Synthesizing Samples for Zero-shot Learning

IJCAI Anonymous Submission 2625

Abstract

Zero-shot learning (ZSL) is to construct recognition models for unseen target classes that have no labeled samples for training. It utilizes the class attributes or semantic vectors as side information and transfers supervision information from related source classes with abundant labeled samples. Existing ZSL approaches adopt an intermediary embedding space to measure the similarity between a sample and the attributes of a target class to perform zero-shot classification. However, this way may suffer from the information loss caused by the embedding process and the similarity measure cannot fully make use of the data distribution. In this paper, we propose a novel approach which turns the ZSL problem into a conventional supervised learning problem by synthesizing samples for the unseen classes. Firstly, the probability distribution of an unseen class is estimated by using the knowledge from seen classes and the class attributes. Secondly, the samples are synthesized based on the distribution for the unseen class. Finally, we can train any supervised classifiers based on the synthesized samples. Extensive experiments on benchmarks demonstrate the superiority of the proposed approach to the state-of-the-art ZSL approaches.

1 Introduction

Recent years have witnessed the tremendous progress of several machine learning and computer vision tasks, such as object recognition, scene understanding, and fine-grained classification, together with the development of deep learning techniques [Krizhevsky *et al.*, 2012; He *et al.*, 2016]. It should be noticed that the learning scheme of them requires sufficient labeled samples for model training, like ImageNet [Russakovsky *et al.*, 2015]. This is affordable when dealing with common objects. However, the objects "in the wild" follow a long-tailed distribution such that the uncommon ones do not occur frequently enough, and the new concepts emerge everyday especially in the Web, which makes it difficult and expensive to collect and label a sufficiently large training set for model learning [Changpinyo *et al.*, 2016]. How to train effective classification models for the uncommon classes without using the labeled samples becomes an important and practical problem and has gathered considerable research interests from the machine learning and computer vision communities.

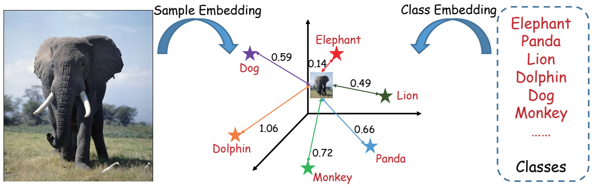


Figure 1: Framework of embedding based ZSL approaches.

It is estimated that humans can recognize approximate $30,000$ basic object categories and many more subordinate ones and they are able to identify new classes given an attribute description [Lampert *et al.*, 2014]. Based on this observation, many zero-shot learning (ZSL) approaches have been proposed [Akata *et al.*, 2015; Al-Halah *et al.*, 2016; Romera-Paredes and Torr, 2015; Zhang and Saligrama, 2016a]. The goal of ZSL is to build classifiers for target unseen classes given no labeled samples, with class attributes as side information and fully labeled source seen classes as knowledge source. Different from many supervised learning approaches which treat each class independently, ZSL associates classes with an intermediary attribute or semantic space and then transfers knowledge from the source seen classes to the target unseen classes based on the association. In this way, only the attribute vector of a target (unseen) class is required and the classification model can be built even without any labeled samples for this class. In particular, an embedding function is learned using the labeled samples of source seen classes that maps the images and classes into a common embedding space where the distance or similarity between them can be measured. Because the attributes are shared by both source and target classes, the embedding function learned by source classes can be directly applied to target classes [Farhadi *et al.*, 2009; Socher *et al.*, 2013]. Finally, given a test image, we map it into the embedding space and measure its distance to each target class and return the class with the minimal distance. An illustration of this ZSL framework is shown in Figure 1.

In reality, given the description of a new unseen object, humans can always imagine and picture some exemplar images of the target object with the help of the knowledge in-

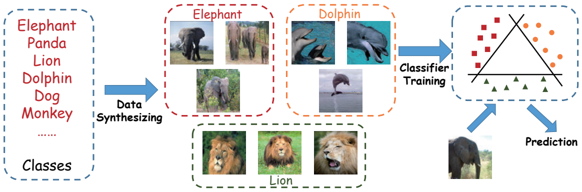


Figure 2: The proposed data synthesizing based ZSL.

duced from the other seen objects, and then utilize them as supervision to guide the future classification [Miller *et al.*, 2000]. Inspired by this observation, we propose a novel ZSL framework based on data synthesis, as shown in Figure 2, which is totally different from existing embedding based approaches. Intuitively, the embedding based ZSL can be regarded as learning how to recognize the characteristics of an image and match them to a class. On the contrary, our framework can be described as learning what a class visually looks like. In particular, the proposed framework has two explicit advantages over the embedding based framework. Firstly, the embedding based framework has to map the test image into an embedding space. It should be noticed that the embedding step may bring in information loss such that the overall performance of the system degrades [Fu *et al.*, 2014; Zhang and Saligrama, 2016b; Lazaridou *et al.*, 2015]. The proposed framework classifies a test image in the original space, which can avoid this problem. Secondly, the supervised learning techniques has been developed rapidly in recent decades but it is not clear how to combine the embedding based framework with most of them. In the proposed framework, labeled samples for the target classes are synthesized. In this way, we turn the ZSL problem into a conventional supervised learning problem such that we can take advantage of the power of supervised learning techniques in the ZSL task.

In particular, we synthesize samples for each target class by probability sampling. Given the labeled samples from source classes, the conditional probability $p(\mathbf{x}|c)$ for each source class c is computed. Then by using the association between the source classes' attributes and target classes' attributes, we estimate the conditional probability for each target class by a linear reconstruction method. Next, based on the distribution, some samples are synthesized. At last, any classification model can be learned in a conventional supervised way with the synthesized samples. The contributions of this paper are:

1. We propose a novel ZSL framework based on data synthesis. By synthesizing samples for each target class, we can turn the ZSL problem into a conventional supervised learning problem such that we can make use of many powerful tools and avoid the information loss from the embedding process.

2. Based on the structure of class attributes and image features, we adopt a simple linear reconstruction method to estimate the conditional probability for each target class and then the samples are synthesized based on the distribution. We empirically demonstrate that the synthesized samples can well approximate the true characteristics of the target classes. To our best knowledge, this is the first work to estimate the conditional probability in the image feature space for ZSL.

3. Comprehensive experimental evidence on four benchmark datasets demonstrates that the proposed approach can consistently outperform the state-of-the-art ZSL approaches.

2 Preliminaries and Related Works

2.1 Problem Definition and Notations

The definition of zero-shot learning is as follows. We are given a set of source classes $\mathcal{C}^s = \{c_1^s, ..., c_{k_s}^s\}$ and n_s labeled source samples $\mathcal{D}^s = \{(\mathbf{x}_1^s, \mathbf{y}_1^s), ..., (\mathbf{x}_{n_s}^s, \mathbf{y}_{n_s}^s)\}$ for training, where $\mathbf{x}_i^s \in \mathbb{R}^d$ is the feature vector and $\mathbf{y}_i^s \in \{0,1\}^{k_s}$ is the corresponding label vector which has $y_{ij} = 1$ if the sample i belongs to class c_j^s or 0 otherwise. We are given some target samples $\mathcal{D}^t = \{\mathbf{x}_1^t, ..., \mathbf{x}_{n_t}^t\}$ 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 ZSL is to build classification models which can predict the label $c(\mathbf{x}_i^t)$ given $\mathbf{x}_i^t$ with no labeled training data for target classes available. To associate source classes and target classes to facilitate knowledge transfer, for each class $c_i \in \mathcal{C}^s \cup \mathcal{C}^t$, we assign a class attribute representation $\mathbf{a}_i \in \mathbb{R}^q$ to it which can be constructed from manual definition or the word2vec tool.

2.2 Related Works

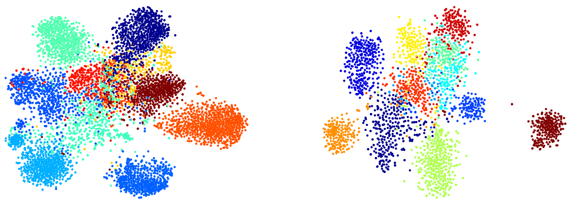
As introduced before, most of the existing ZSL approaches follow the embedding based framework illustrated in Figure 1. Formally, based on the problem definition and notations above, the classification methods of the previous approaches can be summarized into the general function as follows:

$$c(\mathbf{x}^t) = \text{argmax}_{c \in \mathcal{C}^t} sim(\phi(\mathbf{x}^t), \psi(\mathbf{a}_c)) \qquad (1)$$

where ϕ is the embedding function for images, ψ is the embedding function for classes, and $sim(\cdot,\cdot)$ is a similarity or distance measure function between the embedded images and classes. Existing ZSL approaches differ from each other due to different choices of these functions. For example, Lampert *et al.* [2014] adopted linear classifiers, identity function, and Euclidean distance respectively. Romera-Paredes and Torr [2015] used linear projection, identity function and inner product similarity. Fu *et al.* [2015] propose to use a deep model DeViSE [Frome *et al.*, 2013] for image projection and measure the similarity using the semantic manifold distance obtained from absorbing Markov chain process. Zhang and Saligrama [2016a] utilized the unit-ball constrained projection, simplex constrained projection, and aligned inner product similarity. Some approaches have more complicated formulation. But we can also simplify them into the general function. For example, the formulation of Changpinyo *et al.* [2016] can be simplified as the combination of the linear projection by virtual classifiers, exponential transformation, and inner product similarity. Recently many ZSL approaches have been proposed [Akata *et al.*, 2015; Xian *et al.*, 2016; Bucher *et al.*, 2016; Fu and Sigal, 2016]. Because of space limit, we cannot review all of them in detail. But they mostly follow the general function above. To learn these functions, the labeled source samples are used to maximize the function:

$$(\phi, \psi) = \text{argmax}_{(\phi,\psi)} \sum\nolimits_i sim(\phi(\mathbf{x}_i^t), \psi(\mathbf{a}_{c(\mathbf{x}_i^t)})) \qquad (2)$$

Moreover, because the embedding process may lead to critical problems and the distributions of target classes are not effectively described, such as the domain shift problem [Fu *et al.*, 2014] and hubness problem [Lazaridou *et al.*, 2015], many transductive ZSL approaches are proposed which make



(a) AwA (b) aPY

Figure 3: t-SNE visualization of samples from AwA and aPY datasets. Points with the same color belong to the same class.

use of the unlabeled target samples to better capture the target class structure [Kodirov *et al.*, 2015; Guo *et al.*, 2016; Zhang and Saligrama, 2016b]. However, we need to emphasize here that our work focuses on the inductive ZSL setting where no samples in target classes are available at all.

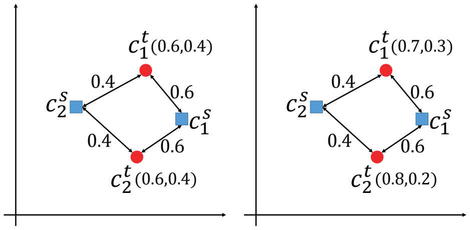
Data synthesis is an effective method to deal with the lack of training data, such as in the learning from imbalanced data problem [He and Garcia, 2009] and few-shot learning problem [Miller *et al.*, 2000; Kwitt *et al.*, 2016]. However, how to apply it to the zero-shot scenario is still a problem. Yu and Aloimonos [2010] made attempt to synthesize data for ZSL using the Author-Topic model [Rosen-Zvi *et al.*, 2010]. However, it should be noticed that their approach can only deal with discrete attributes and discrete visual features like bag-of-visual-word feature. In most of the recent ZSL settings, which are more practical in real world, the attributes and the visual features usually have continuous values, like the $\mathrm{word2vec}$ based attributes and the deep learning based visual features. Obviously, it is unclear and difficult, if not impossible, to apply their approach to these settings, while our approach is capable of handling these practical scenarios.

3 The Propose Approach

3.1 Distribution Estimation by Reconstruction

Because of the lack of labeled samples, it is challenging to train classifiers for target classes in a conventional supervised way. To address this problem, we propose to synthesize some samples for each target class. In particular, for each target class, we wish to estimate its conditional probability $p(\mathbf{x}|c)$ and then it is easy to synthesize samples from it by simple probability sampling. However, if we have no prior about the data distribution, the estimation will be somehow difficult. Therefore, we first briefly investigate the distribution of data.

It is demonstrated that the pre-trained convolutional neural network is a very powerful image feature extractor [Donahue *et al.*, 2014]. Therefore, we choose the VGG-19 network [Simonyan and Zisserman, 2014] and use the $\mathrm{fc7}$ layer outputs as the image feature, which is a $4,096$-dimensional vector. We use the t-SNE [Van der Maaten and Hinton, 2008] to visualize the features of some classes from Animal with Attributes (AwA) [Lampert *et al.*, 2014] and aPascal-aYahoo (aPY) [Farhadi *et al.*, 2009], as shown in Figure 3. Here, it can be observed that the samples from the same class roughly form a cluster. Based on the observation, it is reasonable to assume a Gaussian distribution for each target class, i.e., $p(\mathbf{x}|c) \sim \mathcal{N}(\mathbf{u}_c, \mathbf{\Sigma}_c)$. For source classes, the mean vector $\mathbf{u}_c$ and the covariance matrix $\mathbf{\Sigma}_c$ can be easily obtained from its labeled samples. However, for a target class, we have no more than the attribute vector $\mathbf{a}_c^t$ and thus it is not that straightforward to estimate the parameters like the source classes.



(a) Similarity (b) Reconstruction

Figure 4: The numbers next to lines are the similarity between classes. The numbers in the brackets are the weights to estimate the parameters. In the left subfigure, only the similarity is considered such that two different target classes may have the same estimated distribution. In the right subfigure, the problem is solved since the structure of classes is considered.

There is a saying, "one takes the behavior of one's company." In fact, this idea has been widely accepted by machine learning and computer vision communities. In the image classification task, it is always believed that similar images (short distance between their features) are more likely to belong to the same class, which is the underlying assumption of kNN classifier [Altman, 1992]. Analogously, in the class level, this idea seems reasonable too, indicating that similar classes should have similar properties, like the probability distribution. Fortunately, the similarity between classes can be measured by their attributes. One simple way to measure the similarity between a target class c^t and any source class c_j^s is:

$$s_j = \exp(-\frac{\|\mathbf{a}_c^t - \mathbf{a}_j^s\|_2^2}{\epsilon^2}) \tag{3}$$

where ϵ is the mean value of the distances between attribute vectors of any two source classes. With the similarity, it is straightforward to estimate the distribution parameters for c^t:

$$\mathbf{u}_{c^t} = \frac{1}{z}\sum\nolimits_{j=1}^{k_s} s_j \mathbf{u}_{c_j^s}, \ \mathbf{\Sigma}_{c^t} = \frac{1}{z}\sum\nolimits_{j=1}^{k_s} s_j \mathbf{\Sigma}_{c_j^s} \tag{4}$$

where $z = \sum_j s_j$ is a normalization parameter. In this way, the distribution parameters for a target class can approximately estimated from the information of the source classes'.

However, only considering the similarity seems too simple to well capture the properties of classes. As illustrated in Figure 4(a), because two target classes c_1^t and c_2^t have the same distance to source classes c_1^s and c_2^s, they obtain the same parameters by Eq. (4) even if they are different. In fact, the relative structure of the classes should be also taken into account. To address this issue, we propose a reconstruction method to estimate the parameters. In particular, suppose the parameters are estimated with w_j as the the weights as below:

$$\mathbf{u}_{c^t} = \frac{1}{z}\sum\nolimits_{j=1}^{k_s} w_j \mathbf{u}_{c_j^s}, \ \mathbf{\Sigma}_{c^t} = \frac{1}{z}\sum\nolimits_{j=1}^{k_s} w_j \mathbf{\Sigma}_{c_j^s} \tag{5}$$

To preserve the structure, we hope the weights are constructed such that $\mathbf{a}_{c^t} \approx \frac{1}{z}\sum_j w_j \mathbf{a}_j^s$. To find the optimal weights, it is reasonable to minimize the following reconstruction error:

$$\min_{w_j} \|\mathbf{a}_{c^t} - \mathbf{A}\mathbf{w}\|_2^2 + \mathcal{R}(w_j), \text{ s.t. } \sum\nolimits_j w_j = 1 \tag{6}$$

where $\mathbf{A} = [\mathbf{a}_1^s, ..., \mathbf{a}_{k_s}^s]$ and $\mathcal{R}$ is a regularization term. Obviously, without proper regularization, solving the problem may assign large weights to dissimilar classes. As discussed above, we hope the similar classes have more impact on the target class. Therefore, following the locality constrained reconstruction [Wang *et al.*, 2010], we further incorporate the similarity s_j as a regularization to the weights as follows:

$$\min_{w_j} \|\mathbf{a}_{c^t} - \mathbf{A}\mathbf{w}\|_2^2 + \lambda \sum\nolimits_j w_j/s_j, \text{ s.t. } \sum\nolimits_j w_j = 1 \tag{7}$$

Where λ is a trade-off parameter. Obviously, for dissimilar class with small s_j, minimizing the function will assign small weight w_j. The solution to the above problem is given by:

$$\mathbf{w} = ((\mathbf{A} - \mathbf{1}\mathbf{a}'_{c^t})(\mathbf{A} - \mathbf{1}\mathbf{a}'_{c^t})' + \lambda \text{diag}(s_1, ..., s_{k_s}))^{-1}/z \tag{8}$$

where $z = \sum_j w_j$ is the normalization factor. Moreover, we can take one step further to remove the influence of dissimilar source classes on the target class. In particular, we do not need to use all source classes for reconstruction. Instead, we only need the k-nearest neighbors ($k \ll k_s$) of c^t in $\mathcal{C}^s$, denoted as $\mathcal{N}_k$. In this way, the matrix $\mathbf{A}$ is reduced to $\mathbf{A}_{NN} = [\mathbf{a}_j^s]_{j \in \mathcal{N}_k}$ and we now just solve the subproblem to obtain weights $w_j (j \in \mathcal{N}_k)$ and simply set $w_j = 0 (j \notin \mathcal{N}_k)$. Then with the reconstruction based weights, the probability distribution parameters of c^t can be constructed by Eq. (5).

There is one issue worth discussing about. The covariance matrix contains a large number of parameters. For example, when using the $4,096$-dimensional deep feature, the matrix has about 16 million elements. Therefore, the estimation of it will be complicated and very imprecise if we use the whole matrix without any constraint. Here we consider two convenient simplifications. The first is to assume $\mathbf{\Sigma}_c = \sigma_c \mathbf{I}$ which is the simplest approximation. In fact, Socher *et al.* [2013] also assumed the isotropic Gaussian to prevent overfitting the target class. In this way, we only need to estimate the parameter σ_c for each class. The second is to assume $\mathbf{\Sigma}_c = \text{diag}(\sigma_{c1}, ..., \sigma_{cd})$ where we only consider the diagonal elements and the other elements are assumed to be 0. This is more complicated than the isotropic variance such that it can better fit the data, but much simpler than the whole variance with orders of magnitudes and thus it is more precise and less likely to overfit. In the experiment section, we consistently use the second simplification for the matrix $\mathbf{\Sigma}_c$.

3.2 Classifier Training

For each target class $c^t \in \mathcal{C}^t$, we obtain the estimated conditional probability distribution $p(\mathbf{x}|c^t) \sim \mathcal{N}(\mathbf{u}_{c^t}, \mathbf{\Sigma}_{c^t})$. Then we can perform random sampling with the distribution to synthesize $\mathcal{S}$ samples for each target class, which leads to a labeled training set with $k_t \times \mathcal{S}$ synthesized samples for learning classifiers for target classes. In this way, we turn the ZSL problem into a conventional supervised learning problem. Intuitively, any supervised classifiers can be used based on the synthesized training set, such as kNN classifier, SVM, and L-ogistic Regression. Moreover, some other techniques such as boosting methods like AdaBoost and metric learning methods can be also utilized. Compared to existing embedding based ZSL approaches, it is more straightforward to combine our approach with the supervised learning techniques such that our approach can better take advantage of the power of them.

Here we can notice that our approach falls into the embedding based framework in an extreme case. In particular, if only one sample for each target class is synthesized and we require it to be $\mathbf{u}_{c^t}$ and we use the 1NN classifier, it becomes a standard embedding and similarity measure procedure, which is equivalent to the embedding based framework if we regard the original image feature space as the embedding space. However, in this way, the variance information is not considered. In addition, because only one sample is synthesized, it fails to provide sufficient variability, which is a critical problem for the recognition task [Kwitt *et al.*, 2016].

3.3 Discussion

Now we analyze the error bound of our approach. Denote $\mathcal{D}^{syn}$ as the synthesized labeled samples for target classes, and $\mathcal{D}^t$ as the true samples of target classes. The true labeling function is $h(\mathbf{x})$ and the learned prediction function is $f(\mathbf{x})$. The distribution of $\mathcal{D}^{syn}$ is P_{syn} and of $\mathcal{D}^t$ is P_t. We define the prediction error of f in $\mathcal{D}^{syn}$ and $\mathcal{D}^t$ respectively as:

$$\epsilon_{syn}(f) = \mathbb{E}_{\mathbf{x} \sim P_{syn}}[|h(\mathbf{x}) - f(\mathbf{x})|] \tag{9}$$

$$\epsilon_t(f) = \mathbb{E}_{\mathbf{x} \sim P_t}[|h(\mathbf{x}) - f(\mathbf{x})|] \tag{10}$$

We can consider it as a domain adaptation problem [Ben-David *et al.*, 2006]. Following 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 $f \in \mathcal{H}$, the expected error $\epsilon_t(f)$ is bounded as follows:

$$\begin{aligned} \epsilon_t(f) \leq & \hat{\epsilon}_{syn}(f) + \sqrt{\frac{4}{n}(\bar{d}\log\frac{2en}{\bar{d}} + \log\frac{4}{\delta})} \\ & + d_{\mathcal{H}}(\mathcal{D}^{syn}, \mathcal{D}^t) + \rho \end{aligned} \tag{11}$$

where $\hat{\epsilon}_{syn}(f)$ is the empirical error of f in $\mathcal{D}^{syn}$, $\rho = \inf_{f \in \mathcal{H}}[\epsilon_{syn}(f) + \epsilon_t(f)]$, $d_{\mathcal{H}}(\mathcal{D}^{syn}, \mathcal{D}^t)$ is the distribution distance between $\mathcal{D}^{syn}$ and $\mathcal{D}^t$, e is the base of natural logarithm, and $n = k_t\mathcal{S}$ is the number of synthesized samples.

Our goal is to minimize $\epsilon_t(f)$. In fact, training classifier with $\mathcal{D}^{syn}$ is to minimize $\hat{\epsilon}_{syn}(f)$. For the second term, we can notice that the embedding based case, as discussed above, has $n = k_t \times 1$, while our approach has $n = k_t\mathcal{S}(\mathcal{S} \gg 1)$ indicating that our approach can generalize better, which is consistent with the observation by Kwitt *et al.* [2016]. The third term is very important. In fact, the distribution of $\mathcal{D}^{syn}$ is estimated by the structure of the class attributes. Therefore, if we have high quality attributes that are capable of perfectly preserving the structure of visual similarity among classes, i.e., the attributes and the distribution parameters can be reconstructed by the same weights, the distance between the estimated distribution and the true distribution will be very small, leading to small test error on true samples using the synthesized samples trained classifier. Interestingly, this distance seems to be a good measure of attribute quality. In fact,

Table 1: The statistics of datasets.

	AwA	aPY	SUN	CUB
#source class	40	20	707	150
#source sample	24, 295	14, 140	12, 695	–
#target class	10	12	10	50
#target sample	6, 180	2, 644	200	–
#attributes	85	64	102	312

previous works have paid little attention to evaluate the quality of the attributes in a principled way. The only metric considered before is the test performance. However, since the labels for test samples are not available, this is not feasible for real-world applications. But with this term, we can use the estimated and true distributions of source classes to compute the distance to the measure the quality of attributes, which can further guild the design and choice of the attributes.

4 Experiment

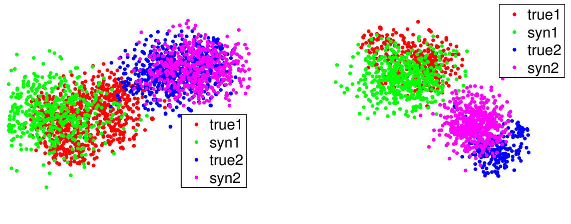
4.1 Datasets and Settings

In this paper, we adopt four benchmark datasets for ZSL. The first is Animal with Attributes (AwA) [Lampert *et al.*, 2014] using a standard source-target split with 40 source classes and 10 target classes. The second is aPascal-aYahoo [Farhadi *et al.*, 2009]. The aPascal subset has 20 objects from VOC challenge and the aYahoo subset has related 12 objects collected from Yahoo image search engine. Following the standard setting, the aPascal provides the source classes and the aYahoo provides the target classes. The third is SUN scene recognition dataset [Patterson and Hays, 2012] which has 717 scenes like "airport" and "palace". Following the standard setting [Jayaraman and Grauman, 2014], 707 scenes are source classes and 10 scenes are target classes. The fourth is Caltech-UCSD-Birds-200-2011 (CUB) [Wah *et al.*, 2011] which has 200 bird species. We follow the suggested split by Akata *et al.* [2015] which uses 150 species as source classes and 50 species as target classes. For each image, we use the VGG-19 network pre-trained on ImageNet [Simonyan and Zisserman, 2014] as feature extractor following Zhang and Saligrama [2016a]. Specifically, we use the 4, 096-dimensional output of the top fully-connected layer of the network as the feature vector. For all datasets, we utilize the attributes provided by the original datasets. The detailed statistics of these four datasets are summarized in TABLE 1.

To determine the model parameters, we employ the class-wise cross-validation method [Zhang and Saligrama, 2016a; Guo *et al.*, 2016]. In particular, we use the labeled source classes to simulate the zero-shot setting by splitting them by class into a training set and a validation set. We use 4-fold CV in this paper. After obtaining the optimal parameters, we use the whole training set to train the final model for evaluation.

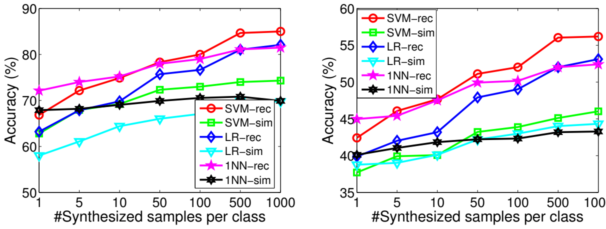
4.2 Analysis

The quality of distribution estimation. We first investigate one key issue of our approach. Specifically, we use the relationship among class attributes to estimate the conditional distribution of each target class. So, it is very important that the estimated distribution can approximate the true distribution, or otherwise the classifiers trained with the synthesized samples perform poorly for true samples. In Figure 5, we use t-SNE to visualize the true samples from two target classes (denoted as true1 and true2) and the synthesized samples for these two classes respectively sampled from the estimated distributions (denoted as syn1 and syn2) for AwA and aPY datasets. It can be observed that the estimated distributions can well approximate the true distributions, which demonstrates that it is challenging but feasible to use class attributes to estimate the data distribution for target classes and the proposed reconstruction method can yield high quality estimation results. The other datasets and classes also have similar results, which builds a solid foundation for our approach.



(a) AwA (b) aPY

Figure 5: Investigation on the quality of the synthesized samples. True1 and true2 denote the true samples from two target classes. Syn1 and syn2 stand for the synthesized samples from the estimated distributions of the corresponding classes.



(a) AwA (b) aPY

Figure 6: Investigation on the influence of different classifiers (SVM, LR, 1NN), different distribution estimation methods (**rec**onstruction using Eq. (5), **sim**imlarity using Eq. (4)), and the number of synthesized samples for each target class.

The effect of estimation method. As discussed before, one important step is to use similar source classes to estimate the distribution of target classes. In Eq. (4), we directly adopt the similarity as the estimation weight. In Figure 4 we illustrate that the similarity based method cannot well preserve the structure of classes. To address this issue we propose to take one more step and employ the reconstruction method based on the similarity to learn the weights in Eq. (7). In Figure 6 we empirically evaluate the influence of these two estimation methods, denoted briefly as "rec" for reconstruction and "sim" for similarity. Obviously, we can notice that the reconstruction based methods achieves higher accuracy than the similarity based methods in almost all circumstance, which demonstrates the superiority and rationality of the reconstruction based method and validates that the reconstruction based method can better estimate the distribution of target classes.

Table 2: Zero-shot classification accuracy on four benchmark datasets.

Approach	Animal with Attributes	aPascal-aYahoo	SUN	Caltech-UCSD-Birds
Akata *et al.* 2015	55.7			50.1
Al-Halah *et al.* 2016	67.5	37.0		
Bucher *et al.* 2016	77.32 ± 1.03	53.15 ± 0.88	84.41 ± 0.71	43.29 ± 0.38
Changpinyo *et al.* 2016	72.9			54.5
Fu *et al.* 2015	66.0			
Guo *et al.* 2017	79.07 ± 0.58	43.59 ± 0.42	83.04 ± 0.19	
Kodirov *et al.* 2015	75.6	26.5		40.6
Lampert *et al.* 2014	57.23	38.16	72.00	
Qiao *et al.* 2016	66.7			50.1
Romera-Paredes and Torr 2015	75.32 ± 2.28	24.22 ± 2.89	82.10 ± 0.32	
Wang *et al.* 2016	82.43			46.24
Xian *et al.* 2016	76.1			47.4
Zhang and Saligrama 2015	76.72 ± 0.83	42.90 ± 0.73	79.50 ± 1.22	30.41 ± 0.20
Zhang and Saligrama 2016a	79.12 ± 0.53	50.35 ± 2.97	83.83 ± 0.29	41.78 ± 0.52
Ours	$\mathbf{84.67 \pm 0.43}$	$\mathbf{55.04 \pm 0.81}$	$\mathbf{85.00 \pm 0.50}$	$\mathbf{56.75 \pm 0.29}$

The effect of classifiers. As an important property, our approach turns the ZSL problem into the conventional supervised learning problem such that we can utilize any powerful supervised tools. In this paper, we simply adopt three kinds of classifiers, SVM, Logistic Regression (LR) and 1NN. We evaluate their performance on AwA and aPY and the results are shown in Figure 6. Typically, SVM performs better than LR and 1NN especially when sufficient samples are synthesized. In fact, there is still difference between the estimated distribution and true distribution although the former can well approximate the latter as shown in Figure 5. Fortunately, the max-margin property of SVM seems to be to somehow robust to the distribution gap. In the future, we plan to incorporate some domain adaptation techniques [Pan and Yang, 2010] in the transductive setting to further improve the performance.

The effect of the number of synthesized samples. We further investigate the impact of the number of synthesized samples for each target class, i.e., $\mathcal{S}$, on the performance, as shown in Figure 6. Generally, the performance gets better with more synthesized samples at first since more information and variability about the target classes are given [Kwitt *et al.*, 2016]. When sufficient samples are synthesized ($\mathcal{S} > 500$), the accuracy stops increasing given more samples finally.

4.3 Benchmark Comparison

Now we compare the proposed approach to the state-of-the-art ZSL approaches on four benchmark datasets. Based on the above analysis, we employ SVM as the classifier. For each target class, 500 samples are synthesized using the reconstruction based distribution. The results are summarized in Table 2. From the results, we can clearly observe the consistently improvements upon the state-of-the-arts given by the proposed approach, which demonstrates the effectiveness of the sample synthesis idea for ZSL. In fact, our framework is based on data synthesis and turns the ZSL problem into a conventional supervised learning setting, which is totally different from the embedding based framework adopted by most ZSL approaches. The results validate the superiority of the proposed framework to the embedding based framework.

Among all baselines, Zhang and Saligrama [2016a] adopts the most joint embedding function, which achieves one of the best results on AwA, aPY, and SUN. The approach of Changpinyo *et al.* [2016] constructs the synthesized classifiers in the image feature space, which is equivalent to using image feature space as the embedding space, achieving best result in baselines on CUB. However, it can be observed that they still perform worse than our approach, which is another important evidence for the superiority of the proposed approach.

Moreover, we observe that the proposed approach is even better than some transductive approaches, like Kodirov *et al.* [2015]. In the transductive setting, the unlabeled target samples are given such that it is easier to capture the properties of target classes compared to the inductive setting where only the attributes of the target classes are available. However, because an embedding is employed, the structure of data is not well preserved. It demonstrates that the embedding step may cause information loss such that the overall performance of the system degrades. Without the embedding, our approach directly synthesizes samples in the original feature space, preventing it from this problem, and leading to better results.

5 Conclusion

In this paper, we propose a novel approach for ZSL. Different from previous embedding based framework, we propose to directly synthesize labeled samples for each target class in the original image space, which turns the ZSL problem into a conventional ZSL problem. Specifically, the conditional probability distribution for each target class is estimated by linear reconstruction based on the structure of the class attributes. Then the samples are synthesized by random sampling with the distribution for each target class. Any classifiers can be trained then with the synthesized samples, making the proposed approach flexible for different situations. We conduct comprehensive empirical analysis on several benchmark datasets. The experimental results demonstrate the superiority of the proposed approach to the state-of-the-art ZSL approaches, which validates its effectiveness for ZSL.

References

[Akata *et al.*, 2015] Z. Akata, S. E. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[Al-Halah *et al.*, 2016] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016.

[Altman, 1992] N. S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3):175–185, 1992.

[Ben-David *et al.*, 2006] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *NIPS*, 2006.

[Bucher *et al.*, 2016] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, 2016.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016.

[Donahue *et al.*, 2014] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014.

[Farhadi *et al.*, 2009] A. Farhadi, I. Endres, D. Hoiem, and D. A. Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

[Frome *et al.*, 2013] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[Fu and Sigal, 2016] Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, 2016.

[Fu *et al.*, 2014] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.

[Fu *et al.*, 2015] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

[Guo *et al.*, 2016] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI*, 2016.

[Guo *et al.*, 2017] Yuchen Guo, Guiguang Ding, Jungong Han, and Yue Gao. Zero-shot recognition via direct classifier learning with transferred samples and pseudo labels. In *AAAI*, 2017.

[He and Garcia, 2009] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE TKDE*, 2009.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Jayaraman and Grauman, 2014] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, 2014.

[Kodirov *et al.*, 2015] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Kwitt *et al.*, 2016] Roland Kwitt, Sebastian Hegenbart, and Marc Niethammer. One-shot learning of scene locations via feature trajectory transfer. In *CVPR*, 2016.

[Lampert *et al.*, 2014] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2014.

[Lazaridou *et al.*, 2015] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *ACL*, 2015.

[Miller *et al.*, 2000] Erik G. Miller, Nicholas E. Matsakis, and Paul A. Viola. Learning from one example through shared densities on transforms. In *CVPR*, 2000.

[Pan and Yang, 2010] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE TKDE*, 2010.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012.

[Qiao *et al.*, 2016] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Less is more: Zero-shot learning from online textual documents with noise suppression. In *CVPR*, 2016.

[Romera-Paredes and Torr, 2015] Bernardino Romera-Paredes and Philip H. S. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[Rosen-Zvi *et al.*, 2010] M.l Rosen-Zvi, C. Chemudugunta, T. L. Griffiths, P. Smyth, and M. Steyvers. Learning author-topic models from text corpora. *ACM TIST*, 2010.

[Russakovsky *et al.*, 2015] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z.g Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[Simonyan and Zisserman, 2014] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[Van der Maaten and Hinton, 2008] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008.

[Wah *et al.*, 2011] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[Wang *et al.*, 2010] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas S. Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[Wang *et al.*, 2016] D. Wang, Y. Li, Y. Lin, and Y. Zhuang. Relational knowledge transfer for zero-shot learning. In *AAAI*, 2016.

[Xian *et al.*, 2016] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[Yu and Aloimonos, 2010] Xiaodong Yu and Yiannis Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, 2010.

[Zhang and Saligrama, 2015] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[Zhang and Saligrama, 2016a] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016.

[Zhang and Saligrama, 2016b] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *ECCV*, 2016.