Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)

Active Learning with Cross-Class Knowledge Transfer

Yuchen Guo, Guiguang Ding, Yuqi Wang, and **Xiaoming Jin**
School of Software, Tsinghua University, Beijing 100084, China
{yuchen.w.guo,wangyuqi10}@gmail.com, {dinggg,xmjin}@tsinghua.edu.cn,

Abstract

When there are insufficient labeled samples for training a supervised model, we can adopt **active learning** to select the most informative samples for human labeling, or **transfer learning** to transfer knowledge from related labeled data source. Combining transfer learning with active learning has attracted much research interest in recent years. Most existing works follow the setting where the class labels in source domain are the same as the ones in target domain. In this paper, we focus on a more challenging **cross-class** setting where the class labels are totally different in two domains but related to each other in an intermediary attribute space, which is barely investigated before. We propose a novel and effective method that utilizes the attribute representation as the seed parameters to generate the classification models for classes. And we propose a joint learning framework that takes into account the knowledge from the related classes in source domain, and the information in the target domain. Besides, it is simple to perform uncertainty sampling, a fundamental technique for active learning, based on the framework. We conduct experiments on three benchmark datasets and the results demonstrate the efficacy of the proposed method.

Introduction

Sometimes, it is expensive and exhaustive to label sufficient samples for training a classifier. For example, training object classifiers for natural images from thousands of categories may require millions of well-labeled samples (Lampert, Nickisch, and Harmeling 2014). It is expected that we can use as few labeled samples as possible to train a classifier which can achieve satisfactory performance. The researchers have exploited two lines to achieve this goal. The first line is **active learning** (Settles 2009). The basic idea of active learning is that training samples have different information, and if the learning algorithm can, and is allowed to, select the most informative samples to label, even a few labeled samples can lead to an effective classifier. Existing works have demonstrated that active learning can significantly reduce the human labeling efforts (Joshi, Porikli, and Papanikolopoulos 2009; Zhuang et al. 2012). The second line is **transfer learning** (Pan and Yang 2010). In transfer learning, an auxiliary domain is available which is always fully labeled. It is related but not the same as the target domain. So we can transfer the supervised information in auxiliary domain into the target domain. By alleviating the influence of marginal and conditional distribution difference between domains, it is demonstrated that we can train accurate classifiers without labeled target data (Long et al. 2014).

The power of active learning and transfer learning motivates researchers to develop methods to combine them for better performance. By simultaneously transferring knowledge from auxiliary labeled domain and selecting the most informative samples from target domain to label, a.k.a., transfer active learning, some promising results have been achieved (Shi, Fan, and Ren 2008; Li et al. 2012; 2013; Chattopadhyay et al. 2013; Zhao et al. 2013). However, existing works assume that the source domain and target domain must have the same classes, while little attention is paid to the more general and challenging case where the classes in source domain are related but different from target domain classes, i.e., cross-class transfer active learning. In fact, the cross-class problem is more practical in real-world scenario. For example, if the target domain contains hundreds of uncommon classes, such as "lophius litulon" and "euchoreutes naso", it is very difficult to collect auxiliary data for all classes exactly and not miss any class. On the other hand, collecting data for common classes, like "fish" and "mouse" is very easy. Therefore if the learning algorithm can transfer knowledge from the common classes in source domain to the uncommon ones in target domain, the effort to collect the auxiliary data can be markedly reduced.

There are two key problems in cross-class transfer active learning. Firstly, because the source domain and target domain do not directly share any class, how to transfer knowledge between different classes? Secondly, how to measure the uncertainty of samples in this task so that the learning algorithm can select samples for labeling? The recent development of attribute-based zero-shot learning (Farhadi et al. 2009; Socher et al. 2013; Norouzi et al. 2013; Lampert, Nickisch, and Harmeling 2014) shows that by building an intermediary layer shared between source domain and target domain, i.e., attributes, the knowledge can be transferred across classes. Following this idea, in this paper we propose a novel usage of attributes. Instead of treating attributes as the intermediate during classification, we regard attributes as the seed parameters to build the classifier. Specifically, we

assume there is a generating function shared among classes which takes the attributes of a class as input and generates the classifier for this class. By using attributes in this way, the knowledge can be transferred from source domain to target domain with the generating function, and explicit classification models are constructed such that we can measure the uncertainty of a sample by the outputs of classifiers as in conventional active learning. Thus the two problems are addressed by the proposed method. Besides, we propose a joint optimization framework which simultaneously takes the information from source domain and target domain into consideration for better performance. In this paper, we make three important contributions listed as follows.

- We study a challenging and practical problem, cross-class transfer active learning where the source and target domains have related but totally different classes. To our best knowledge, our work is the first attempt to enhance active learning by transferring knowledge from different classes.
- We propose a novel method for cross-class transfer active learning. We utilize the attributes shared between source and target domains as the seed parameters to generate the classifiers. Based on this method, the knowledge can be transferred between domains. Besides, the uncertainty of a sample can be easily measured by using these classifiers.
- We carry out extensive experiments on three benchmark datasets. The experimental results demonstrate that the proposed method can significantly reduce the labeling efforts in comparison to traditional active learning methods.

Related Work

Transfer Active Learning

Transfer active learning is a combination of transfer learning and active learning. It simultaneously transfers knowledge from related source domain that are fully labeled and selects the most informative unlabeled samples in target domain for human labeling. By transferring knowledge from source domain, the labeling efforts in target domain can be reduced. In (Shi, Fan, and Ren 2008), the knowledge transferred from source domain is used as often as possible and the human labeling is triggered only when necessary. The likelihood that a sample in target domain can be correctly classified is estimated using the knowledge from source domain. The human labeling is requested when the likelihood for all unlabeled samples is low. In (Li et al. 2012), the shared common latent space between domains is learned from data. The active learning is performed in the latent space such that the knowledge from source domain can be utilized. In (Chattopadhyay et al. 2013), an integrated framework that solves a convex optimization problem is proposed. This framework simultaneously re-weights the source domain samples and selects the target domain samples to minimize a common objective of reducing distribution difference between domains. In (Li et al. 2013), a disjointed learning framework is proposed. Two individual classifiers are learned on source and target domains respectively. The prediction is made based on the decisions from both classifiers and the Query by Committee is adopted as selection strategy. In (Zhao et al. 2013), transfer active learning is utilized for recommendation system by actively identifying entity-correspondences across systems.

Existing works have demonstrated the effectiveness of transfer active learning. However, they all make a strong assumption that the source domain and target domain have the same classes. In real-world applications, it is expected that we can transfer knowledge across classes to reduce labeling efforts. However, existing works fail to handle this problem.

Attribute-based Zero-shot Learning

Zero-shot learning is to construct models for classes without any labeled data. To achieve this goal, the knowledge from some other related classes that are fully labeled is utilized. Specifically, some attributes shared between classes are used as the bridge for knowledge transfer, which is called attribute-based zero-shot learning (Farhadi et al. 2009; Yu et al. 2013; Socher et al. 2013; Norouzi et al. 2013; Lampert, Nickisch, and Harmeling 2014). Take animal classification as an example. We can define some attributes, such as "black", "stripes" and "water". Then the attribute representation for each class (both source and target domain) can be constructed by considering the relationship between class and attributes. By using the labeled data in source domain, the attribute classifier can be trained for each attribute. Because the attributes are shared between classes, the attribute classifiers trained in source domain also work in target domain. For example, a classifier for attribute "stripes" trained with "zebra" and "bear" can also handle images from "tiger" and "bird". For target domain, we can use the attribute classifiers to generate the attribute representation for each image. Finally, the class label is predicted by comparing the similarity between the attribute representation of the test image and all classes in target domain. For example, if a test image is classified as having attributes "stripes", "four legs", "furry", and not "water", it is more likely to be a tiger than a bird or fish. Therefore, by taking advantage of attributes as the intermediary level, the knowledge can be effectively transferred between different classes. Furthermore, with the recent development of word vector (Turney and Pantel 2010; Huang et al. 2012; Mikolov et al. 2013), it is very easy to obtain the intermediary attribute representation for each class.

Attribute-based methods adopt the two-step strategy to transfer knowledge via attributes. However, the two-step strategy may leads to information loss, and the classification performance highly relies on the attribute classifiers that are unreliable in some cases (Jayaraman and Grauman 2014). Besides, it is unclear how to measure the uncertainty based on existing attribute based methods. Hence, how to combine attribute with active learning is an unexplored research issue.

The Proposed Method

Problem Definition and Notation

In this paper, we consider the cross-class transfer active learning problem. We have a set of fully labeled samples from source domain $\mathcal{D}_s = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_s}$, where $\mathbf{x}_i \in \mathbb{R}^d$ is the feature and $\mathbf{y}_i \in \{0,1\}^{c_s}$ is the corresponding label vector. The source domain samples belong to c_s classes

Table 1: Notations and descriptions in this paper.

Notation	Description	Notation	Description
$\mathbf{X}_s, \mathbf{X}_t$	features	n_s, n_t	#samples
$\mathbf{Y}_s, \mathbf{Y}_t$	label matrix	d	#dimension
$\mathbf{A}_s, \mathbf{A}_t$	attribute matrix	m	#attributes
$\mathbf{W}_s, \mathbf{W}_t$	classifiers	c_s, c_t	#classes
$\mathbf{\Theta}_s, \mathbf{\Theta}_t$	weights	f, g	functions
$\mathbf{V}$	factors	α, β	parameters

$\mathcal{C}_s = \{C_j^s\}_{j=1}^{c_s}$. We have $y_{ij} = 1$ if image i belongs to class C_j or 0 otherwise. We also have a set of unlabeled training samples from target domain $\mathcal{D}_t^{tr} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{tr}}$ where the label vector $\mathbf{y}_i$ is unknown unless this sample is selected for human labeling. And there are a set of test samples $\mathcal{D}_t^{te} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^{n_{te}}$ in target domain which are sampled from the same distribution as $\mathcal{D}_t^{tr}$. The test set $\mathcal{D}_t^{te}$ is not available for training. The target domain samples belong to c_t classes $\mathcal{C}_t = \{C_j^t\}_{j=1}^{c_t}$. Our goal is to select as few samples from $\mathcal{D}_t^{tr}$ as possible for human labeling to train a classifier that can achieve satisfactory performance on $\mathcal{D}_t^{te}$. Different from existing transfer active learning that assumes $\mathcal{C}_s = \mathcal{C}_t$, in this paper we consider a more general and challenging setting where $\mathcal{C}_s \cap \mathcal{C}_t = \emptyset$. Besides, for any class $c_j \in \mathcal{C}_s \cup \mathcal{C}_t$, we have the attribute representation $\mathbf{a}_j \in \mathbb{R}^m$ for it which describe the characteristics of the class. Besides, this paper focuses on the multi-class single-label classification problem, i.e., there is only one "1" in each $\mathbf{y}_i$ and the others are all "0". The notations are summarized in Table 1.

Cross-class Transfer Active Learning

Given a set of labeled training samples, the multi-class classifiers is learned by optimizing the objective function below,

$$\min_{f_c} \sum_{\mathbf{x}_i} \sum_{c} \ell(f_c(\mathbf{x}_i), y_{ic}) + \mathcal{R}(f_c) \tag{1}$$

In this paper, we consider the one-vs-the-rest classifier for multi-class classification because of its high classification efficiency and good performance (Fan et al. 2008). Here, f_c is the classifier for class c, $\ell(a, b)$ is the loss function, and $\mathcal{R}$ is the regularization term for classifier parameters. Besides, the prediction of the single-label classification is obtained by

$$c(\mathbf{x}) = \text{argmax}_c f_c(\mathbf{x}) \tag{2}$$

In active learning, the uncertainty of an unlabeled sample $\mathbf{x}$ is computed using the outputs of all classifiers, i.e., we have

$$u(\mathbf{x}) = u(f_1(\mathbf{x}), ..., f_c(\mathbf{x})) \tag{3}$$

Then the samples with the largest uncertainty are selected for human labeling. We can perform active learning only in the target domain. Besides, we can observe from Eq. (1) that the classifier for each class can be learned individually. However, the knowledge in source domain is wasted in both situations. Motivated by attribute-based zero-shot learning, we can utilize attribute as the bridge for knowledge transfer.

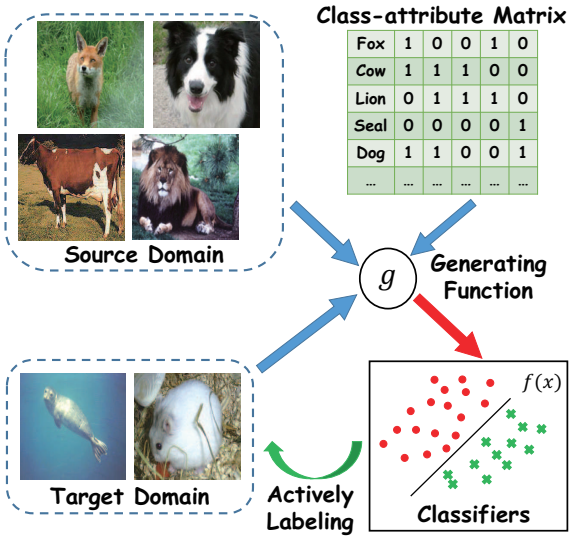


Figure 1: Cross-class transfer active learning.

In this paper, we consider the linear classifiers, i.e., we have $f_c(\mathbf{x}) = \mathbf{x}\mathbf{w}_c'$, where $\mathbf{w}_c \in \mathbb{R}^d$ is the classifier parameters. By Eq. (1), we can obtain the classifiers for source domain classes. However, because we lack the label information for target domain classes, we can not directly use Eq. (1) for training. In this paper, we propose to learn a generating function g that can generate the classifiers given some seed parameters. In fact, one can imagine that the classifier parameters of a class are determined by the properties of the class. Thus given a property description of a class, it is reasonable to assume that there is a transformation function that turns the properties into classifier parameters. Motivated by the success of attribute classification (Socher et al. 2013; Lampert, Nickisch, and Harmeling 2014), we can adopt attributes to characterize the properties of the classes. Therefore, the classifier parameters for class c can be constructed by the attributes and the generating function as $\mathbf{w}_c = g(\mathbf{a}_c)$.

Based on this idea, we illustrate our method in Figure 1. The key in our method is the generating function g. To learn it, we can use the labeled data in source domain $\mathcal{D}_s$ and the training data in target domain $\mathcal{D}_t^{tr}$. Then with the attributes of target domain classes, we can directly generate the classifiers for target domain, which is an important difference between our method and attribute-based learning methods. Because the classifiers are all available, it is straightforward to measure uncertainty of any samples in $\mathcal{D}_t^{tr}$ and perform selection and human labeling in active learning. We can iterate these steps until satisfactory performance is achieved. With the generating function and attributes as bridge, the knowledge from source domain classes can be transferred across classes into target domain classes and thus the labeling efforts in the active learning can be observably reduced.

Learning Generating Function In order to learn the generating function g, we can use the labeled data in $\mathcal{D}_s$ and $\mathcal{D}_t^{tr}$. Besides, based on the theory of semi-supervised learn-

ing (Zhu and Goldberg 2009), considering the unlabeled data in training set can lead to better classification performance (Li et al. 2013; Rohrbach, Ebert, and Schiele 2013). Thus, the learning objective for g can be presented as below,

$$\begin{aligned}&\min_{g,\mathbf{Y}_t} \sum_{\mathbf{x}_s\in\mathcal{D}_s}\sum_{c\in\mathcal{C}_s}\ell(\mathbf{x}_s(g(\mathbf{a}_c))', y^s_{ic})\\&+\alpha(\sum_{j\in\mathcal{L}}\sum_{c\in\mathcal{C}_t}\theta_j\ell(\mathbf{x}^t_j(g(\mathbf{a}_c))', y^t_{jc})\\&+\sum_{j\in\mathcal{U}}\sum_{c\in\mathcal{C}_t}\theta_j\ell(\mathbf{x}^t_j(g(\mathbf{a}_c))', y^t_{jc})) + \mathcal{R}(g(\mathbf{a}))\\&\text{s.t. } \mathbf{y}^t_j \text{ is fixed, } \forall j\in\mathcal{L};\ \|\mathbf{y}^t_j\|_0 = \mathbf{y}^t_j\mathbf{1}'_{c_t} = 1,\ \forall j\in\mathcal{U}\end{aligned} \tag{4}$$

where α is a hyper parameter, $\mathcal{L}$ is the labeled set in $\mathcal{D}^{tr}_t$ and $\mathcal{U}$ is the unlabeled set, θ_i is the weight for the sample $\mathbf{x}^t_i$, and $\|\cdot\|_0$ denotes the ℓ_0-norm of a vector. One may argue that we can learn g with only $\mathcal{D}_s$. However, learning g in this way only considers the information from source domain. Since our ultimate goal is to build classifiers in target domain, we also incorporate the information from the target domain, both labeled and unlabeled, into the objective function for g.

We formulate the learning objective to be general such that one can choose specific settings based on the specific requirement. In this paper, we adopt the linear function for g, i.e., $g(\mathbf{a}) = \mathbf{a}\mathbf{V}'$, where $\mathbf{V}\in\mathbb{R}^{d\times m}$ is the factor for the generating function. Although the linear function seems quite simple, we find out that it works quite well. We leave the other forms for g to our future research. In addition, we use the squared loss for ℓ and the ridge regularization for $\mathcal{R}$. Now we can write the specific objective function as follows,

$$\begin{aligned}&\min_{\mathbf{V},\mathbf{Y}_t}\|\mathbf{X}_s\mathbf{V}\mathbf{A}'_s - \mathbf{Y}_s\|^2_F + \alpha\|\hat{\mathbf{X}}_t\mathbf{V}\mathbf{A}'_t - \hat{\mathbf{Y}}_t\|^2_F + \beta\|\mathbf{V}\mathbf{A}'\|^2_F\\&\text{s.t. } \mathbf{y}^t_j \text{ is fixed, } \forall j\in\mathcal{L};\ \|\mathbf{y}^t_j\|_0 = \mathbf{y}^t_j\mathbf{1}'_{c_t} = 1,\ \forall j\in\mathcal{U}\end{aligned} \tag{5}$$

where $\hat{\mathbf{X}}_t = \mathbf{\Theta}_t\mathbf{X}_t$ and $\hat{\mathbf{Y}}_t = \mathbf{\Theta}_t\mathbf{Y}_t$ are the re-weighted samples, $\mathbf{\Theta}_t = \text{diag}(\theta_1^{\frac{1}{2}}, ..., \theta_{n_t}^{\frac{1}{2}})$ represents the weighting matrix, $\|\cdot\|_F$ denotes the Frobenius norm of matrix, and β is the hyper parameter to control model complexity. We can see that $\mathbf{V}$ appears in both source and target domains. Thus it can bridge domains and transfer knowledge across classes.

To solve Eq. (5) that has two matrix variables, we can adopt an iterative strategy which fixes one variable when optimizing the other. Specifically, when $\mathbf{Y}_t$ is fixed, we simplify and *approximate* the objective function w.r.t. $\mathbf{V}$ as below,

$$\min_{\mathbf{V}}\ \mathcal{O}_{\mathbf{V}} = \|\mathbf{X}\mathbf{V}\mathbf{A}' - \mathbf{Y}\|^2_F + \beta\|\mathbf{V}\mathbf{A}'\|^2_F \tag{6}$$

The notations in the above formulation are defined as below,

$$\mathbf{X} = [\mathbf{X}_s; \sqrt{\alpha}\hat{\mathbf{X}}_t], \mathbf{Y} = \begin{bmatrix}\mathbf{Y}_s & \mathbf{0}_{n_s\times c_t}\\ \mathbf{0}_{n_t\times c_s} & \sqrt{\alpha}\hat{\mathbf{Y}}_t\end{bmatrix}, \mathbf{A} = [\mathbf{A}_s; \mathbf{A}_t]$$

Then the derivative of $\mathcal{O}_{\mathbf{V}}$ w.r.t. $\mathbf{V}$ is calculated as follows,

$$\frac{\partial\mathcal{O}_{\mathbf{V}}}{\partial\mathbf{V}} = 2\mathbf{X}'\mathbf{X}\mathbf{V}\mathbf{A}'\mathbf{A} - 2\mathbf{X}'\mathbf{Y}\mathbf{A} + 2\beta\mathbf{V}\mathbf{A}'\mathbf{A} \tag{7}$$

By setting the derivative to 0, we obtain the solution for $\mathbf{V}$,

$$\mathbf{V}\leftarrow(\mathbf{X}'\mathbf{X} + \beta\mathbf{I}_d)^{-1}\mathbf{X}'\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \tag{8}$$

On the other hand, if we keep $\mathbf{V}$ fixed, we can observe that the optimization problem is row-wise decoupled w.r.t. $\mathbf{y}^t_j$ ($\forall j\in\mathcal{U}$), and each subproblem can be written as follows,

$$\min_{\mathbf{y}^t_j}\ \|\mathbf{x}^t_j\mathbf{V}\mathbf{A}'_t - \mathbf{y}^t_j\|^2_F \text{ s.t. } \|\mathbf{y}^t_j\|_0 = \mathbf{y}^t_j\mathbf{1}'_{c_t} = 1 \tag{9}$$

Solving the above problem leads to the updating rule below,

$$y^t_{jc} = \begin{cases}1, & \text{if } c = \text{argmax}_c\mathbf{x}^t_j\mathbf{V}\mathbf{a}^{t'}_c\\ 0, & \text{otherwise}\end{cases} \tag{10}$$

which is the explicit formulation for Eq. (2). Because we have no label information for unlabeled data, the predicted labels by Eq. (10) are the **pseudo** labels. In fact, more accurate pseudo labels can lead to better $\mathbf{V}$ and vice versa. Thus we can iteratively update $\mathbf{V}$ and $\mathbf{Y}_t$ to gradually refine the pseudo labels and models until convergence (Long et al. 2013). Then we obtain the classifiers which contain knowledge from the source domain classes for the target domain classes by using the generating function g and the attributes.

Uncertainty Sampling Based on the classifiers obtained above, we can perform uncertainty sampling. In this paper, we adopt the Best-vs-Second Best strategy for multi-class uncertainty sampling (Joshi, Porikli, and Papanikolopoulos 2009). Specifically, based on Eq. (10), for sample $\mathbf{x}^t_j$, the output of classifier f_c is $o_{jc} = \mathbf{x}^t_j\mathbf{V}\mathbf{a}'_c$ which is similar to the distance between $\mathbf{x}^t_j$ and the hyperplane $\mathbf{V}\mathbf{a}'_c$ (Tong and Koller 2001). Suppose o_{jc_1} and o_{jc_2} are the largest and second largest outputs. Because the outputs may be negative, we can not use entropy to measure the uncertainty. But we can observe that if there is large difference between o_{jc_1} and o_{jc_2}, the sample is classified as c_1 with high confidence, i.e., it has less uncertainty. On the other hand, if the difference is very small, the sample also has high probability to be c_2 even though it is classified as c_1, i.e., we are uncertain about the classification. Thus, it is reasonable to use the difference between o_{jc_1} and o_{jc_2} to measure the uncertainty as follows

$$u(\mathbf{x}^t_j) = e^{o_{jc_2} - o_{jc_1}} \tag{11}$$

Here we use the exponential function to make the value positive. Then we can select samples with the largest uncertainty for human labeling. Besides, since we focus on the multi-class problem, it is not expected that the selected samples in one iteration belong to the same class, which may result in redundancy. In this paper, we propose to perform the class-wise sampling. Specifically, for each class c, we only consider the samples classified as c by Eq. (10), i.e., $y^t_{jc} = 1$. And we use Eq. (11) to select k samples for human labeling. Therefore in each iteration $c_t k$ samples are selected in total.

Other Issues To learn $\mathbf{V}$ by Eq. (8), we need to know the pseudo labels $\mathbf{Y}_t$ that are generated by Eq. (10) using $\mathbf{V}$, which is a "chicken or the egg" dilemma. To initialize $\mathbf{Y}_t$, we can adopt any existing zero-shot learning method (Lampert, Nickisch, and Harmeling 2014; Jayaraman and Grauman 2014). In this paper, we initialize $\mathbf{V}$ by using the source

Algorithm 1 Cross-class Transfer Active Learning

Input: Source samples $\mathbf{X}_s$; Source labels $\mathbf{Y}_s$;
Target samples $\mathbf{X}_t$; Parameters α and β;
Source attributes $\mathbf{A}_s$; Target attributes $\mathbf{A}_t$;
#iterations T; #selected samples for each class k;
Output: Classifiers $\mathbf{w}_c$ for target domain, $c = 1, ..., c_t$;

Initialize $\mathbf{V}$ using source domain samples;
Initialize $\mathbf{Y}_t$ by Eq. (10);
Initialize $\mathcal{U} = \{1, ..., n_t\}$, $\mathcal{L} = \emptyset$;
for $iter = 1 : T$ **do**
  for $c = 1 : c_t$ **do**
    Select samples with $y_{jc}^t = 1$, $j \in \mathcal{U}$;
    Calculate $u(\mathbf{x}_j^t)$ by Eq. (11);
    Select $\mathcal{S} = \{j_l\}_{l=1}^k$ with the largest uncertainty;
    $\mathcal{L} = \mathcal{L} \cup \mathcal{S}, \mathcal{U} = \mathcal{U} \backslash \mathcal{S}$; // Actively labeling
  end for
  Update $\mathbf{\Theta}$ by Eq. (12);
  repeat
    Update $\mathbf{V}$ by Eq. (8); // Knowledge transfer
    Update y_{jc}^t by Eq. (10), $j \in \mathcal{U}$;
  until Convergence;
end for
Return $\mathbf{w}_c = \mathbf{a}_c \mathbf{V}'$, $c = 1, ..., c_t$;

domain samples and then generate the initial $\mathbf{Y}_t$ by Eq. (10). Then we can iteratively refine them as we introduced above.

Another issue is the weights for all samples in target domain. A simple solution is to set all weights to 1. But this strategy may assign too much weight to very uncertain samples, which may degrade the performance. Thus we should set the weight based on the uncertainty of a sample, such as

$$\theta_j = \frac{1}{1 + \delta u(\mathbf{x}_j^t)} \tag{12}$$

where δ is a scale factor and we set $\delta = 1/\text{median}(\mathbf{\Theta}_t)$ such that it is updated in each iteration. Based on this definition, the weights for certain samples are close to 1 while for uncertain samples are close to 0. Besides, we set $\theta_j = 1$ for labeled samples in target domain. We summarize our method for the cross-class transfer active learning into Algorithm 1.

Experiment

Settings

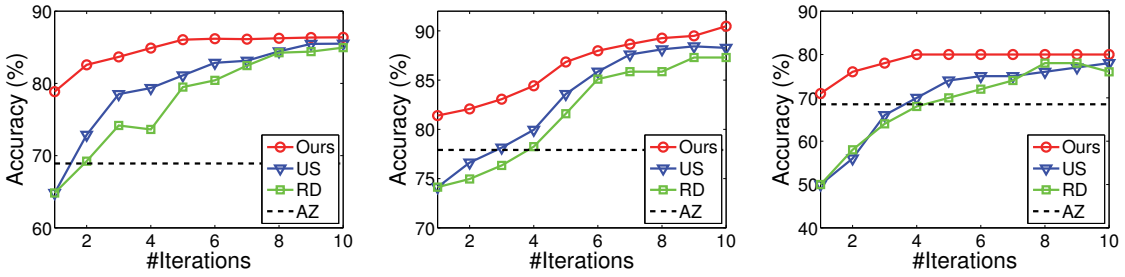
Datasets To demonstrate the effectiveness of the proposed method, we carry out experiments on three benchmark datasets with attributes. The first dataset is Animal with Attributes (AwA) (Lampert, Nickisch, and Harmeling 2014). This dataset has $30,475$ images belonging to 50 animal categories, such as "dog", "dolphin", "bear", and so on. For each class, an 85-dimensional attribute representation is given which contains "brown", "water", and etc. This dataset provides a standard source/target split where 40 classes with $24,295$ samples are in source domain and 10 classes with $6,180$ samples are in target domain. For this dataset, each image is represented by a $4,096$-dimensional deep features extracted by DeCAF (Donahue et al. 2014) without fine-tuning. The second dataset is aPascal-aYahoo (aPY) (Farhadi et al. 2009). This dataset contains two subsets. The first subset is aPascal from PASCAL VOC2008 challenge that has $12,695$ samples from 20 different categories like "people" and "dog". The second subset is aYahoo which is collected from Yahoo image search. aYahoo has 12 categories with $2,644$ images that are similar but different from the categories in aPascal, such as "centaur" and "wolf". In aPY, we follow the standard setting where aPascal works as the source domain and aYahoo is the target domain. In this dataset, each image is annotated by 64 binary attributes, such as "furry" and "pot". We average the attribute representations of images in the same category to generate the class attribute representation. We also use DeCAF to extract a $4,096$-dimensional deep features for each image. The third dataset is the SUN fine-grained scene recognition dataset (Patterson and Hays 2012). This dataset contains 717 different scenes such as "airport", "palace", and etc. There are totally $14,340$ images in this dataset and each class has 20 images. Following the source/target split in (Jayaraman and Grauman 2014), 707 classes form the source domain and the other 10 classes form the target domain. For each image, 102-dimensional binary attributes annotated by human are given which includes "natural", "open", and etc. We average the images' attributes to obtain the attributes for a class. For this dataset, we utilize the author-provided $17,032$-dimensional features for each image which include HOG, color histograms, self similarity, and so on. The statistics of three benchmark datasets are summarized in Table 2.

Table 2: The statistics of datasets.

	AwA	aPY	SUN
#source class	40	20	707
#source sample	24,295	12,695	14,140
#target class	10	12	10
#target sample	6,180	2,644	200
#attribute	85	64	102
#dimension	4,096	4,096	17,032

Following the settings in active learning (Chattopadhyay et al. 2013), we equally split the target domain samples into two parts. We use one part to train classifiers with active learning, i.e., $\mathcal{D}_t^{tr}$. And the other part is the unseen test set, i.e., $\mathcal{D}_t^{te}$. The labeled source domain samples form $\mathcal{D}_s$. All experiments mentioned below share the same train/test split.

Baselines As there is no work for cross-class transfer active learning before, we compare our method to two classical methods. The first one is random sampling (RD) which selects samples randomly from unlabeled data for human labeling. The second is uncertainty sampling (US) (Joshi, Porikli, and Papanikolopoulos 2009) which selects the samples that the current classifiers are most uncertain about. For these two methods, we use Liblinear SVM (Fan et al. 2008) as the base classifier. Besides, we also select a state-of-the-art attribute-based zero-shot learning method (AZ) (Fu et al. 2014) as baseline. In fact, because it adopts the two-step



(a) AwA (b) aPY (c) SUN

Figure 2: Classification accuracy w.r.t. the number of iterations.

strategy, it is unclear how to combine it with active learning. Therefore we just report the zero-shot learning result of AZ.

Implementations For all active learning methods, we select 10, 12, and 10 unlabeled samples for human labeling for AwA, aPY, and SUN datasets respectively in each iteration, i.e., each class has one sample in average. In each iteration, we will perform sampling, labeling and retraining models. The performance is evaluated by the classification accuracy on the unseen test data $\mathcal{D}_t^{te}$ after retraining in each iteration.

There are some parameters to determine for different methods. For RD and US, we need to determine the parameter C for SVM classifier. Here, we use the labeled source domain samples to perform cross-validation to choose a proper value for C, and the candidate set is $\{0.01, 0.1, 1, 10, 100\}$. For our method, we need to determine the hyper parameters α and β. In this paper we propose to perform k-fold cross-validation as follows. We split equally the source domain classes into k parts. In each fold, we use 1 part as the target domain and the other $k-1$ parts as the source domain. Because they are fully labeled, we can simulate the test procedure and evaluate our method under different parameter settings. Then the best setting is utilized for final test. Specifically, we set $k = 4, 4$, and 10 for AwA, aPY and SUN respectively, and the values of α and β are chosen from $\{0.01, 0.1, 1, 10, 100\}$. After the cross-validation, we use the best model for initialization (line 1) in Algorithm 1.

In RD and US, we need to train the multi-class SVM classifier, where at least one sample for each class is required. To ensure this, the initial set (in the first iteration) will be randomly re-generated unless there is one sample for each class. However, this is too demanding when there are a lot of classes. Fortunately, our method can avoid this problem, because the initial model is trained on the source domain and it does not need any labeled samples for the target domain classes.

Results

The classification accuracy curves w.r.t. the number of iterations on three datasets of different methods are plotted in Figure 2. We can observe that our method significantly outperforms the baseline methods and can achieve satisfactory performance with very few labeled data in target domain, and we have some important observations from the results.

Table 3: The #iterations to achieve 80% accuracy.

#Iterations	AwA	aPY	SUN
RD	6	5	> 10
US	5	4	> 10
Ours	2	1	4

Firstly, the cross-class transfer learning methods (AZ and ours) outperform traditional active learning methods (RD and US) when the labeled samples in target domain are extremely insufficient, e.g., fewer than 20. This result indicates that using knowledge from different but related classes helps to train accurate classifiers with insufficient labeled samples.

Secondly, we can observe that US and our method perform better than RD and AZ when we increase the number of labeled samples. This result demonstrates that the active learning can lead to satisfactory performance with just a few labeled samples via uncertainty sampling. Because of the two-step strategy in AZ, it is difficult to combine it with active learning. Hence the attribute-based zero-shot learning methods fail to make use of the advantage of active learning.

Thirdly, our method is much superior to US and RD with very few labeled samples and performs better with more labeled data, which demonstrates that it can not only effectively transfer knowledge from other classes but also benefit from active learning. Therefore, our method can significantly reduce the labeling efforts. Besides, in Table 3, we present the number of iterations each method needs to achieve 80% accuracy for multi-class classification. We can observe that our method can save 60%, 75%, and 60% labeling efforts on three datasets respectively, which validates its effectiveness.

Conclusion

In this paper, we investigate a challenging problem, cross-class transfer active learning. We propose a novel method that utilizes the attribute representation as the seed parameters to directly generate the classifier parameters for a class via a generating function. Based on the classifier, we can

perform uncertainty sampling for active learning. A joint learning algorithm is proposed to take both source domain and target domain, cross-class knowledge transfer and active learning into account. We carried out experiments on three datasets. The results show that our method can significantly outperform baselines and markedly reduce labeling efforts.

Acknowledgements

This research was supported by the National Natural Science Foundation of China (Grant No.61271394 and 61571269) and the National Basic Research Project of China (Grant No. 2015CB352300). At last, the authors would like to sincerely thank the reviewers for their valuable comments and advice.

References

Chattopadhyay, R.; Fan, W.; Davidson, I.; Panchanathan, S.; and Ye, J. 2013. Joint transfer and batch-mode active learning. In *Proceedings of the 30th International Conference on Machine Learning*, 253–261.

Donahue, J.; Jia, Y.; Vinyals, O.; Hoffman, J.; Zhang, N.; Tzeng, E.; and Darrell, T. 2014. Decaf: A deep convolutional activation feature for generic visual recognition. In *Proceedings of the 31th International Conference on Machine Learning*, 647–655.

Fan, R.; Chang, K.; Hsieh, C.; Wang, X.; and Lin, C. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research* 9:1871–1874.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1778–1785.

Fu, Y.; Hospedales, T. M.; Xiang, T.; Fu, Z.; and Gong, S. 2014. Transductive multi-view embedding for zero-shot recognition and annotation. In *Computer Vision - ECCV 2014 - 13th European Conference*, 584–599.

Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *The 50th Annual Meeting of the Association for Computational Linguistics*, 873–882.

Jayaraman, D., and Grauman, K. 2014. Zero-shot recognition with unreliable attributes. In *Annual Conference on Neural Information Processing Systems 2014*, 3464–3472.

Joshi, A. J.; Porikli, F.; and Papanikolopoulos, N. 2009. Multi-class active learning for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2372–2379.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(3):453–465.

Li, L.; Jin, X.; Pan, S. J.; and Sun, J. 2012. Multi-domain active learning for text classification. In *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1086–1094.

Li, S.; Xue, Y.; Wang, Z.; and Zhou, G. 2013. Active learning for cross-domain sentiment classification. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence*.

Long, M.; Wang, J.; Ding, G.; Sun, J.; and Yu, P. S. 2013. Transfer feature learning with joint distribution adaptation. In *IEEE International Conference on Computer Vision*, 2200–2207.

Long, M.; Wang, J.; Ding, G.; Pan, S. J.; and Yu, P. S. 2014. Adaptation regularization: A general framework for transfer learning. *IEEE Trans. Knowl. Data Eng.* 26(5):1076–1089.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *CoRR* abs/1312.5650.

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

Patterson, G., and Hays, J. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 2751–2758.

Rohrbach, M.; Ebert, S.; and Schiele, B. 2013. Transfer learning in a transductive setting. In *27th Annual Conference on Neural Information Processing Systems 2013*, 46–54.

Settles, B. 2009. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison.

Shi, X.; Fan, W.; and Ren, J. 2008. Actively transfer domain knowledge. In *Machine Learning and Knowledge Discovery in Databases, European Conference*, 342–357.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *27th Annual Conference on Neural Information Processing Systems 2013*, 935–943.

Tong, S., and Koller, D. 2001. Support vector machine active learning with applications to text classification. *Journal of Machine Learning Research* 2:45–66.

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *J. Artif. Intell. Res. (JAIR)* 37:141–188.

Yu, F. X.; Cao, L.; Feris, R. S.; Smith, J. R.; and Chang, S. 2013. Designing category-level attributes for discriminative visual recognition. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 771–778.

Zhao, L.; Pan, S. J.; Xiang, E. W.; Zhong, E.; Lu, Z.; and Yang, Q. 2013. Active transfer learning for cross-system recommendation. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*.

Zhu, X., and Goldberg, A. B. 2009. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Zhuang, H.; Tang, J.; Tang, W.; Lou, T.; Chin, A.; and Wang, X. 2012. Actively learning to infer social ties. *Data Min. Knowl. Discov.* 25(2):270–297.