Transductive Zero-shot Recognition via Shared Model Space Learning

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

Abstract

Zero-shot Recognition (ZSR) is to learn recognition models for novel classes without labeled data. It is a challenging task and has drawn considerable attention in recent years. The basic idea is to transfer knowledge from seen classes via the shared attributes. This paper focus on the transductive ZSR, i.e., we have unlabeled data for novel classes. Instead of learning models for seen and novel classes separately as in existing works, we put forward a novel joint learning approach which learns the *shared model space* (SMS) for models such that the knowledge can be effectively transferred between classes using the attributes. An effective algorithm is proposed for optimization. We conduct comprehensive experiments on three benchmark datasets for ZSR. The results demonstrates that the proposed SMS can significantly outperform the state-of-the-art related approaches which validates its efficacy for the ZSR task.

Introduction

Learning a recognition model, e.g., a SVM classifier that tells an object is a dog or not, always requires sufficient manually labeled data for the target class (Bishop and others 2006). However, with the explosive growth of visual data and the huge number of potential classes, such as the images and the annotated tags in Flickr, it is expensive and burdensome to collect well-labeled training data for new classes (Lampert, Nickisch, and Harmeling 2014). To tackle this problem, few-shot recognition (Mensink et al. 2013; Habibian, Mensink, and Snoek 2014), and the extreme situation zero-shot recognition (Palatucci et al. 2009; Lampert, Nickisch, and Harmeling 2009; Akata et al. 2013; Yu et al. 2013; Jayaraman and Grauman 2014; Fu et al. 2014b; Romera-Paredes and Torr 2015), have been proposed and attracted considerable interest from academia in recent years.

The goal of zero-shot recognition is to learn models with no labeled data for novel (target) class. We can notice that there is no labeled data for target class, but the recognition models are generally built in a supervised way. To address this critical problem for ZSR, an intermediary space which is shared among classes is utilized for transferring knowledge from seen (source) classes that have sufficient labeled data to target classes. Two widely used intermediates are attributes (Lampert, Nickisch, and Harmeling 2014) and label semantic representations (Socher et al. 2013). The basic framework of attribute-based[1] ZSR is illustrated in Figure 1. Each class (both source and target) is represented by the attributes (e.g., "black", "water"). At training stage, the attribute recognition models are learned using the source images from source classes and the corresponding attribute representations. At test stage, given an image from target class, we first generate its attribute representation using the attribute models. Because the attributes are shared among classes, the models learned from source classes also work for target classes. Then by comparing the similarity between the attribute representation of this image and the representations of each target class, the recognition result is generated. Based on the shared attributes, the knowledge can be transferred from source classes to target classes, and thus effective recognition models can be constructed for target classes.

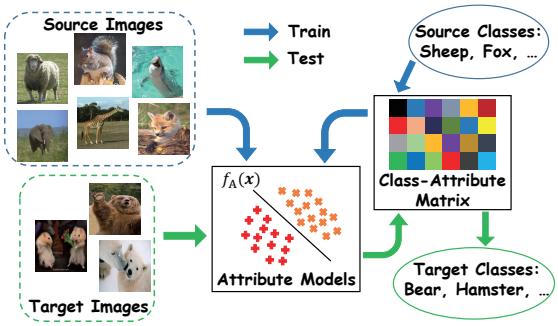


Figure 1: Attribute-based zero-shot recognition.

This paper focuses on ZSR with the transductive setting (Rohrbach, Ebert, and Schiele 2013; Fu et al. 2014a) in which the unlabeled (target) images from the target classes are available. Although the target images have no label information, we can still discover some class structures from them which can promote the recognition accuracy. For example, Rohrbach et al. proposed a label propagation method

[1]Both intermediates represent class labels in a shared space. The only difference is how the representations are constructed. We consistently use the notation *attribute* when there is no ambiguity.

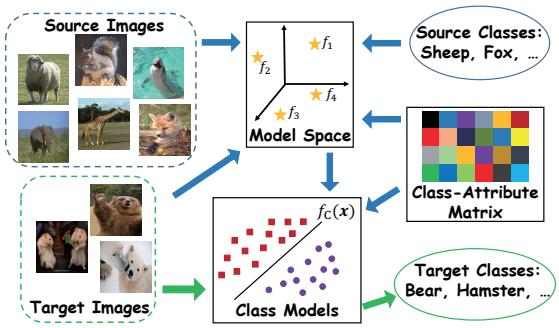


Figure 2: Framework overview.

based on the manifold structure (Belkin, Niyogi, and Sindhwani 2006) of target class. In this paper, we propose a novel transductive ZSR approach based on *shared model space* learning (SMS). The framework is illustrated in Figure 2. Our approach has several important characteristics. Firstly, noticing that the ultimate goal of ZSR is to learn class models, we propose to learn the models that can directly connect images and labels instead of learning attribute models connecting images and attributes. At test stage, we can generate the class labels directly from images without using attributes as intermediary level. Secondly, because we have no labeled data for directly building class models for target class, we propose to learn a shared model space for classes instead of the shared attribute space in existing works. Within this space, the model parameters for a target class can be directly generated using the attribute representation. Thus the knowledge can be transferred into the target class. Thirdly, unlike the disjointed framework in existing approaches that treats source classes and target classes separately, we propose a joint learning framework that takes both into consideration.

In summary, this paper makes several contributions below.

- We propose a novel transductive ZSR approach via shared model space learning. With the shared model space and class attributes, the recognition model which directly generates label for target class can be effectively constructed.
- We propose a joint learning framework that takes both source and target classes into account at training stage which leads to superior ZSR recognition performance. An effective and efficient learning algorithm is also proposed.
- We conduct extensive experiments on three benchmark datasets for ZSR. The results show that the proposed approach can significantly outperform the state-of-the-art related approaches, which demonstrates its effectiveness.

Related Work

Attribute-based ZSR

With class attributes as the intermediate, the knowledge from source classes can be transferred into the target classes because the attributes are shared among classes. In Direct/Indirect Attribute Prediction (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2014) (DAP/IAP), attribute classifiers are learned from source classes and then apply to target images. The recognition result is obtained by comparing the recognized attributes of images with the class attributes. In Attribute Label Embedding (Akata et al. 2013)(ALE), an embedding space maximizing the compatibility between images and labels is learned. In (Yu and Aloimonos 2010), topic model is applied to predict the attributes for target images.

The above approaches use manually defined attributes such as "white" and "water". Another line is to use the word vector (Turney and Pantel 2010; Huang et al. 2012; Mikolov et al. 2013) which represents each label by a vector learned from large-scale textual database like Wikipedia. In Cross-modal transfer (Socher et al. 2013), a regression model between images and word vectors are learned and the recognition is performed using a probabilistic model. In (Norouzi et al. 2013), a convex combination approach is proposed for ZSR. In (Elhoseiny, Saleh, and Elgammal 2013), a constrained optimization formulation combining regression function, knowledge transfer function is proposed for ZSR. In (Fu et al. 2015), the manifold structure in label semantic space is considered using absorbing Markov chain process.

Besides, some other attribute learning approaches have been proposed (Rastegari, Farhadi, and Forsyth 2012; Yu et al. 2013; Guo et al. 2015) to learn latent attributes for class satisfying some specific properties such as predictability and discriminability. They have reached some promising results.

Transductive ZSR

Transductive ZSR is an emerging topic which extends conventional ZSR into a semi-supervised learning scenario where labeled source images and unlabeled target images are available. In Propagated Semantic Transfer (Rohrbach, Ebert, and Schiele 2013) (PST), a projection which maps images to the label semantic space is learned from source classes. Then the target images are projected to the space. Finally, a label propagation step is performed to exploit the manifold structure of data. In Transductive Multi-view embedding (Fu et al. 2014a), an embedding space for both images and attributes is learned to rectify the projection shift. Then a Bayesian label propagation is adopted to generate the label for target images. Both approaches adopts the label semantic space as the intermediate for knowledge transfer while no class model directly connecting images and labels is learned. Besides, they both adopt disjointed steps which learns projections from images to semantic space with only source images and generates labels with only target images.

The Proposed Approach

Problem Definition and Notations

We have a set of source classes $\mathcal{C}^s = \{c_1^s, ..., c_{k_s}^s\}$ and n_s labeled source images $\mathcal{D}^s = \{(\mathbf{x}_1^s, \mathbf{y}_1^s), ..., (\mathbf{x}_{n_s}^s, \mathbf{y}_{n_s}^s)\}$, where $\mathbf{x}_i^s \in \mathbb{R}^d$ is the image feature and $\mathbf{y}_i^s \in \{0,1\}^{k_s}$ is the corresponding label vector which has $y_{ij} = 1$ if the image belongs to class c_j^s or 0 otherwise. We are given target images $\mathcal{D}^t = \{\mathbf{x}_1^t, ..., \mathbf{x}_{n_t}^t\}$ from k_t target classes $\mathcal{C}^t = \{c_1^t, ..., c_{k_t}^t\}$ satisfying $\mathcal{C}^s \cap \mathcal{C}^t = \emptyset$. The goal of transductive ZSR is to build models which can predict the label $c(\mathbf{x}_i^t)$ given $\mathbf{x}_i^t$ with no labeled training data for target

Table 1: Notations and descriptions in this paper.

Notation	Description	Notation	Description
$\mathbf{X}_s, \mathbf{X}_t$	images	n_s, n_t	#images
$\mathbf{Y}_s, \mathbf{Y}_t$	label matrix	d	#dimension
$\mathbf{A}_s, \mathbf{A}_t$	attribute matrix	r	#attributes
$\mathbf{W}^s, \mathbf{W}^t$	recognition model	k_s, k_t	#classes
$\mathbf{V}$	shared parameters	α, β	parameters

classes. For each class $c_i \in \mathcal{C}^s \cup \mathcal{C}^t$, we have an attribute representation $\mathbf{a}_i \in \mathbb{R}^r$ for it. We summarize some notations in this paper and the corresponding descriptions in Table 1.

Shared Model Space Learning

In this paper, we consider the multi-class recognition, whose recognition result is the class whose model can generate the maximal response among all the classes (Hsu and Lin 2002),

$$c(\mathbf{x}) = \text{argmax}_c \ f_c(\mathbf{x}) \tag{1}$$

where $\mathbf{x}$ is the input feature and f_c is the recognition model for class c. To construct the models, we can minimize the following loss function given a set of labeled data $(\mathbf{x}_i, \mathbf{y}_i)$,

$$\min_{f_c} \ \sum_i \sum_c \ell(f_c(\mathbf{x}_i), y_{ic}) + \mathcal{R}(f_c) \tag{2}$$

where $\ell(a, b)$ is a loss function for recognition error, such as hinge loss or squared loss, and $\mathcal{R}$ is the regularization term on model f_c. In this paper, we adopt the linear model, i.e., $f_c(\mathbf{x}) = \mathbf{x}\mathbf{w}_c'$ where $\mathbf{w}_c \in \mathbb{R}^d$ is the model parameter of f_c.

For source classes, we can train a model for each class because the labeled data is available. However, we have no labeled data for target class such that we can not adopt Eq. (2) to learn parameters. So we need to exploit knowledge from the labeled data in source classes and the unlabeled data in target classes simultaneously to learn the parameters.

Consider the class-model matrix $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_k]$. One can naturally notice that the model parameters are affected by the properties of a class. Different classes that have different properties will have different models. The previous research on attributes indicates that attributes can well characterize the properties of a class. Based on this idea, it is reasonable to assume that there is a function g shared by all classes that takes attributes of a class as the input and outputs the class model, i.e., $\mathbf{w}_c = g(\mathbf{a}_c)$. With this key assumption, we get the general learning framework for transductive ZSR,

$$\begin{aligned} \min_{g, \mathbf{Y}_t} \ & \mathcal{R}(g(\mathbf{a}_c)) + \sum_{i=1}^{n_s} \sum_{c=1}^{k_s} \ell(\mathbf{x}_i(g(\mathbf{a}_c))', y_{ic}) \\ & + \alpha \sum_{j=1}^{n_t} \sum_{c=1}^{k_t} \ell(\mathbf{x}_j(g(\mathbf{a}_c))', y_{jc}) \end{aligned} \tag{3}$$

The above learning framework has three important characteristics which are also the key differences from existing works. Firstly, it learns the class models that can directly generate the class labels from images, instead of using attributes as the intermediate. The one-step strategy can lead to less information loss at test stage, and thus we can expect better recognition result. Secondly, it regards attributes as the seed parameters for class models, instead of the recognition target of attribute models. Thirdly, it adopts a joint learning framework to learn models with labeled source data and unlabeled target data together, instead of only using source data. Because the target data is considered, the models can not only effectively transfer knowledge from source data, but also exploit the distribution of target data, which is an ideal property for knowledge transfer (Long et al. 2014).

Now we need to specify the function g. In this paper, we adopt the linear function, i.e., $\mathbf{w}_c = g(\mathbf{a}_c) = \mathbf{a}_c\mathbf{V}'$, where $\mathbf{V} \in \mathbb{R}^{d \times r}$ is the shared parameters. Although the linear function is simple, we find out in our experiment it works very well. Besides, we adopt the squared loss for the loss function ℓ, and the ridge regularization for $\mathcal{R}$, which leads to the specific objective function of the proposed approach:

$$\begin{aligned} \min_{\mathbf{V}, \mathbf{Y}_t} \ & \|\mathbf{X}_s\mathbf{V}\mathbf{A}_s' - \mathbf{Y}_s\|_F^2 + \alpha\|\mathbf{X}_t\mathbf{V}\mathbf{A}_t' - \mathbf{Y}_t\|_F^2 \\ & + \beta\|\mathbf{V}\mathbf{A}'\|_F^2, \text{s.t. } \|\mathbf{y}_j^t\|_0 = \mathbf{y}_j^t\mathbf{1}_{k_t}' = 1, j = 1, ..., n_t \end{aligned} \tag{4}$$

where α and β are hyper parameters, $\|\cdot\|_F$ denotes the Frobenius norm of matrix, $\|\cdot\|_0$ denotes the ℓ_0-norm of a vector, and $\mathbf{1}_k$ is a vector in which all elements are 1. Here, we can observe that the class-model matrix is factorized into the product of two matrices, i.e., $\mathbf{W} = \mathbf{A}\mathbf{V}'$, which is similar to the matrix factorization techniques (Furnas et al. 1988). Analogous to MF, we can regard $\mathbf{a}_c$ as the latent representation of a class model that reveals the properties of a class, and $\mathbf{V}$ as the basis in the latent space shared among classes. Motivated by this explanation, we term our approach as Shared Model Space Learning. In addition, since $\mathbf{V}$ appears in both source classes and target classes, it can work as the bridge to transfer knowledge between them.

Optimization Algorithm

We face the "chicken or the egg" dilemma when we optimize Eq. (4). To predict the label matrix $\mathbf{Y}_t$ for target images, we need to first know the shared parameters $\mathbf{V}$. On the other hand, to learn $\mathbf{V}$, we need to know the labels of target images. To address this problem, we propose an iterative algorithm using the *pseudo* labels. Specifically, we can apply some ZSR approaches, like PST, to generate the initial pseudo labels $\tilde{\mathbf{Y}}_t$. Then we can learn the shared parameters $\mathbf{V}$ given the pseudo labels. Because now $\mathbf{V}$ contains the knowledge from source classes, we can use it to refine the pseudo labels. On the other hand, the refined pseudo labels can further improve the quality of $\mathbf{V}$. Therefore, the refinement procedure can improve the quality of $\mathbf{V}$ and $\tilde{\mathbf{Y}}_s$ in each iteration until convergence. In the Experiment section, we empirically demonstrate the effectiveness of the iterative algorithm. The detailed steps for the algorithm are as below.

Fix $\tilde{\mathbf{Y}}^t$ and refine $\mathbf{V}$. Firstly, we re-denote the notations:

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

Algorithm 1 Transductive ZSR with Shared Model Space

Input: Source images $\mathbf{X}_s$; Source labels $\mathbf{Y}_s$;
Target images $\mathbf{X}_t$; Parameters α and β;
Source attributes $\mathbf{A}_s$; Target attributes $\mathbf{A}_t$;
Output: Shared model space $\mathbf{V}$;
Prediction for target images $\mathbf{Y}_t$;

```
Initialize Y_t using an existing ZSR approach;
repeat
    Update V by Eq. (7);
    Update Y_t by Eq. (9);
until Convergence;
Return V and Y_t;
```

To simplify the problem, we *approximate* Eq. (4) as follows,

$$\min_{\mathbf{V}} \ \mathcal{O}_{\mathbf{V}} = \|\mathbf{X}\mathbf{V}\mathbf{A}' - \mathbf{Y}\|_F^2 + \beta\|\mathbf{V}\mathbf{A}'\|_F^2 \tag{5}$$

Now we can calculate the derivative of $\mathcal{O}_{\mathbf{V}}$ w.r.t. $\mathbf{V}$ as below,

$$\frac{\partial \mathcal{O}_{\mathbf{V}}}{\partial \mathbf{V}} = 2\mathbf{X}'\mathbf{X}\mathbf{V}\mathbf{A}'\mathbf{A} - 2\mathbf{X}'\mathbf{Y}\mathbf{A} + 2\beta\mathbf{V}\mathbf{A}'\mathbf{A} \tag{6}$$

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

$$\mathbf{V} \leftarrow (\mathbf{X}'\mathbf{X} + \beta\mathbf{I}_d)^{-1}\mathbf{X}'\mathbf{Y}\mathbf{A}(\mathbf{A}'\mathbf{A})^{-1} \tag{7}$$

Fix $\mathbf{V}$ and update $\tilde{\mathbf{Y}}_t$. We can observe that Eq. (4) is row decoupled. Thus we can update each row in $\tilde{\mathbf{Y}}_t$ individually. And the objective function w.r.t. $\tilde{\mathbf{y}}_j^t$ can be written as below,

$$\min_{\tilde{\mathbf{y}}_j^t} \ \|\mathbf{x}_j^t\mathbf{V}\mathbf{A}_t' - \tilde{\mathbf{y}}_j^t\|_F^2 \ \text{ s.t. } \|\tilde{\mathbf{y}}_j^t\|_0 = \tilde{\mathbf{y}}_j^t\mathbf{1}_{k_t}' = 1 \tag{8}$$

Solving the above problem leads to the updating rule below,

$$\tilde{y}_{jc}^t = \begin{cases} 1, & \text{if } c = \text{argmax}_c \mathbf{x}_j^t\mathbf{V}\mathbf{a}_c^{t'} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

We can observe that Eq. (9) is the special case of Eq. (1) in which we specify the form for f. We can iterate the above two steps to refine the prediction until convergence. The whole optimization algorithm is summarized in Algorithm 1.

Discussion

We analyze the error bound of our approach. First we denote

$$\mathbf{z}_{ic} = \text{vec}(\mathbf{x}_i'\mathbf{a}_c) \in \mathbb{R}^{dr}, \ \mathbf{u} = \text{vec}(\mathbf{V}) \in \mathbb{R}^{dr} \tag{10}$$

where $\text{vec}(\cdot)$ is the vectorization operation which turns a matrix into a vector. With the above transformation, it is easy to verify $\mathbf{x}_i\mathbf{V}\mathbf{a}_c' = \mathbf{z}_{ic}\mathbf{u}'$. So we can rewrite Eq. (2) as follows,

$$\min_{\mathbf{u}} \sum_{\mathbf{x}_i \in \mathcal{D}^s \cup \mathcal{D}^t} \sum_{c \in \mathcal{C}^s \cup \mathcal{C}^t} \ell(\mathbf{z}_{ic}\mathbf{u}', y_{ic}) + \mathcal{R}(\mathbf{u}) \tag{11}$$

which leads to a domain adaptation problem (Ben-David et al. 2006). Now denote the true labeling function as $h(\mathbf{z})$, and the learned prediction function as $f(\mathbf{z})$ as in Eq. (9). We can define the expected error of f in $\mathcal{D}^s$ and $\mathcal{D}^t$ respectively as

$$\begin{aligned} \epsilon_s(f) &= \mathbb{E}_{\mathbf{z}\sim P_s}[|h(\mathbf{z}) - f(\mathbf{z})|] \\ \epsilon_t(f) &= \mathbb{E}_{\mathbf{z}\sim P_t}[|h(\mathbf{z}) - f(\mathbf{z})|] \end{aligned} \tag{12}$$

In this paper we follow the Theorem 1 in (Ben-David et al. 2006). Suppose the hypothesis space $\mathcal{H}$ containing f is of VC-dimension $\bar{d}$, then with probability at least $1-\delta$, for every $f \in \mathcal{H}$, its expected error in $\mathcal{D}^t$ is bounded as follows

$$\begin{aligned} \epsilon_t(f) \leq & \hat{\epsilon}_s(f) + \sqrt{\frac{4}{n}(\bar{d}\log\frac{2en}{\bar{d}} + \log\frac{4}{\delta})} \\ & + d_{\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^t) + \lambda \end{aligned} \tag{13}$$

where e denotes the base of natural logarithm, $\hat{\epsilon}_s(f)$ is the empirical error of f in $\mathcal{D}^s$, $\lambda = \inf_{f\in\mathcal{H}}[\epsilon_s(f) + \epsilon_t(f)]$, and $d_{\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^t)$ is the distribution distance between $\mathcal{D}^s$ and $\mathcal{D}^t$.

Here we are interested in the first term $\hat{\epsilon}_s(f)$ and the third term $d_{\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^t)$ in the error bound. In fact, these two terms reveal the two important factors that affect the ZSR performance. The first factor is the quality of the attributes which controls the first term. In this paper, a key assumption we make is that the attributes can well characterize the properties of a class, and thus the class model can be derived from the attributes. From Eq. (10), we can observe that we transform the original features into another feature space using the attributes. If we have good attributes, it is expected that the transformed features have legible class structures such that we can learn a model that has small error. On the other hand, the bad attributes, e.g., random attributes, may result in indistinguishable transformed features such that the learned model may have large error. The second factor is the relatedness between source classes, target classes, and attributes, which controls the third term. The fundamental assumption in ZSR is that the attributes are shared between source and target classes. In the best situation where source images and target images have the same conditional distribution given an attribute, and suppose all images are i.i.d., the distribution distance, such as Maximum Mean Discrepancy (MMD) (Pan, Kwok, and Yang 2008) will be small. On the other hand, if the attributes in the source classes does not appear in the target classes at all and vice versa, there will be very large distribution distance. For example, we define a set of attributes that half for animals and the other half for face. The source classes are for animal recognition, but the target classes are for face recognition. Although the attributes can well distinguish different classes, such attributes can not transfer knowledge at all. Based on the above analysis, we can summarize two principles to design attributes for ZSR. Firstly, to guarantee small error $\hat{\epsilon}_s(f)$, the attributes should well characterize the properties of classes. Secondly, to reduce the distribution distance $d_{\mathcal{H}}(\mathcal{D}^s, \mathcal{D}^t)$, the attributes should be strongly related to both source and target classes.

Experiment

Datasets and Settings

To evaluate the efficacy of the propose approach, we conduct extensive experiments on three benchmark datasets for ZSR. The first dataset is Animal with Attributes (AwA) (Lampert, Nickisch, and Harmeling 2014). This dataset contains $30,475$ images from 50 animal classes, such as "dog" and "fox". It contains 85 binary attributes and the class-attribute matrix is given. In this dataset, 40 classes with $24,295$ images are adopted as the source classes and 10 images with

Table 2: The statistics of datasets.

	AwA	aPY	SUN
#source class	40	20	707
#source image	24,295	12,695	14,140
#target class	10	12	10
#target images	6,180	2,644	200
#attribute	85	64	102
#dimension	4,096	9,751	17,032

Table 3: ZSR results.

Accuracy (%)	AwA	aPY	SUN
DAP	51.00	18.12	52.50
ZUA	53.75	26.02	56.00
PST	54.10	24.11	64.50
TME	69.91	28.47	68.50
SMS	**78.47**	**39.03**	**82.00**

$6,180$ images are adopted as the target classes. We follow the standard setting from the dataset for the source/target split. The second dataset is aPascal-aYahoo (aPY) (Farhadi et al. 2009). aPascal dataset has 20 types of objects for PASCAL VOC2008 challenge, such as "people" and "dog". It contains $12,695$ images. aYahoo dataset was collected from Yahoo image search. It has 12 classes which are similar but different from the ones in aPascal, such as "centaur" and "wolf". It contains $2,644$ images. There are 64 binary attributes and each image is annotated by them. We average the attribute representations of the images belonging to the same class to obtain the class-attribute representation. In aPY, the aPascal is adopted as the source, and the aYahoo is adopted as the target. The third dataset is SUN scene recognition dataset (Patterson and Hays 2012). This is a fine-grained dataset in which the difference between classes is quite small. It has 717 scenes such as "airport" and "palace", and each scene has 20 images. There are 102 attributes in this dataset and each image is annotated by them. We also average the image attributes to obtain the class attributes. For this dataset, we use 707 classes as the source and 10 classes as the target. The source/target split follows the setting in (Jayaraman and Grauman 2014). For AwA dataset, we extract the deep feature using the DeCAF (Donahue et al. 2014). For aPY and SUN dataset, we use the author-provided features, including HOG, color histograms, and so on. The statistics of three benchmarks are shown in Table 2.

We compare the proposed SMS to the following state-of-the-arts. The attribute-based approaches include Direct Attribute Prediction (Lampert, Nickisch, and Harmeling 2014) (DAP) and ZSR with Unreliable Attributes (Jayaraman and Grauman 2014) (ZUA). Besides, the transductive ZSR approaches have Propagated Semantic Transfer (Rohrbach, Ebert, and Schiele 2013) (PST), and Transductive Multi-view Embedding (Fu et al. 2014a) (TME). Besides, we adopt as evaluation metric the recognition accuracy on target classes.

Implementation Details

Our approach has two hyper parameters, α and β. In this paper, we adopt the cross validation (CV) to determine the values for them. For AwA and aPY datasets, we perform 4-fold CV. For SUN dataset, we perform 10-fold CV. Specifically, to perform k-fold CV, we split the source classes equally into k parts. In each fold, we choose one part as the validation set and the other $k-1$ parts form the training set. In addition, the values for α and β are selected from $\{0.01, 0.1, 1, 10, 100\}$.

Our optimization algorithm needs to initialize the pseudo labels $\tilde{\mathbf{Y}}_t$. As we mentioned above, we can adopt any existing ZSR approaches to generate the initial pseudo labels. For fair comparison, in this paper, we use the model that is learned in CV with the best CV result for self-initialization.

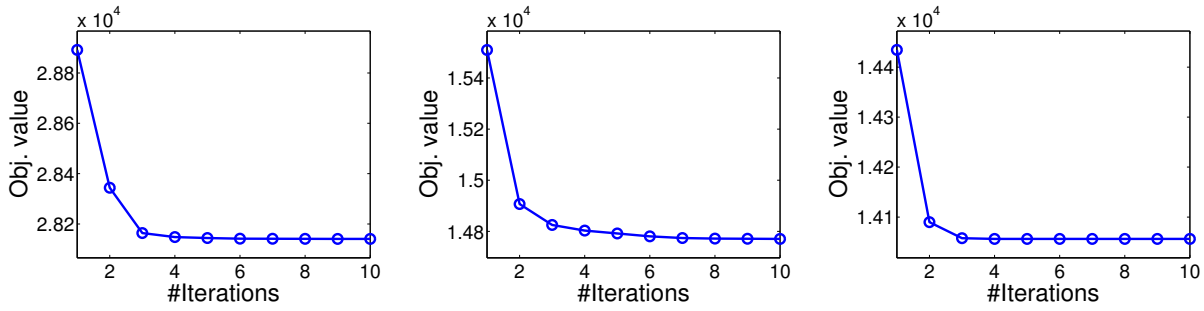
Results on Benchmarks

The ZSR accuracy of the proposed SMS and four baseline approaches on three benchmark datasets are summarized in Table 3. We can observe that SMS achieves much better performance than the four baseline approaches with statistical significance. The recognition accuracies of SMS on three datasets are 78.47%, 39.03%, and 82.00%, and the performance improvements compared with the best baseline approach TME are 8.56%, 10.56%, and 13.50% respectively, i.e., the error reductions are 28.45%, 14.76%, and 42.86%.

Firstly, SMS achieves more than 80% recognition accuracy on AwA and SUN dataset. This is a remarkable result for 10-class classification considering we have no labeled data for target classes. One may argue that the result on AwA is due to the powerful visual feature. Here we admit that the DeCAF feature indeed improves the performance on AwA. But we can also notice the big performance gap between SMS and baseline approaches which indicates that the difference in approaches plays an important role for this result.

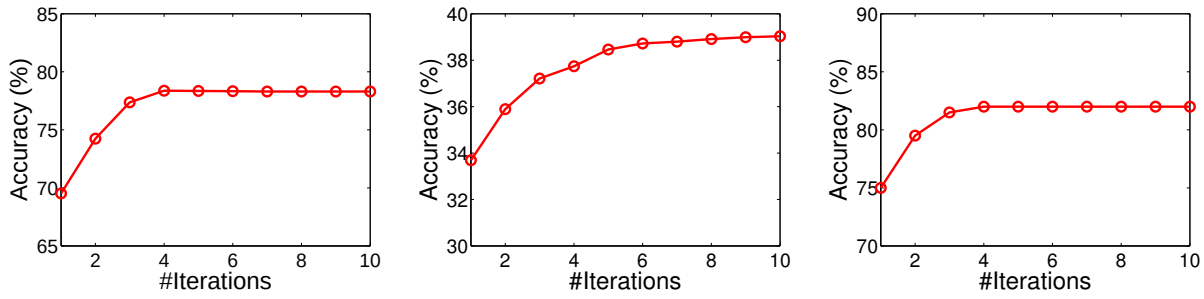
Secondly, ZUA achieves comparable performance to PST even though it makes no use of target images at all. In ZUA, the unreliability of attributes is taken into consideration. In our discussion about Eq. (13), we mentioned that the recognition accuracy on target classes partially relies on the quality of attributes. ZUA tries to construct robust models for target classes that accounts for the unreliability of attributes by utilizing the error tendencies of the attributes. Because it alleviates the effect of bad attributes, it can achieve comparable results to transductive ZSR approaches. The proposed SMS does not consider this situation. Thus it is expected that SMS can achieve better performance if the unreliability is considered. We leave this problem to our future research.

Lastly, SMS can significantly outperform the other two transductive ZSR approaches. As we have mentioned above, SMS has three important characteristics which are also the main differences from PST and TME. (1) SMS directly learns the class models instead of the attribute models. (2) The attributes are utilized as the parameters to generate the class models, instead of as the intermediary level in the recognition procedure. (3) SMS adopts a joint learning framework that considers source and target images simultaneously, while the others utilize disjoint methods. The results demonstrate the effectiveness of above characteristics.



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

Figure 3: Objective function value in Eq. (4) w.r.t. the number of iterations..



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

Figure 4: Recognition accuracy w.r.t. the number of iterations in Algorithm 1.

More Verification

To further verify the effectiveness of SMS, we inspect the performance of SMS and the optimization algorithm. We plot the objective function value in Eq. (4) and the recognition accuracy on target classes w.r.t. each iteration in Algorithm 1, which are shown in Figures 3 and 4 respectively.

Firstly, we can observe from Figure 3 that the objective function value can decrease steadily with more iterations and can converge with about 10 iterations, which validates the effectiveness of Algorithm 1. Besides, each iteration takes 8.81, 9.70, and 15.83 seconds on AwA, aPY and SUN datasets respectively, which is quite efficient compared with PST and TME that need to perform label propagation on a large graph. The speed is measured using a computer with Intel Core i7-2600 3.40 GHz CPU and 16GB memory.

Secondly, we adopt an iterative refinement procedure. In Figure 4, we can see that the recognition accuracy can increase steadily with more iterations, which indicates that we obtain more correct pseudo labels after each iteration. By iteratively refining the pseudo labels, the model can better capture the structure of target classes and improve the recognition performance. The results validate the efficacy of the iterative refinement. Besides, the recognition accuracy and the objective function value have similar trend. This phenomenon indicates that our objective function can reflect the ZSR performance. Thus, if we find more effective optimization algorithm, we can expect better ZSR performance. In our future research, we will investigate this interesting issue.

Last but not least, one key characteristic of the proposed approach is the joint learning framework. The fact that the final accuracy is higher than the one after self-initialization using the model learned with only source classes, and the increasing accuracy both demonstrate our motivation that considering the knowledge from the source classes and the structure of target classes in a unified way is indeed better than treating the source classes and target classes separately.

Conclusion

This paper investigates the transductive ZSR problem where labeled source images and unlabeled target images are available. We propose a novel joint leaning approach to learn the shared model space such that the knowledge can be effectively transferred between classes. It is different from existing works in three ways. Firstly, we learn class models that can direct generate the class labels from images instead of attribute models that recognize attributes. Secondly, we learn a shared model space for source and target classes such that the recognition model can be generated simply using the attributes of classes. Thirdly, we adopt a joint learning framework considering both source and target classes simultaneously. An effective and efficient optimization algorithm is proposed. Extensive experiments on three benchmark datasets demonstrate the efficacy of the proposed approach.

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

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 819–826.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR* 7:2399–2434.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2006. Analysis of representations for domain adaptation. In *NIPS*, 137–144.

Bishop, C. M., et al. 2006. *Pattern recognition and machine learning*, volume 1. Springer, New York.

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.

Elhoseiny, M.; Saleh, B.; and Elgammal, A. M. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *IEEE ICCV*, 2584–2591.

Farhadi, A.; Endres, I.; Hoiem, D.; and Forsyth, D. A. 2009. Describing objects by their attributes. In *CVPR*, 1778–1785.

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

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2014b. Learning multimodal latent attributes. *IEEE Trans. Pattern Anal. Mach. Intell.* 36(2):303–316.

Fu, Z.; Xiang, T.; Kodirov, E.; and Gong, S. 2015. Zero-shot object recognition by semantic manifold distance. In *IEEE Conference on Computer Vision and Pattern Recognition*.

Furnas, G. W.; Deerwester, S. C.; Dumais, S. T.; Landauer, T. K.; Harshman, R. A.; Streeter, L. A.; and Lochbaum, K. E. 1988. Information retrieval using a singular value decomposition model of latent semantic structure. In *SIGIR*, 465–480.

Guo, Y.; Ding, G.; Jin, X.; and Wang, J. 2015. Learning predictable and discriminative attributes for visual recognition. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 3783–3789.

Habibian, A.; Mensink, T.; and Snoek, C. G. M. 2014. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *ACM MM*, 17–26.

Hsu, C., and Lin, C. 2002. A comparison of methods for multiclass support vector machines. *TNN* 13(2):415–425.

Huang, E. H.; Socher, R.; Manning, C. D.; and Ng, A. Y. 2012. Improving word representations via global context and multiple word prototypes. In *ACL*, 873–882.

Jayaraman, D., and Grauman, K. 2014. Zero-shot recognition with unreliable attributes. In *NIPS*, 3464–3472.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 951–958.

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.

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.

Mensink, T.; Verbeek, J. J.; Perronnin, F.; and Csurka, G. 2013. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Trans. Pattern Anal. Mach. Intell.* 35(11):2624–2637.

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.

Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *NIPS*, 1410–1418.

Pan, S. J.; Kwok, J. T.; and Yang, Q. 2008. Transfer learning via dimensionality reduction. In *AAAI*, 677–682.

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.

Rastegari, M.; Farhadi, A.; and Forsyth, D. A. 2012. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 876–889.

Rohrbach, M.; Ebert, S.; and Schiele, B. 2013. Transfer learning in a transductive setting. In *NIPS*, 46–54.

Romera-Paredes, B., and Torr, P. H. S. 2015. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2152–2161.

Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. Y. 2013. Zero-shot learning through cross-modal transfer. In *NIPS*, 935–943.

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

Yu, X., and Aloimonos, Y. 2010. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, 127–140.

Yu, F. X.; Cao, L.; Feris, R. S.; Smith, J. R.; and Chang, S. 2013. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 771–778.