Semi-supervised Zero-Shot Learning by a Clustering-based Approach

Anonymous CVPR submission

Paper ID 3035

Abstract

In some of object recognition problems, labeled data may not be available for all categories. Zero-shot learning utilizes auxiliary information (also called class signatures) describing all categories in order to find a classifier that can also recognize samples from categories with no labeled instance. In this paper, we propose a novel semi-supervised zero-shot learning method that works on a representation space corresponding to abstract deep visual features. We use the idea that the rich deep visual features provide a representation space in which samples of each class are usually condensed in a cluster. We seek a transformation on signatures to map them onto the visual features, such that the mapped signatures of the seen classes are close to labeled samples of the corresponding classes and unlabeled data are also close to the mapped signatures of one of the unseen classes. The effectiveness of the proposed method is demonstrated through extensive experiments on four public benchmarks and we show that our method improves the state-of-the-art prediction accuracies.

1. Introduction

Zero-shot learning [17, 23, 16, 9] is an extension to the conventional supervised learning scenario that releases the assumption of having ample labeled instances for all categories. It addresses the recognition problem in which no labeled instance is available for some classes but some sort of description that is called *class signature* is available for all categories. Example of class signatures include human-annotated discriminative attributes or textual description of the categories. The problem addressed by zero-shot learning rises naturally in practice wherever it is not feasible to acquire abundant labeled instances for all categories (e.g., fine-grained classification problems). To describe the task more precisely, in the training phase, labeled instances for some categories called seen classes are provided while for other categories, called unseen ones, there is no labeled instance available. In the test phase, unlabeled instances should be classified into seen or unseen categories. In this work, we focus on the most popular version of zero-shot recognition in which test instances belong only to unseen categories. Most existing methods for zero-shot learning focus on using labeled instances (i.e. images) to learn a compatibility function indicating how similar an instance is to each label embedding obtained from class signatures [3, 26, 35]. Each instance will then be labeled with the category having the most compatible signature. On the other hand, recent advances in deep neural networks provide rich visual features with high discrimination capability [28]. We found out through experiments that the space of deep visual features is a rich space in which instances of different categories usually form natural clusters. Nonetheless, little attention has been paid to exploiting this property of deep visual features in the context of zero-shot learning.

In this paper, we propose a semi-supervised zero-shot learning method called *Joint Embedding and Clustering* (JEaC) that uses both labeled instances of seen classes and unlabeled instances of unseen classes to find a more proper representation of class signatures in the space of deep visual features. In this method, a linear transformation is learned to map the class signatures to the space of abstract visual features and jointly assignments of unlabeled samples to unseen classes are found (Figure 1). The linear mapping is learned so that the mapped signature of each seen class tends to be representative for samples of the corresponding class and simultaneously assignments of unlabeled samples to unseen classes are leaned such that the mapped signature of each unseen class also tends to be representative of assigned samples to that class. Using unlabeled samples from unseen classes, we can substantially mitigate the *domain shift problem* introduced in [11] that impairs the zero-shot recognition performance. We also propose a simpler method called *Independent Embedding and Clustering* (IEaC) in which label assignment and class embedding in the image space are not learned jointly. Instead, after finding the mapping according to just instances of seen classes, a clustering algorithm is used to assign labels to instances of unseen classes.

We present experimental results on four popular zero-shot classification benchmarks and see that the proposed

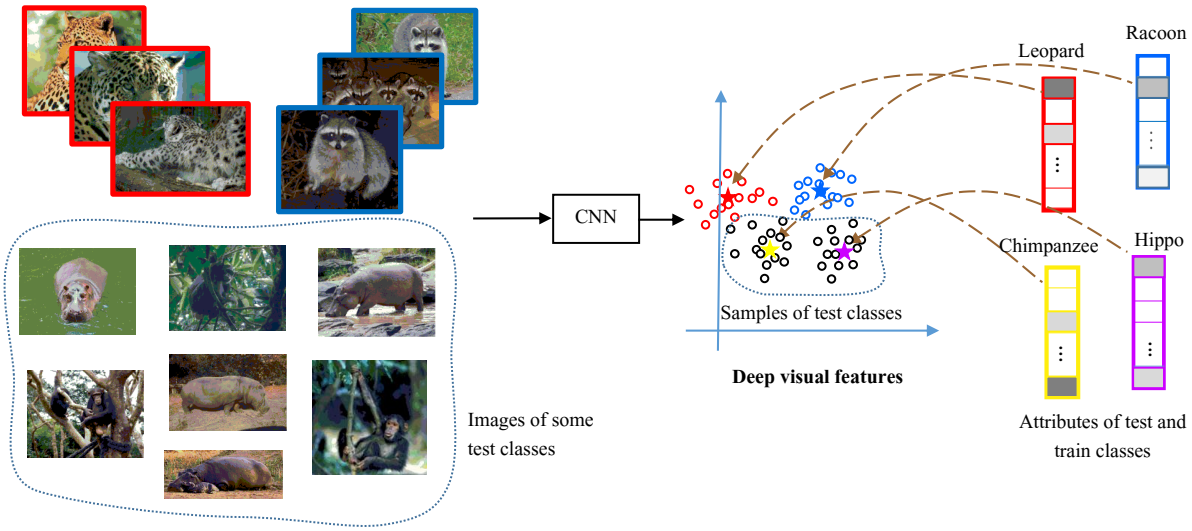

Figure 1: The proposed method maps attribute vectors of classes to the space of deep visual features such that the mapped attribute vectors are good representatives for the samples of the corresponding classes. The mapped attributes are depicted by stars.

method outperforms state-of-the-art methods on these datasets.

2. Related Work

A notable body of work in zero-shot recognition belongs to attribute prediction from images [16, 34, 20, 31, 29]. In these methods, the semantic label embeddings are considered to be externally provided attributes. Thus, label embeddings are already available and the task is just to map images to the semantic space, i.e. predicting attributes for the images. Early methods, like [16] do not consider dependency of attributes and train binary attribute classifiers. Probabilistic graphical models have been utilized to model and/or learn correlations among different attributes [34, 31] to improve the prediction. In [14], a random forest approach has been employed that accounts for unreliability in attribute predictions for final class assignment. In [2], a max-margin objective function is defined for attribute-based image classification.

More recent works often exploit bilinear models [33, 10, 22, 35, 26, 19]. Several objective functions have been proposed for learning such bilinear models. In [26], the sum of the squared error on the label prediction is used and clever regularization terms that compensate undesirable characteristics of this cost function are also utilized. In [18, 19], a max margin objective function for learning the bilinear mapping is used. These two methods learn labels for test instances simultaneously and so they differ from almost all of other existing methods in this way. They provide the possibility of leveraging unsupervised information available in test images. In [19], the distribution of unlabeled instances is entered through using a Laplacian regularization term that penalizes similar objects assigned to different classes.

Designing label embeddings in the multi-class classification problem is another line of research that can also be used in zero-shot recognition. In [33], an objective function is proposed to derive such label embeddings based on information about similarities among categories. A relatively popular embedding for labels is to describe unseen categories as how similar they are to the seen ones. In [22], the outputs from the softmax layer of a CNN trained on seen categories are used to score similarity between test instances and seen classes. Using these outputs as weights, images are represented in the semantic space as a convex combination of seen class labels embedding. In [35], a histogram showing seen class proportions is used for label embedding and then a max margin framework is defined to embed images in this space. This work is further extended in [36] and a supervised dictionary learning formulation is presented to jointly learn embedding for images and labels. The idea of combining already available classifiers to create new ones for unseen categories is also used in [6] but rather than using seen categories as basis, they define a set of (possibly smaller) *phantom* classes and learn base classifiers on them.

Although most of the works on zero-shot recognition consider attributes as auxiliary information, textual descriptions or name of the categories can also be utilized as class signatures. [10] introduce a bilinear model to find the compatibility score of deep visual features and Word2vec [21] representation of class names. [5] proposes nonlinear map-

pings modeled by neural networks on the image and the text inputs to find their compatibility. [8] presents an objective function to predict classifier parameters from textual descriptions. In [3], some different label embeddings and also a combination of these embeddings have been considered along with a bilinear compatibility function. In [32], this work is extended further to model nonlinear compatibility functions that can be expressed as a mixture of bilinear models. In [12], a set of vocabulary much larger than just seen and unseen class names is used and mapping from images to word embeddings is learned by maximizing the margin with respect to all words in the vocabulary; this framework can be used in zero-shot and also supervised and open set learning problems. In [1], authors propose to use multiple auxiliary information and also semantic part annotations in the image domain to compensate for weaker supervision in textual data. Convolutional and recurrent neural networks have also been used for text embedding in [27].

The most related methods to ours are [18, 15, 19] that are semi-supervised zero-shot learning methods. Here, we briefly specify the differences between these methods and ours. First, we use abstract visual features obtained by deep learning as the semantic space as opposed to [18, 19]. These two methods learn a max margin classifier on the image space classifying both seen and unseen instances while we propose a clustering-based approach in the semantic space of deep visual features that uses a ridge regression to map signatures to this semantic space. Our approach results in a much simpler optimization problem to solve. We also explicitly account for domain shift problem in our objective function and thus achieve better results compared to these methods. There are major differences between our work and [15] using a dictionary learning scheme in which coding coefficients are considered to be label embeddings (attribute vectors) and a sparse coding objective is used to map images into this representation space. Most importantly, in our method, labels of unseen instances are jointly learned with the mapping of the signatures to the semantic space while in [15] the label prediction is accomplished using the nearest neighbor or the label propagation on embeddings of images. Moreover, we do not need to learn embedding of test instances in the semantic space as opposed to [15], alternatively we learn just the representation of class signatures in the visual domain.

3. Proposed Method

In this section, we introduce two zero-shot learning methods that use deep visual features as the semantic space and learn a mapping from class signatures to this semantic space. First, we propose Independent Embedding and Clustering (IEaC) as a simple and efficient semi-supervised zero-shot learning method. Then, we further extend our method to jointly learn the embedding of signatures and class assignments of unlabeled samples. We call this method Joint Embedding and Clustering (JEaC). We formulate an optimization problem for JEaC and present an iterative method to solve it. IEaC is used to find a proper starting point for this optimization procedure (i.e. labels found by IEaC for instances of unseen classes are considered as initial label assignment to these instances).

3.1. Notation

Let $X, \mathbf{x}$, and x denote matrices, column vectors, and scalars respectively. $\|X\|_F^2$ shows the squared Frobenius norm of a matrix and $X_{(i)}$ denotes its ith column. $\mathbf{1}_k$ denotes a column vector whose k-th element is one and other elements are zero. Suppose there are n_s seen categories and n_u unseen categories. For each category y, auxiliary information $a_y \in \mathbb{R}^r$ is available. We assume that labels $\{1, \ldots, n_s\}$ correspond to seen categories.

Let $X_s \in \mathbb{R}^{d \times N_s}$ and $X_u \in \mathbb{R}^{d \times N_u}$ denote matrices whose columns are seen and unseen images respectively where d is the dimension of image features and N_s (or N_u) shows the whole number of the images in the seen (or unseen) classes. $S_s = [a_1, \ldots, a_{n_s}]$ presents the matrix of signatures for seen classes and S_u is defined similarly for unseen classes. $Z_s = [\mathbf{z}_1, \ldots, \mathbf{z}_{N_s}]$ contains labels of training data in the one-hot encoding format. $\mathbf{r}_n$ also denotes the label assigned to the n-th instance by our algorithm in the one-hot coding format.

3.2. Independent Embedding and Clustering

Our first method can be roughly summarized in three steps:

1. Using data from seen classes, we learn a linear mapping from attribute vectors to the semantic space.

2. We find a data clustering using our proposed semi-supervised clustering algorithm.

3. For each cluster, we find the label whose mapped signature in the semantic visual space is the nearest one to the center of that cluster and assign the corresponding label to all of its instances.

We use a simple ridge regression to map class signatures to deep visual features. We intend to a mapping such that each mapped (seen) class signature is close to the samples of that class in this space in average. The linear mapping is found using the following optimization problem:

$$W^* = \arg\min_W \|X_s - WY_s\|_F^2 + \gamma \|W\|_F^2, \qquad (1)$$

where columns of $Y_s \in \mathbb{R}^{r \times n_s}$ are the class signatures of the samples lied in the columns of X_s (i.e. $Y_s = S_s L$ where columns of L contain the one-of-n_s encoding of the labels

for instances of seen classes). This optimization problem is known to have the following closed form solution:

$$W = X_s Y_s^T (Y_s Y_s^T + \gamma I)^{-1}. \quad (2)$$

The parameter γ is determined through cross validation as we will describe in the Experiments section.

Here, we intend to find labels for the instances belonging to unseen classes. To this end, we find a clustering of instances in the space of deep visual features and then assign a label to each cluster according to the distance between the center of that cluster and the mapped signatures of unseen classes. The label whose mapped signature is the closest one to the cluster center is selected to be assigned to all instances in the cluster. More precisely, let $\boldsymbol{\mu}_k$ be the center of the k-th cluster and $c(\mathbf{x}_n)$ denote the cluster number to which $\mathbf{x}_n$ is assigned. The label assigned to $\mathbf{x}_n$ in our method would be:

$$\underset{i=n_s+1,\ldots,n_u+n_s}{\arg\min} \left\| W(Y_s)_{(i)} - \boldsymbol{\mu}_{c(\mathbf{x}_n)} \right\|_2^2. \quad (3)$$

To find a better clustering of instances belonging to unseen classes, we can also incorporate labeled instances of seen classes. The clustering problem over unseen instances with which we encounter here is different from the conventional semi-supervised learning problem [7]. Here, all labeled data are from seen classes and there is no labeled sample for unseen classes that is due to the special characteristic of zero-shot learning problem. Therefore, we propose a semi-supervised clustering method which can be seen as an extension of k-means that is properly adapted for zero-shot learning problem. We try to find a clustering such that labeled instances tend to be assigned to the corresponding classes and all instances tend to be close to the center of the clusters to which they are assigned:

$$\min_{R,\boldsymbol{\mu}_1,\ldots,\boldsymbol{\mu}_k} \sum_{n,k} r_{nk} \|\mathbf{x}_n - \mu_k\|_2^2 + \beta \sum_{n=1}^{N_s} \mathbb{1}(\mathbf{r}_n \neq \mathbf{z}_n), \quad (4)$$

where $\boldsymbol{\mu_i}$s are cluster centers and $R = [\mathbf{r}_1, \ldots, \mathbf{r}_{N_s+N_u}]$ contains cluster assignments of all instances in one-hot encoding format. The objective function is similar to that of the k-means clustering algorithm but for each labeled instance there is a penalty of β if the assigned cluster number is different from its label. Thus, this objective function encourages the first n_s clusters be corresponding to the seen classes.

Parameters β and k can be determined via cross validation. However, in our experiments, we found out the model is not very sensitive to them so we fix $\beta = 1$ when data have been normalized. Moreover, we set $k = (n_s + 2n_u)$ which can allow more than one cluster per class for unseen categories. This, to some extent, copes with diversity of instances in a class.

To solve the optimization problem in Eq. 4, we use an iterative procedure (similar to k-meams) shown in Algorithm 1. In each iteration, μ_is are updated as:

$$\mu_i = \frac{\sum_{n=1}^{N_s+N_u} \mathbb{1}(r_{ni} = 1)\mathbf{x}_n}{\sum_{n=1}^{N_s+N_u} \mathbb{1}(r_{ni} = 1)}, \quad (5)$$

and instances of unseen classes are assigned to the nearest cluster to them.

To initialize $\boldsymbol{\mu}_i$'s, for clusters corresponding to seen classes the centers are set as the mean of instances of those classes. Centers of other clusters are initialized using k-means++ [4] on unlabeled instances.

A key distinction between the clustering-based method presented here and other existing methods lies in the nature of the compatibility function. The compatibility function in other works is usually a similarity measure between each instance and class description. This measure is found independently for different instances. Here, the compatibility function relies strongly on the distribution of instances in the semantic space and the compatibility of a label for an instance is found according to the similarity of the cluster center to which this instance is assigned and the mapped signature of that label. Therefore, by considering the distribution of data samples (via clustering) in designing the compatibility function, we can reach a more reliable measure.

Although the above method outperforms the state-of-the-art methods on most zero-shot recognition benchmarks, it uses only instances of the seen classes to find the linear transformation from class signatures to the visual feature space and thus the proposed method may suffer from the domain shift problem introduced in [11]. To overcome the domain shift problem more substantially, in the next subsection, we propose an optimization problem for finding the linear transformation from class signatures to the visual feature space that uses instances of both seen and unseen classes.

3.3. Joint Embedding and Clustering

In this section, we present an optimization problem for jointly learning a linear transformation from class signatures to the visual feature space and assigning of unlabeled instances to unseen classes. Accordingly, we seek a linear transformation (and simultaneously cluster assignments of unseen data) such that the mapped signatures are good representatives of the assigned instances to the corresponding classes. We then present an iterative procedure to solve the proposed optimization problem.

Algorithm 1: Training Procedure of IEaC

input : X_s, Y_s, Z_s, X_u, S_u
output: Z_u (label predictions for X_u)

Initialize $\boldsymbol{\mu}_k$ by Eq. (5), $k = 1, \ldots, n_s$;
Initialize $\boldsymbol{\mu}_k$ by kmeans++, $k = n_s + 1, \ldots, n_s + n_u$;
repeat
  $c_n \leftarrow \arg\min_i \|x_n - \mu_i\|_2, \quad n \in \{1, 2, \ldots, N_s + N_u\}$;
  $\boldsymbol{\mu}_k \leftarrow \frac{\sum_n \mathbf{x_n} \mathbb{1}(c_n = k)}{\sum_n (\mathbb{1}(c_n = k)}, \quad k \in \{n_s + 1, 2, \ldots, n_s + n_u\}$;
until *convergence to local minimum*;
$W \leftarrow X_s Y_s^T (Y_s Y_s^T + \gamma I)^{-1}$;
// array l maps cluster numbers to labels
$l[k] \leftarrow \arg\min_j \|\boldsymbol{\mu_k} - (W S_u)_{(j)}\|_2$;
$(\mathbf{Z_u})_{(\mathbf{n})} \leftarrow \mathbf{1}_{l[c_n]}$;

The objective function of JEaC is formulated as follows:

$$\min_{R,W} \|X_s - W Y_s\|_F^2 + \lambda \left\|X_u - W S_u R^T\right\|_F^2 + \gamma \|W\|_F^2, \quad (6)$$

$$\text{s. t.} \quad R \in \{0,1\}^{N_u \times n_u}.$$

where R shows the cluster assignment of unlabeled instances in one-hot encoding format. The first term in the above optimization problem is identical to Eq.(1) and the second one incorporates unlabeled data for learning the mapping W. Therefore, we seek class assignments for instances of unseen classes and simultaneously learn a mapping on the class signatures such that the mapped signatures provide good representatives for instances of both the seen and unseen classes. By enforcing the signatures of unseen classes to be mapped close to a set of test instances, the second term helps us to confront the domain shift problem. More precisely, in the above problem, for the seen classes, the sum of the squared distances of instances from the mapped signature of the corresponding class is minimized. For instances of unseen classes, the mapping and class assignment are jointly learned to minimize such term (i.e. sum of their squared distances from representation of the class they are assigned to). The second term in Eq. 6 can be essentially considered as a clustering objective with two additional advantages. First, the number of clusters is no longer a parameter and is determined by the number of unseen classes. Second, the cluster centers are set to be the mapped signatures of the test classes and thus we can learn mapping W and class assignments R jointly.

3.3.1 Optimization

The optimization problem in Eq. (6) is not convex and considering that R is a partitioning of instances, the global optimization requires an exhaustive search over all possible

Algorithm 2: Training Procedure for JEaC

input : X_s, Y_s, Z_s, X_u, S_u
output: Z_u (label predictions for X_u)

Initialize R by output of Algorithm 1 ;
repeat
  update W by Eq. (7) ;
  updata R by Eq. (8) ;
until *no element of R changes*;
output $Z_u \leftarrow R$

labeling of test data with n_u labels. Therefore, we use a simple coordinate descent method (like k-means). We alternate between optimizing R and W while keeping the other one fixed. Having fixed the labeling R, the problem becomes a simple multi-task ridge regression which has the following closed-form solution:

$$W = (X_s Y_s^T + \lambda X_u R S_u^T)(Y_s Y_s^T + \lambda S_u R^T R S_u^T + \gamma I)^{-1}. \quad (7)$$

By fixing W, the optimal R can be achieved via assigning each instance to the closest class representative:

$$r_{ij} = \mathbb{1}[j = \arg\min_k \|X_{u(i)} - W S_{u(k)}\|_2]. \quad (8)$$

Whenever a row of R contains no ones, i.e. an empty cluster appears, a percentage of instances are randomly assigned to that cluster. We continue alternating between updates of W and R till R remains constant, i.e. no label changes. To evade poor local minima, we use a good starting point that initializes R with cluster assignments found by IEaC.

4. Experiments

In this section, we conduct experiments on the popular benchmarks to obtain results of the proposed method on these benchmarks and compare them with those of the other recent methods.

4.1. Setup

Datasets. We evaluate our methods on four popular public benchmarks for zero-shot classification. (1) Animal with Attributes (AwA) [16]: There are images of 50 mammal species in this dataset. Each class is described by a single 85−dimensional attribute vector. We use the continuous attributes rather than the binary ones as it has been proved to be more discriminative in previous works like [3]. The train/test split provided by the dataset is used accordingly. (2) aPascal/aYahoo [9]: 20 categories from Pascal VOC 2008 [13] are considered as seen classes and categories from aYahoo are considered to be unseen. As this dataset provides instance level attribute vectors, for class signatures we use the average of the provided instance attributes. (3) SUN Attribute [24]: The dataset consists of 717

Table 1: Adjusted Rand Index for a simple k-means clustering on test instances from four benchmark datsets.

Clustering Method	Animals with Attributes	CUB-2011	aPascal-aYahoo	SUN Attribute
k-means	76.99	38.59	77.85	55.49
Proposed Clustering	**70.74±0.32**	**42.63±0.07**	**69.93± 3.4**	**45.50±1.32**

Table 2: Classification accuracy (in percent) on four public datasets: Animals with Attributes, CUB-2011, aPascal-aYahoo, and SUN in the form of average ± std.

Feature	Method	Animals with Attributes	CUB-2011	aPascal-aYahoo	SUN
Shallow	Li and Guo [18]	38.2±2.3			18.9±2.5
	Li *et al*. [19]	40.05±2.25		24.71 ±3.19	
	Jayaraman and Grauman [14]	43.01 ± 0.07		26.02 ± 0.05	56.18 ± 0.27
GoogleNet	Akata *et al*. [3]	66.7	50.1		
	Changpinyo *et al*. [6]	72.9	54.5		62.7
	Xian *et al*. [32]	71.9	45.5		
VGG-19	Khodirov *et al*. [15]	73.2	39.5	26.5	
	Akata *et al*. [3]	61.9	50.1		
	Zhang and Saligrama [35]	76.33±0.53	30.41 ±0.20	46.23 ± 0.53	82.50 ± 1.32
	Zhang and Saligrama [36]	80.46±0.53	42.11 ±0.55	50.35 ± 2.97	83.83 ± 0.29
	IEaC (k-means)	86.34±0.13	52.48±0.60	48.03±1.56	75.75±1.06
	IEaC (Proposed Clustering)	90.52±0.74	53.10±0.43	48.00±0.69	80.66±0.76
	JEaC (init W)	91.29	57.55	43.36	70.10
	JEaC (init R)	***92.98 ±0.14***	***60.42±0.72***	**53.20±3.12**	***85.33±0.57***

categories and all images are annotated with 102 attributes. We just use the average attributes among all instances of each categories for our experiments and the same train/test spilt as in [14] where 10 classes have been considered as unseen. (4) Caltech UCSD Birds-2011 (CUB) [30]: This is a dataset for fine-grained classification task. There are 200 species of birds where each image has been annotated with 312 binary attributes. Again, we average over instances to get continuous class signatures. The same train/test split as in [2] (and many other following works) is used to make comparison possible.

Visual features. As our method relies on meaningful structure in visual features domain, we use features from a deep CNN known to be more discriminative than *shallow* features like SIFT or HOG. We report results using $4096-$dimensional features from the first fully connected layer of 19 layer VGG network [28] pre-trained on imagenet, provided publicly by [35].

Cross Validation. To adjust parameters γ and β in Eq. (7) and parameter γ in Eq. (2), training data are partitioned into train and validation sets. We choose a number of categories randomly from training data as validation categories. For each dataset, the ratio of the categories in the validation set to those in the whole train set is the sames as the ratio of the test categories to the total of the train and test ones. In our experiments, we used $10-$fold cross validation, i.e. average results from ten different validation splits are used to decide on optimal parameters. Once optimal values for parameters are determined through the grid search by testing on validation set, the model is then trained on all seen categories.

4.2. Experimental Results

Clustering experiment. First, to give evidence for the key assumption of our method that samples of each class usually form a cluster in the space of deep visual features, we conduct an experiment in which samples of unseen categories (in the space of deep visual features) are clustered using a simple kmeans algorithm. The number of clusters in this experiment is set to the number of unseen classes. The kmeans implementation available in Scikit-learn library [25] is used and the algorithm is run with 20 different initializations and the best score according to the cost function of kmeans is selected. To evaluate clustering results, the adjusted Rand Index is used and the obtained results are reported in Table 1. These results show that the cluster structure assumption in the deep visual space is valid to some extent and thus we can use this assumption for the unsupervised task of labeling samples of unseen classes.

Compared methods The proposed methods are compared with the most recent methods of zero-shot recognition shown in Table 2. Among these methods, [19, 18] are semi-supervised zero-shot learning methods that use unlabeled samples of unseen classes during the training. For other methods, we use the results reported in their original papers. Since we did not re-implement any of the other

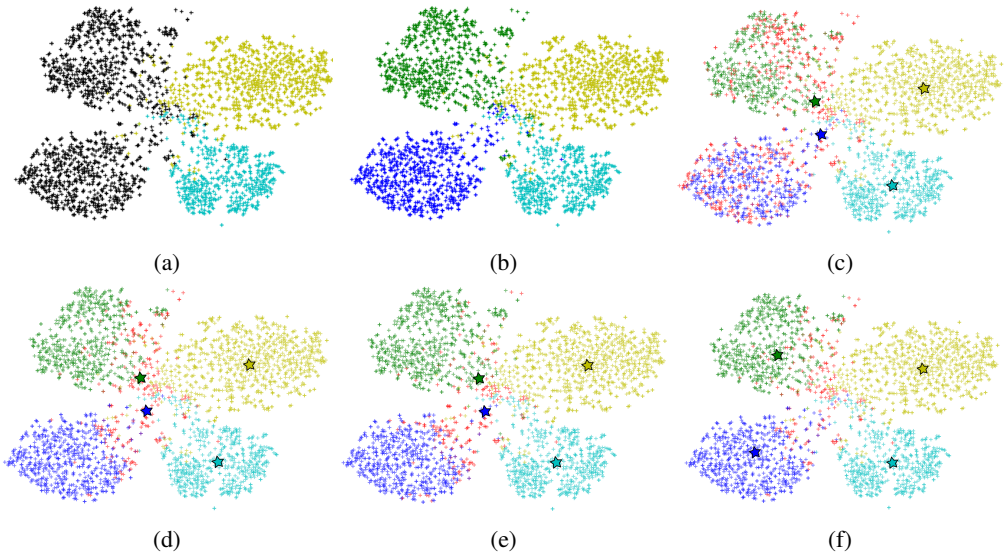

Figure 2: t-SNE embedding of the deep features for samples of four classes from AwA, two seen: antelope (cyan), grizzly bear (yellow) and two unseen: chimpanzee (blue), giant panda (green). Image samples are shown by plus signs and embedding of class signatures by stars. In Figs. c-f, red points denote assignment to a class other the four classes considered here. In Figs. c-e, the linear transformation in Eq. (2) is used to map signatures to the visual feature space. **a)** Data of seen classes are colored and those of unseen classes are black, **b)** Samples colored according to their ground truth labels, **c)** Classification by finding the nearest mapped signature, **d)** Classification done by our compatibility function on cluster assignments found by k-means, **e)** Classification by our compatibility function using our semi-supervised clustering, **f)** Cluster assignment and mapped signatures found by JEaC .

methods, if the original paper does not report results on a dataset we leave the corresponding cell as blank.

In Table 2, *IEaC (K-means)* and *IEaC (Proposed Clustering)* are the two versions of the IEaC method proposed in Section 3.2. In fact, *IEaC (K-means)* uses the simple k-means clustering on samples of unseen classes and then utilizes Eq. 3 to assign labels to the clustered samples. However, *IEaC (Proposed Clustering)* that has been shown in Algorithm 1 performs the proposed semi-supervised clustering instead of the simple kmeans clustering. For our JEaC method, we introduced an iterative procedure in Section 3.3.1 that starts by initializing R. Nonetheless, we can start by initializing either W or R. *JEaC(init W)* corresponds to initializing W using Eq. (2) and *JEaC(init R)* denotes JEaC that initializes R by the output of IEaC.

Results. Results of the compared methods have been summarized in Table 2. For our methods, the average and the standard deviation of different runs have been reported in Table 2. As it can be seen, the initialization of R done by IEaC has an important role on the performance of JEaC. This can be justified by noting that the information about the distribution of unlabeled data is leveraged when initializing R (by the cluster assignment R found in our IEaC method) while such information is absent in initializing W as in Eq. 2. According to the results in Table 2, our JEaC (init R) method outperforms the other methods on all the four benchmarks. Although our IEaC method performs generally better than the existing zero-shot learning methods, our JEaC (init R) method outperforms IEaC too.

The effectiveness of different components of our methods is further illustrated in Figure 2. As it can be seen in Figure 2c merely using mapping from Eq. (2) results in poor signature embeddings and domain shift problem occurs. However, using the compatibility function proposed in Eq. 3 that is based on clustering of unseen data improves label assignments (Figure 2d) although there is no change in the mapped signatures. Label assignments can also slightly be improved when the clustering method proposed in Section 3.2 is used (Figure 2e) instead of k-means (Figure 2d). Finally, as Figure 2f shows using the mapping and class assignments found by JEaC, the domain shift problem is substantially alleviated and better signature embeddings are achieved.

More Analysis. We conduct more experiments to show that the performance of IEaC is not such sensitive to the number of clusters. In the aforementioned experiments, we had set the number of clusters to $n_s + 2n_u$ for this method. However, according to Fig. **??**, we see that the IEaC performance is not such sensitive to the number of clusters and for a rather wide range of numbers results do not change significantly. Moreover, it is worth to mention that in our JEaC method number of clusters is no longer a parameter since we directly assign samples to the unseen classes in this method.

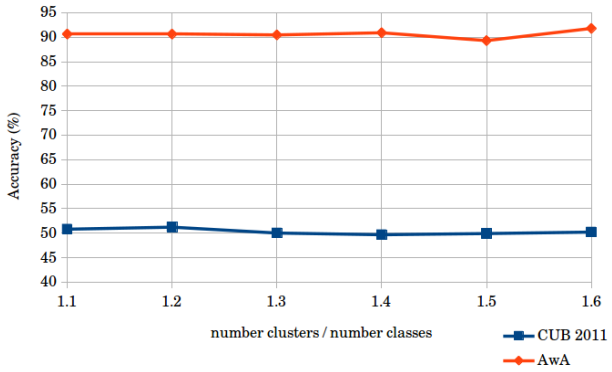

Figure 3: Effect of number of clusters on multi class prediction accuracy for AwA and CUB 2011

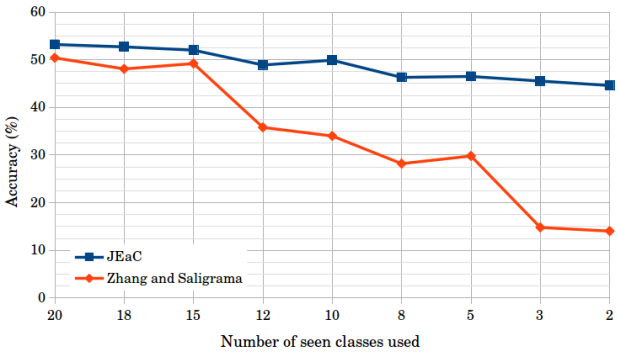

Figure 4: Comparing effect of number on seen classed used in training on zero-shot prediction accuracy in aPascal/aYahoo dataset between JEaC and Zhang and Saligrama [36].

Since the performance of zero-shot classification methods depends on the number of seen classes, we intend to evaluate how much the performance of the proposed method degrades by decreasing the number of seen classes. We compare performance of our JEaC method and those of [36] (having highest accuracy among comparing methods in Table 2) on aPascall/aYahoo dataset while fixing unseen categories and varying the number of seen categories used from the dataset. The results are presented in Fig. **??** showing that JEaC is far less sensitive to number of seen categories and the gap between performance of JEaC and that of [36] widens when decreasing number of seen categories used in training.

5. Conclusion

In this paper, we proposed semi-supervised methods for zero-shot object recognition. We used the space of deep visual features as a semantic visual space and learned a linear transformation to map class signatures to this space such that the mapped signatures provide good representative of the corresponding instances. We utilized this property that the rich deep visual features provide a representation space in which samples of each class are usually condensed in a cluster. In the proposed method that jointly learns the mapping of class signatures and the class assignments of unlabeled data, we used also unlabeled instances of unseen classes when learning the mapping to alleviate the domain shift problem. Experimental results showed that the proposed method outperformed other recent methods with a large margin.

References

[1] Z. Akata, M. Malinowski, M. Fritz, and B. Schiele. Multi-Cue Zero-Shot Learning with Strong Supervision. In *CVPR*, pages 59–68, 2016. 3

[2] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013. 2, 6

[3] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of Output Embeddings for Fine-Grained Image Classification. In *CVPR*, pages 2927–2936, 2015. 1, 3, 5, 6

[4] D. Arthur and S. Vassilvitskii. k-means++: the advantages of careful seeding. In *Eighteenth Annual ACM-SIAM symposium on Discrete Algorithms*, pages 1027–1035, 2007. 4

[5] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting Deep Zero-Shot Convolutional Neural Networks using Textual Descriptions. In *ICCV*, pages 4247–4255, 2015. 2

[6] S. Changpinyo, W. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. pages 5327–5336, 2016. 2, 6

[7] O. Chapelle, B. Schölkopf, and A. Zien. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006. 4

[8] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, pages 2584–2591, 2013. 3

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by Their Attributes. In *CVPR*, pages 1778–1785, 2009. 1, 5

[10] 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*, pages 2121–2129, 2013. 2

[11] Y. Fu, T. M. Hospedales, T. Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, volume 6315. 2014. 1, 4

[12] Y. Fu and L. Sigal. Semi-supervised Vocabulary-informed Learning. In *CVPR*, pages 5337–5346. 3

[13] D. Hoiem, S. K. Divvala, and J. H. Hays. Pascal voc 2008 challenge, 2008. 5

[14] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *NIPS*, pages 3464–3472. 2014. 2, 6

[15] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised Domain Adaptation for Zero-Shot Learning. In *ICCV*, pages 2452–2460, 2015. 3, 6

[16] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009. 1, 2, 5

[17] H. Larochelle, D. Erhan, and Y. Bengio. Zero-data learning of new tasks. In *AAAI*, pages 646–651, 2008. 1

[18] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *AISTATS*, pages 626–634, 2015. 2, 3, 6

[19] X. Li, Y. Guo, and D. Schuurmans. Semi-Supervised Zero-Shot Classification with Label Representation Learning. In *ICCV*, pages 4211–4219, 2015. 2, 3, 6

[20] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, pages 1227–1234, 2011. 2

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119. 2013. 2

[22] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014. 2

[23] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418. 2009. 1

[24] G. Patterson, C. Xu, H. Su, and J. Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014. 5

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 6

[26] B. Romera-Paredes and P. H. S. Torr. An Embarrassingly Simple Approach to Zero-shot Learning. In *ICML*, pages 2152–2161, 2015. 1, 2

[27] B. S. Scott Reed, Zeynep Akata, Honglak Lee. Learning Deep Representations of Fine-Grained Visual Descriptions. In *CVPR*, pages 49–58, 2016. 3

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1, 6

[29] M. Suzuki, H. Sato, S. Oyama, and M. Kurihara. Transfer learning based on the observation probability of each attribute. In *Systems, Man and Cybernetics (SMC), IEEE International Conference on*, pages 3627–3631, 2014. 2

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 6

[31] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *ICCV*, pages 2120–2127, 2013. 2

[32] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent Embeddings for Zero-shot Classification. In *CVPR*, pages 69–77, mar 2016. 3, 6

[33] F. X. Yu, L. Cao, R. S. Feris, J. R. Smith, and S.-F. Chang. Designing Category-Level Attributes for Discriminative Visual Recognition. In *CVPR*, pages 771–778, 2013. 2

[34] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, volume 6315, pages 127–140. 2010. 2

[35] Z. Zhang and V. Saligrama. Zero-Shot Learning via Semantic Similarity Embedding. In *ICCV*, 2015. 1, 2, 6

[36] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, pages 6034–6042, 2016. 2, 6, 8