Learning Discriminative Latent Attributes for Zero-Shot Classification

Huajie Jiang[1,2,3], Ruiping Wang[1], Shiguang Shan[1], Yi Yang[4], Xilin Chen[1]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]Shanghai Institute of Microsystem and Information Technology, CAS, Shanghai, 200050, China
[3]ShanghaiTech University, Shanghai, 200031, China
[4]Huawei Technologies Co., Ltd., Beijing, 100085, China
huajie.jiang@vipl.ict.ac.cn, {wangruiping, sgshan, xlchen}@ict.ac.cn, yangyi16@huawei.com

Abstract

*Zero-shot learning (**ZSL**) aims to transfer knowledge from observed classes to the unseen classes, based on the assumption that both the seen and unseen classes share a common semantic space, among which attributes enjoy a great popularity. However, few works study whether the human-designed semantic attributes are discriminative enough to recognize different classes. Moreover, attributes are often correlated with each other, which makes it less desirable to learn each attribute independently. In this paper, we propose to learn a latent attribute space, which is not only discriminative but also semantic-preserving, to perform the ZSL task. Specifically, a dictionary learning framework is exploited to connect the latent attribute space with attribute space and similarity space. Extensive experiments on four benchmark datasets show the effectiveness of the proposed approach.*

1. Introduction

Visual recognition has made tremendous progress in the past few years with the rapid growing of data scales and progressing of classification methods. However, traditional approaches to visual recognition are mainly based on supervised learning, which needs large numbers of labeled samples to obtain a high performance classification model. It is well known that collecting large scale of labeled samples is difficult, especially when the required labels are fine-grained, which hinders further development of visual recognition. It is therefore important and desirable to develop recognition systems that can recognize categories with few or no labeled samples, thus ZSL approaches attract more and more attentions in the past few years.

Inspired by the ability of humans to recognize unseen objects, ZSL aims to recognize categories that have never been seen before [17, 9]. A general assumption for ZSL is that both the seen and unseen classes share a common semantic space, where the samples and class prototypes are projected, to perform the recognition task. In terms of different mid-level representations exploited in the learning process, current ZSL approaches can be categorized into four groups. The first group is **attribute-based** methods, which use attributes to build up the relationships between the seen and unseen classes [17, 9, 43, 13]. For example, attributes, such as *black* and *furry*, are shared among different animals. The cross-category property of attributes makes it possible to transfer knowledge from the seen classes to the unseen classes. The second group is **text-based** approaches, which automatically mine the relationships of different classes via ample text corpus [7, 10, 3, 30]. These approaches reduce the human labor for defining attributes, thus ZSL can be applied to large scale settings. The third one is based on **class similarity**, which directly mines the similarities between the seen and unseen classes to bridge up their relationships [23, 22, 24, 44]. The similarities may be derived from the hierarchical category structure or from the semantic descriptions of each class. The last group is to **combine different mid-level representations** to learn more robust relationships [11, 12, 19, 15]. These works build upon a common idea that different mid-level representations will catch complementary information of the data, which can be used to reduce the domain difference between the seen and unseen classes.

In this paper, we focus on the **attribute-based** approach. Traditional process for such methods mainly focuses on how to learn the semantic embeddings or what strategy to utilize to perform the recognition task. However, there are three aspects which are merely considered in previous works, as is shown in Figure 1. First, whether the human-designed semantic attributes are discriminative enough to recognize different classes. Second, whether it is reasonable to learn each attribute independently since attributes

2380-7504/17 $31.00 © 2017 IEEE
DOI 10.1109/ICCV.2017.453

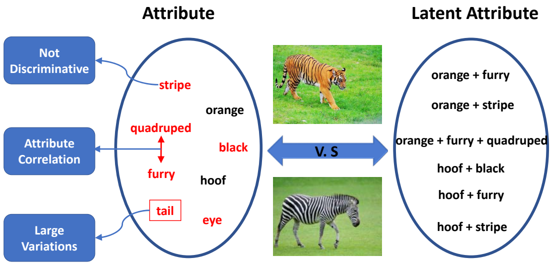

Figure 1. Motivations for learning latent attributes. Attributes in red font are not discriminative. Attributes with rectangle have large variations. Attributes connected by the double arrow are correlated with each other.

are often correlated with each other. Third, the variations within each attribute may be quite large making it difficult to learn the attribute classifiers. For the first aspect, [43] proposes to learn discriminative category-level attributes. However, these attributes are learned on fixed categories and do not care semantic meanings. When new classes appear, the class-level representations have to be relearned. For the second aspect, [14] incorporates the attribute relationships into the learning process. However, such relationships are human-defined and they are usually too complex in real world to define beforehand. For the third aspect, [15] utilizes domain adaptation approaches to finetune the attribute models. However, the target domain samples are mandatory for such models.

In order to tackle the problems described above simultaneously, we propose to learn latent attributes. Specifically, the proposed method automatically explores discriminative combinations of different attributes, where each combination is viewed as a latent attribute. On the one hand, the latent attributes are required to be discriminative enough, thus to classify different classes more reliably. On the other hand, the latent attributes should be semantic-preserving, thus to enable building up the relationships between different classes. Moreover, the attribute correlations are also implicitly considered in latent attributes. For example, *furry* often correlates with *black* and *white*, thus it is not favorable to learn *furry* alone. In contrast, our latent attributes have the ability to find the combination of *furry* + *black* and *furry* + *white*, thus the variation within each latent attribute becomes smaller than that within each attribute.

To learn the latent attribute space for performing ZSL task, we exploit dictionary learning framework to directly model the latent attribute space, where the images can be reconstructed by some latent attribute dictionary items. In order to preserve the semantic information, a linear transformation is utilized to build up the relationships between attributes and latent attributes, thus the latent attributes can be viewed as different combinations of attributes. Moreover, to make the latent attributes discriminative, seen class classifiers are utilized to classify different classes, where the probability outputs can be viewed as similarities to seen classes. Thus we can transform the image representations from the latent attribute space to the similarity space.

The rest of the paper is organised as follows: Section 2 discusses the related works. Section 3 describes the formulation and optimization of our proposed latent attribute dictionary (**LAD**) approach in detail. Section 4 extensively evaluates the proposed method on four benchmark ZSL datasets. Section 5 gives concluding remarks.

2. Related Work

In this section, we briefly review the related works on attributes and zero-shot learning.

2.1. Attributes

Attributes are general descriptions of images and have drawn much attention in different computer vision tasks in recent years, such as image description [9], image captioning [16], image retrieval [35] and image classification [41, 21, 27]. Earlier works on attribute learning often consider it as a binary classification problem and learn each attribute independently [9]. Due to the fact that attributes are often correlated with each other, [14] incorporates the attribute relationships into the learning framework. Moreover, attributes are related to the categories, [21, 1] propose to learn attributes and the class labels jointly. As deep learning becomes increasingly popular in recent years, [8] makes an analysis about the relationship between visual attributes and different layers of convolutional networks. In order to make the attributes discriminative, [43, 29] exploit the discriminative attributes to do the classification task. However, these attributes do not have semantic meanings.

2.2. Zero-Shot Learning

ZSL tackles the problem of recognizing the classes that have never been seen before. With the growth of data scales and the difficulty of image annotation, this application is becoming increasingly popular in recent years. ZSL, first proposed by [17] and [9] in parallel, is accomplished by attributes, which utilizes the cross-category property of attributes to build up the relationships between seen and unseen classes. Then other mid-level semantic descriptions are proposed to tackle such problem, such as word vector [7, 10, 3] and class similarity [23, 22, 24, 44].

An intuitive way to do zero-shot recognition is to train different **attribute classifiers** and recognize an image by the attribute prediction results and unseen class descriptions [17, 9]. Considering the unreliability of the attribute classifiers, [13] proposes a random forest approach to make more robust predictions and [41, 21, 40] model the relationships between attributes and classes to improve the attribute prediction results. To make use of the rich intrinsic structure on the semantic manifold, [12] proposes semantic manifold distance to recognize the unseen class samples. Another widely used approach is **label embedding**, which projects the images and labels into a common semantic space and performs the classification task via nearest neighbor approach [1, 2, 26, 44, 19, 42]. In order to expand ZSL to large-scale settings, [25, 10, 37, 3] use neural networks to learn more complicated non-linear embeddings. Some other works use **transfer learning** techniques to transfer knowledge from seen classes to unseen classes [33, 32, 31, 37, 24, 15]. Recently, [45] proposes to capture the latent relationships between the seen and unseen classes in the similarity space. [5] proposes to synthesize the unseen class classifier directly by sharing the representations between the semantic space and feature space. [4] proposes a metric learning approach to tackle the ZSL problem. [6] expands the traditional ZSL problem to the generalized ZSL problem, where the seen classes are also considered in the test process.

Another popular assumption for ZSL is that the unseen-class samples are available in the problem settings [11, 19, 15, 46]. To make use of the complementary information among different semantic descriptions, [11] proposes a multi-view embedding approach, where graph models are constructed using both the seen and unseen class samples to reduce the domain difference between seen and unseen classes. [19] proposes a semi-supervised framework to learn the unseen classifiers directly where semantic information can be incorporated as side information. [15] utilizes domain adaption approaches to tackle the domain shift problem between seen and unseen classes. Inspired by the clustering property of samples within one category, [46] leverages the structured prediction approach to recognize the unseen class samples. It is important to point out that our approach is not in such settings.

3. Proposed Approach

We propose a latent attribute dictionary (**LAD**) learning process for ZSL. There are some motivations to design the objective function. First, the latent attributes should preserve the semantic information, and thus are able to relate the seen and unseen classes. Second, the representations in the latent attribute space should be discriminative to recognize different classes. Based on these considerations, a framework proposed for **LAD** is shown in Figure 2.

3.1. Problem Formulation

Suppose there are c_s seen classes with n_s labeled samples $\Phi_s = \{X_s, A_s, Z_s\}$ and c_u unseen classes with n_u unlabeled samples $\Phi_u = \{X_u, A_u, Z_u\}$. Each sample x_i is represented as a d-dimensional feature vector. Then we have $X_s \in \mathbb{R}^{d \times n_s}$ and $X_u \in \mathbb{R}^{d \times n_u}$, where $X_s = [\boldsymbol{x_1}, ..., \boldsymbol{x_{n_s}}]$ and $X_u = [\boldsymbol{x_1}, ..., \boldsymbol{x_{n_u}}]$. Z_s and Z_u are the class labels of the seen and unseen class samples. In zero-shot recognition settings, the seen and unseen classes are disjoint: $Z_s \bigcap Z_u = \emptyset$. A_s and A_u are the m-dimensional semantic representations (*i.e.* attribute annotations) of seen and unseen class samples, where $A_s \in \mathbb{R}^{m \times n_s}$ and $A_u \in \mathbb{R}^{m \times n_u}$. The semantic information about the seen class samples A_s is provided and that for the unseen class samples A_u is unknown. Given the semantic descriptions of the classes $P \in \mathbb{R}^{m \times (c_s + c_u)}$, the goal of ZSL is to predict Z_u.

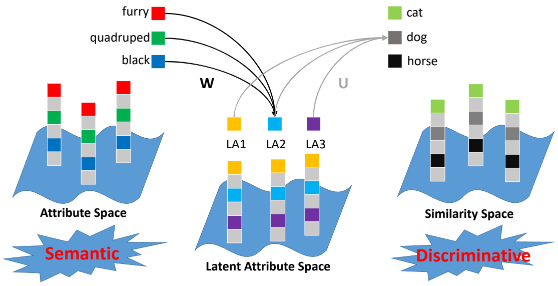

Figure 2. The latent attribute dictionary learning framework for ZSL. The proposed latent attribute space is connected with both the attribute space and the similarity space. The attribute space makes the latent attributes semantic-preserving and the similarity space makes the latent attributes discriminative.

3.2. Latent Attribute Dictionary Learning

The key issue to perform ZSL task is to find a common space which can build up the relationship between seen and unseen classes. Traditional approaches select the attribute space to perform the recognition task. For example, [15] proposes to learn the attribute dictionary directly as :

$$\arg\min_{D, Y_s} \|X_s - DY_s\|_F^2, \quad s.t. \quad ||d_i||_2^2 \leq 1, \forall i, \tag{1}$$

where $\|.\|_F$ denotes the Frobenius norm, d_i is the ith column of the learned dictionary D, and Y_s is the reconstruction coefficient. By forcing Y_s to be A_s, the learned dictionary are viewed as representations of the attributes. However, this constraint is too strong. If we relax the semantic constraint and the objective function can be formulated as:

$$\begin{aligned} &\arg\min_{D, Y_s} \|X_s - DY_s\|_F^2 + \alpha \|Y_s - A_s\|_F^2, \\ &s.t. \quad ||d_i||_2^2 \leq 1, \forall i. \end{aligned} \tag{2}$$

The second term in Eq. 2 encourages Y_s to be similar to the attribute representations A_s, thus to ensure that the learned bases depict attribute dictionary items.

Although attributes are widely used in the recognition task, there are two things that should be considered. First, the user-defined attributes are not always the same important for discrimination, thus it may be less desirable to learn each attribute directly. Second, there are correlations among the attributes, thus it is not suitable to learn each attribute independently. To address such problems, we propose to learn latent attributes. Specifically, we use dictionary learning framework to model the latent attribute space directly. To preserve the semantic information, a linear transformation matrix W is utilized to build up the relationship between latent attributes and attributes, as is shown in Figure 2:

$$\begin{aligned} \arg \min_{D,Y_s,W} & \|X_s - DY_s\|_F^2 + \alpha \|Y_s - WA_s\|_F^2 , \\ s.t. \quad & ||d_i||_2^2 \le 1, \quad ||w_i||_2^2 \le 1, \forall i, \end{aligned} \tag{3}$$

where w_i is the ith column of W. It can be inferred from Eq. 3 that the latent attributes can be viewed as linear combinations of the semantic attributes, which will implicitly combine strongly correlated attributes.

In order to make more effective recognition task, the latent attributes should be discriminative. In other words, we want to find the most discriminative attribute combinations to classify different categories. Thus we utilize the seen-class classifiers to make the latent attributes more discriminative. Specifically, a linear mapping U is learned from the latent attribute space to the seen categories, as is shown in Figure 2:

$$\begin{aligned} \arg \min_{D,Y_s,W,U} & \|X_s - DY_s\|_F^2 + \alpha \|Y_s - WA_s\|_F^2 \\ & + \beta \|H - UY_s\|_F^2 , \\ s.t. \quad & ||d_i||_2^2 \le 1, \quad ||w_i||_2^2 \le 1, \quad ||u_i||_2^2 \le 1, \forall i, \end{aligned} \tag{4}$$

where $H = [h_1, h_2, ..., h_{n_s}] \in \mathbb{R}^{c_s \times n_s}$ and $h_i = [0...0\,1\,0...0]^T$ is a one hot vector which shows the class label of sample x_i. Here the column vectors h_i of H form the similarity space $\mathbb{R}^{c_s}$, where each dimension represents the similarity to one seen class. U can be viewed as seen-class classifiers in the latent attribute space. The third term in Eq. 4 aims to make the latent attribute representations discriminative enough to classify different classes. It implicitly pulls samples from the same class together and pushes those from different classes away from each other.

In summary, the latent attributes learned by the proposed method are not only discriminative to classify different classes but also semantic-preserving. As is shown in Figure 2, the resulted latent attribute space is connected with two kinds of semantic spaces. First, a linear transformation matrix W is utilized to connect the latent attribute space with the semantic attribute space. Through this transformation matrix, we can recover the attribute representations by the latent attributes, thus the latent attributes have semantic meanings. Second, the category classifiers U in the latent attribute space can be viewed as the connection between latent attribute space and the similarity space. This constraint encourages the latent attributes to be discriminative.

3.3. Optimization

It is obvious that Eq. 4 is not convex for D, Y_s, W and U simultaneously, but it is convex for each of them separately. We thus employ an alternating optimization method to solve it. In particular, we alternate between the following subproblems:

(1) Fix D, W, U and update the latent attribute representations Y_s. The subproblem can be formulated as:

$$\arg \min_{Y_s} \left\| \tilde{X} - \tilde{D} Y_s \right\|_F^2 , \tag{5}$$

where

$$\tilde{X} = \begin{bmatrix} X_s \\ \alpha W A_s \\ \beta H \end{bmatrix}, \tilde{D} = \begin{bmatrix} D_s \\ \alpha I \\ \beta U \end{bmatrix},$$

and I is the identity matrix. Forcing the derivative of Eq. 5 to be 0 and the closed-form solution for Y_s is

$$Y_s = (\tilde{D}^T \tilde{D})^{-1} \tilde{D}^T \tilde{X}. \tag{6}$$

(2) Fix Y_s, W, U and update the latent attribute dictionary D. The subproblem can be formulated as:

$$\arg \min_{D} \|X_s - DY_s\|_F^2 , \quad s.t. \quad ||d_i||_2^2 \le 1. \tag{7}$$

This problem can be optimized by the Lagrange dual. Thus the analytical solution for Eq. 7 is

$$D = (X_s Y_s^T)(Y_s Y_s^T + \Lambda)^{-1}, \tag{8}$$

where Λ is a diagonal matrix constructed by all the Lagrange dual variables.

(3) Fix D, Y_s, U and update the embedding W. The subproblem can be formulated as:

$$\arg \min_{W} \|Y_s - WA_s\|_F^2 , \quad s.t. \quad ||w_i||_2^2 \le 1. \tag{9}$$

This problem can be solved in the same way as Eq. 7. The analytical solution for W is

$$W = (Y_s A_s^T)(A_s A_s^T + \Lambda)^{-1}, \tag{10}$$

where Λ is a diagonal matrix constructed by all the Lagrange dual variables.

(4) Fix D, Y_s, W and update the embedding U. The subproblem can be formulated as:

$$\arg \min_{U} \|H - UY_s\|_F^2 , \quad s.t. \quad ||u_i||_2^2 \le 1. \tag{11}$$

This problem can be solved in the same way as Eq. 7. The analytical solution for U is

$$U = (HY_s^T)(Y_s Y_s^T + \Lambda)^{-1}, \tag{12}$$

Algorithm 1 Latent Attribute Dictionary Learning for ZSL

```
Input: $X_s, A_s, \alpha, \beta$
Output: $D, Y_s, W$ and $U$
Initialize $D, W, U$ randomly.
while not converge do
    Update $Y_s$ by Eq. 6.
    Update $D$ by Eq. 8.
    Update $W$ by Eq. 10.
    Update $U$ by Eq. 12.
end while
```

where Λ is a diagonal matrix constructed by all the Lagrange dual variables.

The complete algorithm is summarized in Algorithm 1. In our experiments, the optimization process always converges after tens of iterations, usually less than 50 [1].

3.4. ZSL Recognition

Since our latent attribute space is associated with attribute space and similarity space, we can perform ZSL in multiple spaces.

Recognition in the latent attribute space. In order to perform ZSL in the latent attribute space, we must obtain the latent attribute representations of the test samples and the unseen class prototypes. Given a test sample with its feature vector x^u, we obtain its latent attribute representation y^u by

$$y^u = \min_{y^u} \|x^u - Dy^u\|_F^2 + \gamma \|y^u\|_2^2 , \tag{13}$$

where D is the latent attribute dictionary learned from the training data. γ is a weight for regularization term. For the unseen class prototypes, we project their attribute representations to the latent attribute space by the transformation matrix W. Then the distance between the test sample and the unseen classes can be computed. In the end, we perform ZSL recognition by nearest neighbour approach using the cosine distance.

Recognition in the similarity space. As is described above, we can use U to transform the latent attribute representations y^u to the similarity space, where each dimension represents the similarity to one seen class. Thus each image can be denoted by a similarity vector h. Moreover, we can also obtain the similarity representation of an unseen class prototype by mapping the class attribute vectors to histogram representations of seen class proportions, as similarly done in [44]. Thus nearest neighbour approach can be performed to classify a test sample to an unseen class.

Recognition in the attribute space. We can recover the attribute representation of an image a^u by its latent attribute representation y^u:

$$a^u = \min_{a^u} \|y^u - Wa^u\|_F^2 + \lambda \|a^u\|_2^2 . \tag{14}$$

Thus the attributes of each image can be obtained. Then we can classify a test image to an unseen class by the attribute representations.

Combining multiple spaces. Since different spaces may contain complementary information of an unseen class, we can combine different spaces to perform the ZSL task. In this paper, we simply concatenate different vector representations of an image to form the final representation and the same process is done for the unseen class prototypes. Then ZSL task can be performed by the same approach proposed above.

3.5. Discussion

Difference from other works. Our approach is mainly inspired by JLSE [45] and JSLA [29]. JLSE proposes the latent similarity space to build up the relation between feature space and similarity space, while we utilize the similarity space to make our latent attributes more discriminative. JSLA proposes to learn user-defined attributes and discriminative attributes separately, where background information is also modeled. However, it is not designed for ZSL task since background information can not be utilized in ZSL settings. In contrast, we embark from the actual problems that exist in ZSL task and merge the semantic and discriminative information into the latent attributes aiming to make more effective recognition.

Further improvements. The proposed approach is based on linear models. It is well known that deep learning based nonlinear models would be more powerful. Thus we use the deep network to extract the image representations and ensure favorable performance to a large extent. It is believed that learning the semantic transformation directly through a deep model will help to boost the performance and this remains a promising direction for our future work.

4. Experiments

In this section, we evaluate the proposed method on four benchmark datasets and then analyze the effectiveness of the method.

4.1. Datasets and Settings

Datasets. We perform experiments on four benchmark ZSL datasets to verify the effectiveness of the proposed method, *i.e.* aPascal & aYahoo (**aP&Y**) [9], Animals with Attributes (**AwA**) [17], Caltech-UCSD Birds-200-2011 (**CUB-200**) [39], and SUN Attribute (**SUN-A**) [28]. The statistics of these datasets are shown in Table 1. (a) **aP&Y** consists of two attribute datasets: aPascal contains 12,695 images and aYahoo has 2,644 images. A 64-D

[1] The source code of the proposed *LAD* approach is available at *http://vipl.ict.ac.cn/resources/codes*.

Table 1. Statistics of different datasets, where 'b' and 'c' stand for binary value and continuous value respectively.

Database	Instance	Attributes	Seen/Unseen
aP&Y	15,339	64(c)	20 / 12
AwA	30,475	85(c)	40 / 10
CUB-200	11,788	312(b)	150 / 50
SUN-A	14,340	102(c)	707 / 10

attribute vector is provided for each image. There are 20 object classes for aPascal and 12 for aYahoo, which are disjoint. For ZSL task, the categories in aPascal dataset are used as seen classes and those in aYahoo are used as unseen classes. (b) **AwA** is an animal dataset, which contains 50 animal categories, with 30,475 images. There are 85 attributes annotated for each class. The standard split for ZSL is to use 40 categories as seen classes and the other 10 categories as unseen classes. (c) **CUB-200** is a bird dataset for fine-grained recognition. It contains 200 classes with 11,788 images, where 312 binary attributes are provided for each image. Following the same setting as [1], we take 150 categories as seen classes and the other 50 as unseen classes. (d) **SUN-A** is created for high-level scene understanding and fine-grained scene recognition, which contains 717 classes with 14,340 images collected. There are 102 real-valued attribute annotations for each image, which are produced by a voting process. Following [13], we select the same 10 classes as unseen classes.

Parameter Settings. For all the datasets, we utilize the 4096-dimension CNN feature vectors extracted by the imagenet-vgg-verydeep-19 pre-trained model [36]. We use the multi-class accuracy as the evaluation metric. The dictionary size for **AwA** is chosen as 300 and those for the other three datasets are 500 to 800. Other parameters α and β are obtained using five-fold cross-validation. γ and λ can also be tuned and we set them to be 1 for simplicity. More details can be found in the supplementary material.

4.2. Effectiveness of the Proposed Framework

The proposed latent attribute space is associated with the attribute space and the similarity space, which makes it not only semantic-preserving but also discriminative enough to classify different classes. To demonstrate the effectiveness of each component, we compare five different approaches and the results on **aP&Y** are shown in Figure 3. (1) Recognition in the attribute space (**A**) merely, by learning the attribute dictionary directly, as is proposed in Eq. 2. (2) Recognition in the similarity space (**S**) merely, by removing the semantic constraint (*i.e.* the second term) in Eq. 4. (3) Recognition in the latent attribute space but without discriminative constraint (**LA**$^-$), by removing the discriminative constraint (*i.e.* the third term) in Eq. 4. (4) Recognition in the latent attribute space with discriminative constraint (**LA**), as is proposed in Eq. 4. (5) Recognition by combining the latent attribute space and similarity space (**LAS**), as is proposed in Section 3.4.

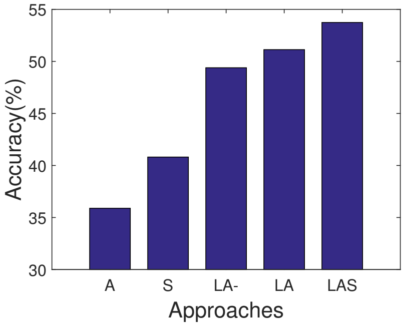

Figure 3. Comparisons of five approaches on **aP&Y**.

By comparing the performance of **A**, **S** and **LA**$^-$, we can infer that the latent attributes are more successful in the ZSL task. Moreover, by imposing the discriminative constraint in the objective function, the recognition accuracy improves, as is shown by the performance of **LA**. Furthermore, the combination of latent attributes and seen class similarities also improves the final recognition performance, as is shown by the result of **LAS**.

4.3. Benchmark Comparisons and Evaluations

In this part, we compare our method with several popular approaches. Our settings are the same as those in [44]. Table 2 shows results on four datasets, where the blank spaces indicate that the corresponding methods were not tested on the datasets in their original papers. Here, for our LAD method, we report the results of two variants **LA** and **LAS**, as described in Section 4.2. It is important to point out that the first three approaches are in transductive ZSL settings, where the information of unseen class samples are utilized, while our approach and other competing methods are in traditional ZSL settings. From this table, we can see that our approach achieves the best performance on three datasets, which shows the effectiveness of the proposed method. Note that '*' indicates that the approaches use different features from ours, where [11, 15] use the OverFeat [34] features and [2, 5, 6] use the GoogleNet [38] features.

From Table 2, we can see that great improvement is made on **CUB-200** dataset. It should be contributed by the good reconstruction property of latent attribute dictionary. This dataset is designed for fine-grained recognition, where the sample variation is not very large. Thus the dictionaries learned on the seen class samples can have a good reconstruction of the unseen class samples. With the help of discriminative property of latent attributes, the recognition performance of unseen classes can be improved further. Our result on **AwA** is slightly lower than that of JSLA [29]. The reason may lie in the class-level attribute annotations, which can not cover the variations of images within each class.

Table 2. Comparisons with published results in multi-class accuracy (%) for the task of zero-shot learning. '*' denotes the features are different from ours. '†' denotes the data split is different from ours.

Method	aP&Y	AwA	CUB-200	SUN-A
TMV-HLP* [11]		80.50	47.90	
UDA* [15]		75.60	40.60	
SP-ZSR [46]	**69.74**	**92.08**	**55.34**	**89.50**
DAP [18]	38.16	57.23		72.00
ESZSL [26]	24.22	75.32		82.10
SJE* [2]		66.70	50.10	
SSE-ReLU [44]	46.23	76.33	30.41	82.50
JLSE [45]	50.35	80.46	42.11	83.83
SynC* [5]		72.90	54.70†	62.70†
JSLA [29]		**82.81**	49.87	
Chao *et al.** [6]		73.40	54.40†	
Bucher *et al.* [4]	53.15	77.32	43.29	84.41
LAD (LA)	51.13	80.49	56.36	83.50
LAD (LAS)	**53.74**	82.48	**56.63**	**85.00**

The reconstruction accuracy of dictionary may thus be influenced by the limited attribute representations.

4.4. Semantic-Preserving Property

Different from previous approaches which learn discriminative attributes with no semantic meanings [43], our latent attributes are semantic-preserving. This can be reflected by the following two aspects.

First, we can recover the attributes of images by their latent attribute representations, as is shown in Eq. 14. In order to test whether the recovered attributes are good or not, we perform an attribute prediction task on **AwA**. Specifically, we binarize the attributes by the thresholds provided by the original dataset and measure the performance by Area Under Curve(AUC). Figure 4 shows the attribute prediction results of unseen classes. Blank spaces indicate that there are no such attributes in the unseen classes. We can figure out that most of the attributes recovered from the latent attributes have relatively good performance, which demonstrates the semantic-preserving property of the latent attributes. The mean AUC of attributes recovered by our approach is 0.729, which is comparative to other attribute prediction results [29]. Since the main purpose of our approach is not attribute prediction, no further detailed comparison is conducted.

Second, the latent attributes can be viewed as different combinations of attributes. To have an intuitive understanding of what the combinations are, we visualize some latent attribute dictionaries. Specifically, given a latent attribute dictionary item, we show the images which have the largest and smallest correspondance over this item. Figure 5 shows the visualization results on **AwA** using the unseen class samples. Besides, we also show the attributes with the largest and smallest activations over the latent attributes in Figure 5. It can be observed that the images, which have high correspondence over the latent attribute, are in accordance with the attribute combinations. For example, *seal* is highly correlated with attributes in the green blocks of 'LA1' and has little relevance to those in red blocks. We can figure out that a specific latent attribute may group some highly correlated attributes, thus the variation within each latent attribute becomes smaller. It is desirable that a latent attribute should have strong activations only on a small numbers of attributes. While we do not explicitly impose such sparsity constraint, the combination coefficients are found to be mostly sparse in our experimental study. Due to limited space, detailed statistic results and analyses are provided in the supplementary material.

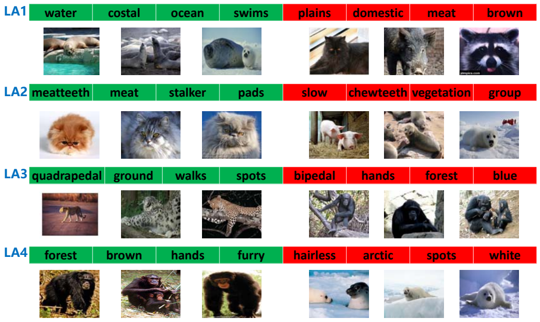

Figure 5. Latent attribute dictionary visualization on **AwA** using the unseen class samples. 'LA' is the latent attribute. Green blocks show attributes with largest activations and red ones show attributes with smallest activations. The first three pictures are randomly selected from images which have largest activations over the latent attribute and the last three are selected from images with smallest activations.

4.5. Discriminative Property

From our objective function in Eq. 4, it can be seen that a discriminative constraint is achieved by learning seen class classifiers in the latent attribute space. The probability outputs of seen-class classifiers for each image can be viewed as a vector representation in the similarity space, where each dimension represents the similarity to one seen class. Figure 6 shows the top five most similar seen classes corresponding to some unseen class samples on **aP&Y**, where the probability outputs are obtained by softmax function. It can be seen that most of the similarities are human comprehensible. For example, *wolf* is most similar to *cat* and *jetski* is most similar to *motorbike*.

In order to perform ZSL task, we also need the similarity representations of unseen class prototypes. This can be obtained by the approach proposed by [44], as is mentioned in Section 3.4. Figure 7 shows the similarities of unseen classes to the seen classes on **aP&Y**, where the values are normalized. It can be inferred that most of the similar-

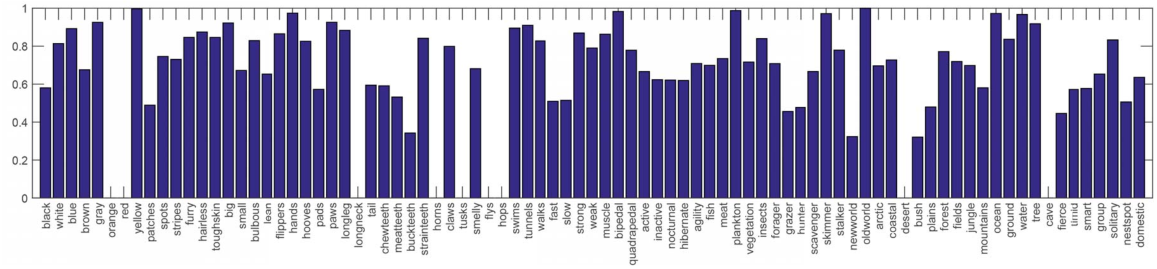

Figure 4. Attribute prediction results on **AwA** with the measure of AUC.

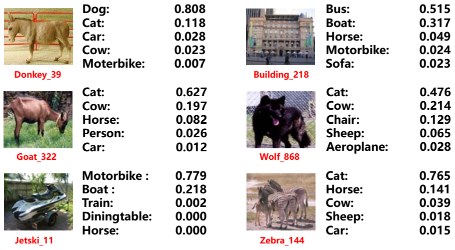

Figure 6. Similarity representations of unseen class samples on **aP&Y**. Top five similar classes are shown.

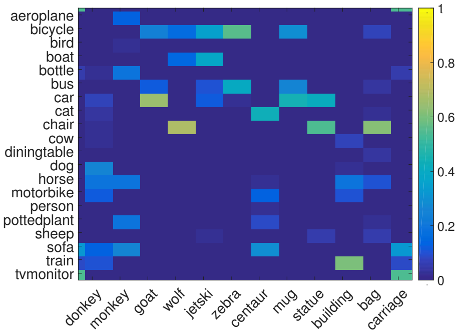

Figure 7. Similarity representations of unseen class prototypes on **aP&Y**. Each column shows the similarity of an unseen class to the seen classes.

ities are also in accordance with human knowledge. Since the similarities grasp some information of unseen classes, we can also use such semantic information to perform ZSL task. However, due to the fact that there are only limited numbers of seen classes, some similarities are not that comprehensible, such as the similarity representation of *bag*.

As mentioned above, in order to make the latent attribute discriminative, we utilize seen-class classifiers to connect the latent attribute space to the similarity space. To explore whether the latent attributes are discriminative or not, we visualize the unseen class samples in **AwA** by their latent attribute representations learned by our approach. Specifically, we use t-SNE [20] to project the learned latent attribute representations of each unseen class sample to a 2-D plane. Figure 8 shows the visualization results. We can see that images of the same class are grouped together and those

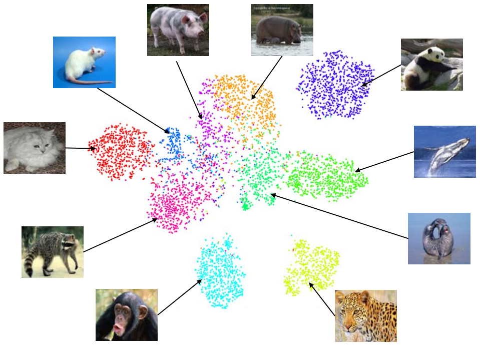

Figure 8. Visualization of the unseen class samples on **AwA** using their latent attribute representations. Each color represents a class and each point represents an image.

from different classes are separated, which indicates that the latent attributes are discriminative for the recognition task. It is also very interesting to find that visually similar classes have small distances. For example, the cluster of *pigs* is near that of *hippopotamuses* and the cluster of *humpback whales* is near that of *seals*. This again shows that the latent attributes preserve the semantic information.

5. Conclusion

In this paper, we propose a novel ZSL approach which is accomplished by latent attribute dictionary learning (**LAD**). The proposed approach shows its effectiveness on four datasets. We attribute the promising performance of **LAD** to three aspects. First, the proposed latent attribute space is connected with the attribute space, so it preserves the semantic information. Thus it is possible to perform ZSL task in the latent attribute space. Second, the latent attribute space is connected with the similarity space. This makes the latent attribute space discriminative to recognize different categories. Third, the latent attributes can be viewed as different combinations of semantic attributes, which implicitly deals with the attribute correlation problem.

Acknowledgements. This work is partially supported by Natural Science Foundation of China under contract Nos. 61390511, 61379083, 61650202; 973 Program under contract No. 2015CB351802; and Youth Innovation Promotion Association CAS No. 2015085.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[3] J. Ba, K. Swersky, S. Fidler, and R. Salakhutdinov. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proc. of International Conference on Computer Vision*, pages 4247–4255, 2015.

[4] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classification. In *Proc. of European Conference on Computer Vision*, pages 730–746, 2016.

[5] S. Changpinyo, W. L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *Proc. of Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[6] W. L. Chao, S. Changpinyo, B. Gong, and F. Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *Proc. of European Conference on Computer Vision*, pages 52–68, 2016.

[7] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proc. of International Conference on Computer Vision*, pages 2584–2591, 2013.

[8] V. Escorcia, J. C. Niebles, and B. Ghanem. On the relationship between visual attributes and convolutional networks. In *Proc. of Computer Vision and Pattern Recognition*, pages 1256–1264, 2015.

[9] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proc. of Computer Vision and Pattern Recognition*, pages 1778–1785, 2009.

[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 *Proc. of Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[11] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(11):2332–2345, 2015.

[12] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *Proc. of Computer Vision and Pattern Recognition*, pages 2635–2644, 2015.

[13] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Proc. of Advances in Neural Information Processing Systems*, pages 3464–3472, 2014.

[14] D. Jayaraman, F. Sha, and K. Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proc. of Computer Vision and Pattern Recognition*, pages 1629–1636, 2014.

[15] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *Proc. of International Conference on Computer Vision*, pages 2452–2460, 2015.

[16] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. of Computer Vision and Pattern Recognition*, pages 951–958, 2009.

[18] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[19] X. Li, Y. Guo, and D. Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *Proc. of International Conference on Computer Vision*, pages 4211–4219, 2015.

[20] L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[21] D. Mahajan, S. Sellamanickam, and V. Nair. A joint learning framework for attribute models and object descriptions. In *Proc. of International Conference on Computer Vision*, pages 1227–1234, 2011.

[22] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 2441–2448, 2014.

[23] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Proc. of European Conference on Computer Vision*, pages 488–501. 2012.

[24] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proc. of International Conference on Learning Representations*, 2014.

[25] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *Proc. of Advances in Neural Information Processing Systems*, pages 1410–1418, 2009.

[26] B. R. Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *Proc. of International Conference on Machine Learning*, pages 2152–2161, 2015.

[27] G. Patterson and J. Hays. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. of Computer Vision and Pattern Recognition*, pages 2751–2758, 2012.

[28] 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.

[29] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang. Joint learning of semantic and latent attributes. In *Proc. of European Conference on Computer Vision*, pages 336–353, 2016.

[30] R. Qiao, L. Liu, C. Shen, and A. V. D. Hengel. Less is more: zero-shot learning from online textual documents with noise

suppression. In *Proc. of Computer Vision and Pattern Recognition*, pages 2249–2257, 2016.

[31] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Proc. of Advances in Neural Information Processing Systems*, pages 46–54, 2013.

[32] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proc. of Computer Vision and Pattern Recognition*, pages 1641–1648, 2011.

[33] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *Proc. of Computer Vision and Pattern Recognition*, pages 910–917, 2010.

[34] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. of International Conference on Learning Representations*, 2014.

[35] B. Siddiquie, R. S. Feris, and L. S. Davis. Image ranking and retrieval based on multi-attribute queries. In *Proc. of Computer Vision and Pattern Recognition*, pages 801–808, 2011.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of International Conference on Learning Representations*, 2015.

[37] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Proc. of Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[38] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proc. of Computer Vision and Pattern Recognition*, pages 1–9, 2015.

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

[40] X. Wang and Q. Ji. A unified probabilistic approach modeling relationships between attributes and objects. In *Proc. of International Conference on Computer Vision*, pages 2120–2127, 2013.

[41] Y. Wang and G. Mori. A discriminative latent model of object classes and attributes. In *Proc. of European Conference on Computer Vision*, pages 155–168, 2010.

[42] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. A. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *Proc. of Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[43] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *Proc. of Computer Vision and Pattern Recognition*, pages 771–778, 2013.

[44] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *Proc. of International Conference on Computer Vision*, pages 4166–4174, 2015.

[45] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proc. of Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.

[46] Z. Zhang and V. Saligrama. Zero-shot recognition via structured prediction. In *Proc. of European Conference on Computer Vision*, pages 533–548, 2016.