the adaptive knowledge transfer model is a critical component in our unified transfer learning framework.

Conclusion

This paper proposes a novel approach of Adaptive Knowledge Transfer for Multiple Instance Learning (AKT-MIL) in image classification. The new method leverages cross-category knowledge for improving the learning process in the target category by exploring the knowledge contained in the source categories in multiple instance setting. We design a unified learning framework with a data-dependent mixture model, which adaptively combines the transferred knowledge from sources with the weak classifier built in the target domain. An iterative coordinate descent scheme together with a Constrained Concave-Convex Procedure (CCCP) is proposed as the optimization method. Experimental results on two datasets demonstrate the advantage of the proposed AKT-MIL approach against several state-of-the-art multiple instance learning and transfer learning methods. In future, we plan to develop theoretical analysis of the convergence rate of the proposed learning algorithm. We also plan to extend the current binary classification problem to the multi-label case.

Acknowledgments

This work is partially supported by NSF research grants IIS-0746830, DRL-0822296, CNS-1012208, IIS-1017837 and CNS-1314688. This work is also partially supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370.

References

Andrews, S.; Tsochantaridis, I.; and Hofmann, T. 2002. Support vector machines for multiple-instance learning. In *NIPS*, 561–568.

Bergeron, C.; Moore, G. M.; Zaretzki, J.; Breneman, C. M.; and Bennett, K. P. 2012. Fast bundle algorithm for multiple-instance learning. *IEEE Trans. Pattern Anal. Mach. Intell.* 34(6):1068–1079.

Bunescu, R. C., and Mooney, R. J. 2007. Multiple instance learning for sparse positive bags. In *ICML*, 105–112.

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

Chen, Y.; Bi, J.; and Wang, J. Z. 2006. Miles: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(12):1931–1947.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *CIVR*.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89(1-2):31–71.

Fu, Z., and Robles-Kelly, A. 2009. An instance selection approach to multiple instance learning. In *CVPR*, 911–918.

Hu, Y.; Li, M.; and Yu, N. 2008. Multiple-instance ranking: Learning to rank images for image retrieval. In *CVPR*.

Kuzborskij, I.; Orabona, F.; and Caputo, B. 2013. From n to n+1: Multiclass transfer incremental learning. In *CVPR*, 3358–3365.

Maron, O., and Lozano-Pérez, T. 1997. A framework for multiple-instance learning. In *NIPS*.

Maron, O., and Ratan, A. L. 1998. Multiple-instance learning for natural scene classification. In *ICML*, 341–349.

Nocedal, J., and Wright, S. J. 2006. *Numerical Optimization*. New York: Springer, 2nd edition.

Pan, S. J., and Yang, Q. 2010. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22(10):1345–1359.

Qi, G.-J.; Aggarwal, C. C.; Rui, Y.; Tian, Q.; Chang, S.; and Huang, T. S. 2011. Towards cross-category knowledge propagation for learning visual concepts. In *CVPR*, 897–904.

Ray, S., and Craven, M. 2005. Supervised versus multiple instance learning: an empirical comparison. In *ICML*, 697–704.

Raykar, V. C.; Krishnapuram, B.; Bi, J.; Dundar, M.; and Rao, R. B. 2008. Bayesian multiple instance learning: automatic feature selection and inductive transfer. In *ICML*, 808–815.

Tommasi, T.; Orabona, F.; and Caputo, B. 2010. Safety in numbers: Learning categories from few examples with multi model knowledge transfer. In *CVPR*, 3081–3088.

Viola, P. A.; Platt, J. C.; and Zhang, C. 2005. Multiple instance boosting for object detection. In *NIPS*.

Wang, H.; Nie, F.; Huang, H.; and Ding, C. H. Q. 2011. Dyadic transfer learning for cross-domain image classification. In *ICCV*, 551–556.

Wang, Q.; Si, L.; and Zhang, D. 2012. A discriminative data-dependent mixture-model approach for multiple instance learning in image classification. In *ECCV (4)*, 660–673.

Wang, Q.; Tao, L.; and Di, H. 2010. A globally optimal approach for 3d elastic motion estimation from stereo sequences. In *ECCV (4)*, 525–538.

Xue, X.; Zhang, W.; Zhang, J.; Wu, B.; Fan, J.; and Lu, Y. 2011. Correlative multi-label multi-instance image annotation. In *ICCV*, 651–658.

Yao, Y., and Doretto, G. 2010. Boosting for transfer learning with multiple sources. In *CVPR*, 1855–1862.

Yuille, A. L., and Rangarajan, A. 2003. The concave-convex procedure. *Neural Computation* 15(4):915–936.

Zha, Z.-J.; Hua, X.-S.; Mei, T.; Wang, J.; Qi, G.-J.; and Wang, Z. 2008. Joint multi-label multi-instance learning for image classification. In *CVPR*.

Zhang, Q., and Goldman, S. A. 2001. Em-dd: An improved multiple-instance learning technique. In *NIPS*, 1073–1080.

Zhang, D., and Si, L. 2009. Multiple instance transfer learning. In *ICDM Workshops*, 406–411.

Zhang, D.; Liu, Y.; Si, L.; Zhang, J.; and Lawrence, R. D. 2011. Multiple instance learning on structured data. In *NIPS*, 145–153.

Zhang, D.; He, J.; Si, L.; and Lawrence, R. 2013. Mileage: Multiple instance learning with global embedding. In *ICML*.

Zhang, D.; He, J.; and Lawrence, R. D. 2013. M2ls: Multi-instance learning from multiple information sources. In *KDD*.

Zhou, Z.-H.; Sun, Y.-Y.; and Li, Y.-F. 2009. Multi-instance learning by treating instances as non-i.i.d. samples. In *ICML*, 157.