Low-Rank Embedded Ensemble Semantic Dictionary for Zero-Shot Learning*

†Zhengming Ding ‡Ming Shao †♯Yun Fu
†Department of ECE, College of Engineering, Northeastern University, Boston, USA
‡Computer and Information Science, University of Massachusetts Dartmouth, USA
♯College of Computer and Information Science, Northeastern University, Boston, USA
allanding@ece.neu.edu, mshao@umassd.edu, yunfu@ece.neu.edu

Abstract

Zero-shot learning for visual recognition has received much interest in the most recent years. However, the semantic gap across visual features and their underlying semantics is still the biggest obstacle in zero-shot learning. To fight off this hurdle, we propose an effective Low-rank Embedded Semantic Dictionary learning (LESD) through ensemble strategy. Specifically, we formulate a novel framework to jointly seek a low-rank embedding and semantic dictionary to link visual features with their semantic representations, which manages to capture shared features across different observed classes. Moreover, ensemble strategy is adopted to learn multiple semantic dictionaries to constitute the latent basis for the unseen classes. Consequently, our model could extract a variety of visual characteristics within objects, which can be well generalized to unknown categories. Extensive experiments on several zero-shot benchmarks verify that the proposed model can outperform the state-of-the-art approaches.

1. Introduction

Visual recognition algorithms assume that the training and test data share the same classes/labels/tags and feature space, so that the learned classifier can be reused for the test data without any change. However, it is a bottleneck to collect a large number of well-labeled images for each class, especially when visual recognition task is moving towards a fine-grained scenario. In addition, labeling work for such collections is expensive, and requires either large quantities of attributes or expert opinions [20, 21, 34, 1, 23].

To that end, zero-shot learning (ZSL) has been developed recently which attracts great attention due to its appealing performance. ZSL is inspired by the learning mechanism of human brain and attempts to recognize new classes which are not observed in the training stage [30, 13, 37, 17, 3, 33, 25, 10, 24]. For example, one can recognize a new species of animal after being told what it looks like and how it is similar to or different from other observed animals. The reason is simple: humans can explore the relationship across different objects through secondary information, and adapt the knowledge from known classes to unknown ones. Likewise, ZSL aims to uncover the intrinsic semantic relationship across seen and unseen classes. In general, three fundamental elements are needed: (1) visual representation conveying nontrivial yet informative visual features; (2) semantic representation reflecting the relationship across different classes; (3) learning model properly linking visual features with the underlying semantics.

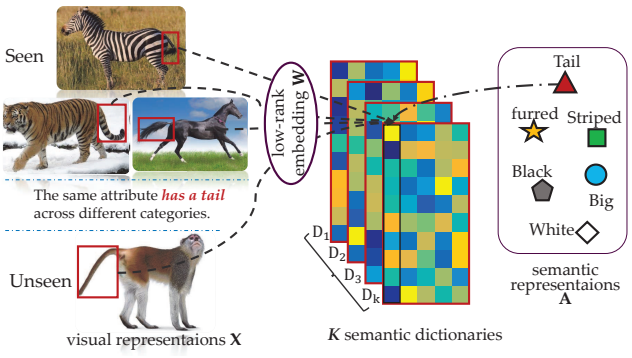


Figure 1. Illustration of our proposed framework, where low-rank projection W maps visual features X into a new space, thus similar features, e.g., "has a tail", would gather together. Simultaneously, multiple semantic dictionaries D_k are learned with the constraint $WX \approx DA$ to connect visual features and their semantic representations. In this way, multiple transferable semantic dictionaries could constitute the latent basis for the unseen classes.

While ZSL is promising in simulating the human learning process, it has two degenerating factors. First, the distribution of samples in visual feature space is often distinct from that of their underlying semantic space as visual features in various forms may convey the same concept. Such

*This work is supported in part by the NSF IIS award 1651902, ONR Young Investigator Award N00014-14-1-0484, U.S. Army Research Office Young Investigator Award W911NF-14-1-0218, and NIJ Award 2016-R2-CX-0013.

semantic gap traps the knowledge transfer from the observed classes to unseen classes. Secondly, the "hubness" [27] is recently identified as a factor that accounts for the poor performance, which is exacerbated by a lack of training instances of unknown classes in visual domain. Hence, the domain shift problem, i.e., the distribution difference between training and test data, raises a challenge in ZSL [9, 37, 4].

Convectional ZSL approaches typically consider that there exists a shared semantic latent space where both the visual features and the class labels of the seen and unseen classes lie in [30, 13, 37, 17, 3, 33, 25, 10, 24]. Specifically, the information learned through the observed data is usually captured by a mapping function, e.g., embedding, that transforms each low-level feature vector to its class prototype. Through such a mapping, the captured knowledge could be adapted to the unseen data in the evaluation stage.

In this paper, we develop an effective Low-rank Embedded ensemble Semantic Dictionary learning (LESD) to handle issues in zero-shot learning (Figure 1). Our main assumption is that the latent semantic dictionary for unseen data should share its majority with semantic dictionary for the seen data[1], which can be identified in the low-rank embedding space. In addition, multiple transferable dictionaries learned for the unseen data will have better chance to recover the latent semantic dictionary. Finally, we summarize our contributions in three folds:

- First, we identify a low-rank embedding to transfer the intrinsic knowledge and shared features from the seen categories. In this way, a better latent semantic dictionary for the unseen categories can be recovered.
- Second, ensemble strategy is exploited to learn multiple semantic dictionaries, which is able to complete the latent semantic dictionary to mitigate the distribution divergence across seen and unseen classes.
- Computationally, we adopt a novel low-rank reframing approach to overcome the existing sparse singular values issues to secure a better low-rank embedding space. We also design a nontrivial solution for efficiency.

2. Related Work

Zero-shot learning (ZSL) manages to build models of visual concepts without test images containing these concepts. As visual knowledge from such test classes is unobservable during training, ZSL requires auxiliary information to make up for the unknown visual knowledge. Attribute-based descriptions are the most well-known characteristics shared across various classes [20, 21, 34, 1, 23], which provide a secondary representation linking the low-level visual features with the semantic labels. Given the low-level visual representations of images and their underlying high-level semantics, the key problem in ZSL turns to "how to adapt knowledge from the visual data of observed classes to those of unobserved ones" [30, 13, 37, 17, 3, 33, 25]. Generally, there are three lines of ZSL approaches in terms of the strategy to bridge the semantic gap.

First of all, direct mapping is designed to seek a projection function from visual features to their corresponding semantic representations [1, 13]. Along this line, Direct Attribute Prediction as well as Indirect Attribute Prediction adopted the hidden layer of attributes as variables decoupling the images from the layer of labels [15]. Further, Gan et al. proposed to seek a representation transformation in visual space to enhance the attribute-level discriminative capacity for attribute prediction [11].

Secondly, common space learning tries to find new spaces where visual features and semantic representations enjoy the maximum similarities for instances of the same class. The learned common space is either interpretable [36] or latent [9]. Following this, Zhang et al. developed a model by treating any instance in unseen classes as a mixture of those in known classes in both visual and semantic spaces [36]. More recently, Zhang et al. further presented a probabilistic framework for learning joint similarity latent embedding where both visual and semantic embedding along with a class-independent similarity measure are learned simultaneously [37].

Thirdly, parameter mapping aims to estimate model parameters for unseen classes by "tuning" model parameters learned from observed classes. Essentially, it exploits the inter-class relationship between observed and unseen classes in semantic space [19, 4]. Along this line, Mensink et al. employed co-occurrences statistics of visual concepts within images and adopted the co-occurrences to design a new classifier [19]. Furthermore, Changpinyo et al. proposed to gain model parameters for unseen classes by aligning the topology of all the classes in semantic and model parameter spaces [4].

However, all these methods pay less attention to discriminative knowledge in the unseen classes given high intra-class variability, and may fail to discover shared semantics across different domains. Our proposed approach follows in the direct mapping category, which is similar to the regression problem as "dictionary learning + sparse coding" [13]. Moreover, recent research efforts show the appealing superiority of ensemble learning in dictionary learning [35, 38, 26], in which a set of base classifiers are trained and integrated as an ensemble classifier to obtain extra performance. Differently, we jointly optimize low-rank embedding and semantic dictionary to capture shared discriminative features across seen and unseen classes. Furthermore, ensemble strategy helps recover the complete latent seman-

[1] Both seen and unseen data in our work share lots of semantics.

tic space that cannot be fulfilled by a single dictionary.

3. The Proposed Algorithm

In this section, we will present our novel low-rank embedded semantic dictionary learning via ensemble strategy, followed by an efficient solution.

Suppose there are C seen classes with n labeled samples $\mathcal{S} = \{X, A, y\}$ and C_u unseen classes with n_u unlabeled samples $\mathcal{U} = \{X_u, A_u, y_u\}$. Each sample is denoted as visual feature with dimension d. Assume there are n samples in the seen training data and n_u samples in the unseen test data, and thus, the visual features are represented as $X \in \mathbb{R}^{d\times n}$ and $X_u \in \mathbb{R}^{d\times n_u}$, while their corresponding class label vectors are $y \in \mathbb{R}^n$ and $y_u \in \mathbb{R}^{n_u}$. In ZSL setting, the observed and unobserved classes have no label overlap, i.e., $y \cap y_u = \emptyset$. $A \in \mathbb{R}^{m\times n}$ and $A_u \in \mathbb{R}^{m\times n_u}$ are the m-dimensional semantic representations of instances in the seen and unseen datasets, respectively. For the seen dataset, A is provided in advance since seen samples X are labeled with either attribute features or word2vector representations corresponding to their class labels y. On the other hand, A_u needs to be estimated since the unseen data are unlabeled. The task of ZSL is to predict A_u and y_u given visual features X_u using the classifier learned from seen classes.

3.1. Low-rank Embedded Semantic Dictionary Learning

While seen data X and unseen data X_u sampled from different categories lie in different feature spaces, A and A_u may share similar semantics. For example, in attribute-based description, both seen and unseen data can be represented with pre-defined attributes with different weights, e.g., binary or continuous values. The intuition behind zero-shot learning is that the classifier would be able to capture the relationship between the visual-input space and the individual dimensions of the semantic feature space [20].

Since we are not accessible to the data of test classes during the training stage, we are encouraged to discover shared knowledge generalized to the unseen data from the seen ones. Inspired by the recent work [13] considering semantic representation A as the encoded coefficients of X based on a semantic dictionary, we develop an effective low-rank embedded semantic dictionary learning formula that integrates the merits of both semantic representation learning and low-rank discriminative embedding:

$$\begin{aligned} \min_{W,D} \|WX - DA\|_{\mathrm{F}}^2 + \alpha \mathrm{rank}(W) \\ \text{s.t. } \|d_j\|_2^2 \leq 1, \forall j, \end{aligned} \tag{1}$$

where α is the balance parameter, $\|\cdot\|_{\mathrm{F}}$ is the Frobenius norm, and $d_j \in \mathbb{R}^d$ is the j-th atom of semantic dictionary $D \in \mathbb{R}^{d\times m}$. $\mathrm{rank}(\cdot)$ is the rank operator of a matrix.

Remarks: In brief, the rank constraint on $W \in \mathbb{R}^{d\times d}$ enforces a new low-rank representation for seen data to highlight shared semantics across different categories. For example, attribute "it has a tail" would be assigned to many different categories, e.g., horse, monkey, tiger. Low-rank constraint on W will help collect such visual features which underlie the embedding space. In this way, discriminative and descriptive features from seen categories could be adapted to unseen ones. Mathematically, the low-rankness will be propagated to DA in Eq. (1), and thus yield a low-rank semantic dictionary D, which includes shared semantics across categories from seen data.

3.2. Rank Constraint Re-framing

Rank minimization in Eq. (1) is a well-known NP-hard problem, and considerable approaches have been proposed. Majority of them focuses on seeking a surrogate to solve instead. One of appealing strategies is to adopt trace norm $\|W\|_*$ to solve the term $\mathrm{rank}(W)$ [5, 6, 7]. Specifically, trace norm has been corroborated to achieve low-rank matrix structure in the matrix completion literature, which equals the sum of all singular values of W. However, it does not allow an explicit control on the rank of W. That is, the non-zero singular values of matrix W will change along with $\|W\|_*$, but the rank of W may remain unchanged. In this sense, trace norm may not be a good surrogate to obtain the minimal rank matrix.

Alternatively, we exploit a regularization term that guarantees that the rank of optimized W will no larger than a targeted rank r. This skillfully converts the problem to minimizing the square sum of r-smallest singular value of W. When the non-zero singular values increase largely however, they are excluded by our proposed term such that the norm value keeps constant. Mathematically, the new formula with fixed rank constraint can be written as:

$$\begin{aligned} \min_{W,D} \|WX - DA\|_{\mathrm{F}}^2 + \alpha \sum_{i=r+1}^{d} \big(\sigma_i(W)\big)^2 \\ \text{s.t. } \|d_j\|_2^2 \leq 1, \forall j, \end{aligned} \tag{2}$$

where $\sigma_i(W)$ is the i-th singular value of W. Such solutions will naturally converge to a subspace corresponding to the r most significant singular values. As the rank of W is the size of its non-zero singular values, the proposed regularization term allows an explicit constraint over the rank of W. In addition, the novel term can handle the sparse singular values issues raised by existing works[2]. Thanks to the term of square sum of r-smallest singular values, we are

[2]Interestingly, Hu et al. [12] explored the truncated trace norm by minimizing the sum of r-smallest singular values, which can also avoid the effect of large singular values and is better than the traditional trace norm. However, minimizing the sum of r-smallest singular values is an $l_1$minimization problem, which results in sparse solution, i.e., some r-smallest singular values will be zero, but some may get large values.

able to shrink the singular values and make sure all of them shrink down near to zeros.

Specifically, we find that $\sum_{i=r+1}^{d}\left(\sigma_i(W)\right)^2$ equals $\mathrm{tr}(\Gamma^\top WW^\top\Gamma)$, in which $\mathrm{tr}(\cdot)$ is the trace operator of matrix and Γ denotes the singular vectors corresponding to the smallest d-r singular values of $WW^\top$. In this way, we can transform Eq. (1) into the following formulation as:

$$\begin{aligned}\min_{W,D,\Gamma}\|WX-DA\|_{\mathrm{F}}^2+\alpha\mathrm{tr}(\Gamma^\top WW^\top\Gamma)\\ \text{s.t. }\|d_j\|_2^2\leq 1,\forall j.\end{aligned}\tag{3}$$

3.3. Ensemble Discriminative Dictionary Learning

While the dictionary learned in Eq. (3) is able to reconstruct the seen categories in semantic space for each sample pair $\{x_i,a_i\}_{i=1}^n$, it fails to capture the discriminative features within each class. In zero-shot learning, we expect not only a shared dictionary in embedding space, but also discriminative features extended to unseen categories. Namely, the semantic dictionary could also well reconstruct the unseen data in the testing stage. Moreover, we have sufficient pairs $\{x_i,a_j\}_{i,j=1}^n$ sampling the joint space of X and A. These matched pairs and semantics therein will play critical roles in the unseen category learning. For example, if x_i is in class c, then x_i will encoded all the semantics from class c in A too, i.e., A_c with corresponding weights.

To this end, we introduce a new term Z into the dictionary learning to couple the discriminative information from the seen data, which can be written as:

$$\begin{aligned}\min_{W,D,\Gamma}\|WX-DAZ\|_{\mathrm{F}}^2+\alpha\mathrm{tr}(\Gamma^\top WW^\top\Gamma)\\ \text{s.t. }\|d_j\|_2^2\leq 1,\forall j,\end{aligned}\tag{4}$$

where $Z\in\mathbb{R}^{n\times n}$ is weight matrix with its element $z_{ip}=\frac{1}{n_c}$ when x_i and a_p are from the same class (n_c is the sample size for class c), otherwise $z_{ip}=0$. In this way, the semantic dictionary would be more discriminative by preserving more class-wise knowledge from the seen data.

Notably, it is difficult to include necessary semantics for zero-shot learning by a single dictionary D, as little has been known about the unseen data, which will possibly degrade overall performance. This also has been revealed in [13] where a poor performance was identified for the unseen data. Even worse, as we have no access to the unseen data during training, no adaptation can be employed in this problem. To approach ideal semantic dictionary for unseen data, we propose to generate multiple semantic dictionaries through *ensemble learning* [38, 18, 26] in the training stage. We adapt Eq. (4) to achieve this purpose by optimizing the followed formulation:

$$\begin{aligned}\sum_{k=1}^{K}\|WXQ_k-D_kAQ_kZ_k\|_{\mathrm{F}}^2+\alpha\mathrm{tr}(\Gamma^\top WW^\top\Gamma)\\ \text{s.t. }\|d_k^j\|_2^2\leq 1,\forall j,\end{aligned}\tag{5}$$

where d_k^j is the j-th atom of D_k and $Q_k\in\mathbb{R}^{n\times n}$ is column sampling matrix with values only on the diagonal. If $Q_{k,ii}=1$, the i-th sample is selected, otherwise not. Given multiple semantic dictionaries, we have better chance to build the latent semantic space for unseen data. Note as the sample size for each class may change we update Z_k for each sampling. Specifically, we sample $\frac{2}{K}\times 100$ percentage of instances in each class every time.

3.4. Solutions and Optimization

As the formulation in Eq. (5) is not joint convex over all variables, there is no close solution. Thus, we resort to an iterative optimization to update a single unknown variable each time. We further split into two sub-problems, i.e., ensemble semantic dictionary learning D_k by fixing W,Γ; and low-rank embedding learning W,Γ with D_k fixed.

Semantic Dictionary Refinement: When W is fixed, we could optimize the semantic dictionaries D_k as:

$$\begin{aligned}D_k=\arg\min_{D_k}\|WXQ_k-D_kAQ_kZ_k\|_{\mathrm{F}}^2\\ \text{s.t. }\|d_k^j\|_2^2\leq 1,\forall j.\end{aligned}\tag{6}$$

By applying projected gradient descent, we update the j-th dictionary atom d_k^j as follows:

$$\begin{cases}s_k^j=d_k^j-\frac{1}{\mu}\nabla_{d_k^j}\mathcal{F}(W,D_k),\\ d_k^j=\arg\min_{\|d_k^j\|_2^2=1}\|d_k^j-s_k^j\|_2=\frac{s_k^j}{\|s_k^j\|_2},\end{cases}\tag{7}$$

where μ is the step size, $\mathcal{F}(W,D_k)=\|WXQ_k-D_kAQ_kZ_k\|_{\mathrm{F}}^2$.

Learning Low-Rank Embedding: When D_k is fixed, we could update W,Γ.

Update W:

$$\begin{aligned}W=\arg\min_{W}\sum_{k=1}^{K}\|WXQ_k-D_kAQ_kZ_k\|_{\mathrm{F}}^2\\ +\alpha\mathrm{tr}(\Gamma^\top WW^\top\Gamma).\end{aligned}\tag{8}$$

We then calculate the deviation to W and set it to zero:

$$\begin{aligned}&\sum_{k=1}^{K}(WXQ_k-D_kAQ_kZ_k)(XQ_k)^\top+\alpha\Gamma\Gamma^\top W=0,\\ \Rightarrow\ &W\sum_{k=1}^{K}XQ_kX^\top+\alpha\Gamma\Gamma^\top W=\sum_{k=1}^{K}D_kAQ_kZ_kQ_kX^\top,\end{aligned}\tag{9}$$

which is a standard Sylvester equation, that can be effectively addressed through existing tools such as the BartelsStewart algorithm [2].

Update Γ:

When W is updated, we could optimize Γ with the eigenvectors related to the $(d-r)$-smallest singular values of $WW^\top$. To compute Γ, we require singular value

Algorithm 1 Solving Problem (5)

Input: X, A, Z_k, Q_k, α

Initialize: $W, D_k, \Gamma, \mu = 10^{-1}, \epsilon = 10^{-5}, t = 0.$

while not converged **do**

1. Optimize D_k via Eq. (6) by fixing others.
2. Optimize W via Eq. (9) by fixing others.
3. Optimize $\Gamma\Gamma^\top$ by fixing others.
4. Check the convergence conditions $|\mathcal{J}_{t+1} - \mathcal{J}_t| < \epsilon$.
5. $t = t + 1$.

end while

output: W, D_k.

decomposition (SVD) of $WW^\top$. Suppose singular value decomposition of $WW^\top = U_w \Sigma_w U_w^\top$, we further define $U_w = [U_w^1, U_w^2]$, in which $U_w^1 \in \mathbb{R}^{d\times(d-r)}$ and $U_w^2 \in \mathbb{R}^{d\times r}$, and therefore, we could easily obtain $\Gamma_{t+1} = U_w^1$.

Actually, we do not need to directly calculate Γ, but rather to compute the values of $\Gamma\Gamma^\top$. Given the fact that $U_w U_w^\top = U_w^1 {U_w^1}^\top + U_w^2 {U_w^2}^\top = \mathrm{I}_d$, we have $\Gamma\Gamma^\top = \mathrm{I}_d - U_w^2 {U_w^2}^\top$. Since $WW^\top$ is a matrix with low-rankness, r should be a small value ($r \ll d$). A direct computing of Γ would cost $\mathcal{O}((d-r)^2 d) \approx \mathcal{O}(d^3)$ due to the matrix multiplication of $\Gamma_{t+1}\Gamma_{t+1}^\top$. Thanks to a simpler matrix multiplication $U_w^2 {U_w^2}^\top$, our newly optimized approach would only cost $\mathcal{O}(r^2 d) \approx \mathcal{O}(d)$.

So far, we build the optimization rules for all the variables. Then, we iteratively update all the variables until converge. For clarity, we list the detailed steps of the optimization in **Algorithm 1**.

3.5. Zero-shot Learning via Ensemble

In zero-shot learning scenario, we need to verify the predicted class label given reference data. Given a test data x_u^i and semantic representation A_u with C_u classes, we could use reconstruction error with semantic dictionary to assign the label to x_u^i in the following way:

$$c_u^k = \arg\min_{c=1}^{C_u} \|W x_u^i - D_k A_u^c\|_2^2, \tag{10}$$

where A_u^c is the average semantic representation for class c and c_u^k is the prediction result of x_u^i on the k-th semantic dictionary. Then we adopt voting strategy to obtain the final result. For all classes, we measure the overall recognition performance in terms of accuracy.

4. Experiment

In this section, we experiment on popular ZSL benchmarks to testify the proposed approach by comparing it with several state-of-the-art ZSL approaches.

4.1. Dataset & Experimental Setting

Four standard benchmarks are experimented for zero-shot learning and their statistics are listed in Table 1.

Table 1. Statistics of the 4 ZSL benchmarks.

Dataset	aPaY	AwA	CUB	SUN
#Training classes	20	40	150	707
#Test classes	12	10	50	10
#Instances	15,339	30,475	11,788	14,340
#Attributes	64	85	312	102

aPascal-aYahoo (aP&aY) [8] contains 20 objects classes from the PASCAL VOC 2008 dataset and 12 object classes collected with the Yahoo image search engine. Following previous work [28, 36, 37, 3], we treat PASCAL VOC 2008 as seen data for model training, and evaluate on Yahoo images. Specifically, there are 64 attributes shared by two datasets to describe the object images.

Animal with Attribute (AwA) [16] includes 50 animals categories, each with over 92 instances. Each category is paired with a human annotated 85-attribute semantic feature.

Caltech-UCSD Birds-200-2011 (CUB) [32] is a fine-grained bird dataset with 200 different bird species and 11,788 image samples. For semantic representation, there are 312 visual attributes to annotate those birds in class level.

SUN scene attribute dataset (SUN) [22] is a fine-grained dataset, which shows less variations across different classes. There are 717 scene categories, each with 20 images. In total, 102 attributes are adopted to annotate those images.

In fact, each sample from the aP&aY, CUB and SUN benchmarks has its specific attribute description, that is, any two samples within the same class could have relatively different descriptions. However, for AwA, all the samples from the same class share a single class-wise description. We adopt the continuous attributes as the semantic representation since it works better than the binary one [4].

Regarding the representation of images, we adopt the following deep features: AlexNet [14], VGG-VeryDeep-19 [29], and GoogLeNet [31]. Specifically, for AlexNet, we take the 7-th layer (FC_7) as visual features with dimensions 4,096. For VGG-VeryDeep-19, we adopt the top layer as visual features with 4,096-dimensional activations[3]. For GoogLeNet, we utilize the 1,024-dimensional units as visual features [4].

In our experiments, we follow previous ZSL approaches to tune the parameters using cross-validation [28, 36, 37, 3]. Specifically, we split the seen training data into three subsets and then choose the parameters based on the performance of one subset with other two as training. We repeat three times and report the average evaluation accuracies.

[3] https://zimingzhang.files.wordpress.com/2014/10/cnn-features.key

Table 2. Zero-shot classification accuracy (%) of the comparisons on the four datasets. There are three kinds of features: AlexNet CNN features **[ALEX]**, VGG-VeryDeep-19 CNN features **[VGG]** and GoogLeNet features **[GGL]**.

Features	Methods	aP&aY	AwA	CUB	SUN
[VGG]	DAP [16]	38.2	57.2	39.8	72.0
	ESZSL [28]	24.2	75.3	-	82.1
	SSE [36]	46.2	76.3	30.4	82.5
	JSLE [37]	50.4	**80.5**	42.1	83.8
	ISEC [3]	53.2	77.3	43.3	84.4
	KDICA [11]	-	73.8	43.7	-
	Ours	**55.2**	82.8	45.2	86.0
[ALEX]	DAP [16]	-	53.2	31.4	-
	UDA [13]	-	73.2	39.5	-
	COSTA [19]	-	55.2	36.9	-
	SJE [1]	-	61.9	40.3	-
	ESZSL [28]	-	53.2	37.2	-
	ISEC [3]	46.1	-	42.0	75.5
	SynC [4]	-	64.8	47.1	-
	Ours	48.9	71.4	43.9	77.1
[GGL]	DAP [16]	-	60.5	39.1	-
	COSTA [19]	-	61.8	40.8	47.9
	SJE [1]	-	66.7	50.1	87.0
	ESZSL [28]	-	59.6	44.0	82.1
	SynC [4]	-	72.9	**54.5**	90.0
	Ours	58.8	76.6	56.2	**88.3**

4.2. Zero-Shot Classification

In this part, we mainly compare with several state-of-the-art zero-shot learning methods, including: DAP [16], ESZSL [28], SSE [36], JSLE [37], ISEC [3], KDICA [11], UDA [13], COSTA [19], SJE [1] and SynC [4]. Note that partial results are directly cropped from the published papers. The classification performance in term of accuracy is listed in Table 2.

From the results, we notice that the proposed approach outperforms other competitors in most cases of four datasets with remarkable margins. Compared with three kinds of visual features, we notice that VGG-VeryDeep-19 and GoogLeNet features work better than AlexNet CNN FC_7, which indicates that these two deep features are more powerful in representing images. Comparing VGG-VeryDeep-19 and GoogLeNet features, we observe that VGG-VeryDeep-19 shows superiority on AwA dataset, while GoogLeNet features are more effective in aP&aY, CUB and SUN. Furthermore, we also notice that all models work better on AwA and SUN than on aP&aY and CUB. The reason we consider is that the class connection in AwA and SUN is much stronger than in aP&aY, since AwA only includes animal classes, SUN only contains scene classes, whereas aP&aY consists of random object classes. Thus, it is easier to capture the shared information across the categories in AwA/SUN than in aP&aY. Besides, the semantic attributes of AwA are provided to tailor for animals with special descriptions, however, the provided attributes of aP&aY cannot describe an object comprehensively. In this way, more effective information could be adapted from the observed categories to the unobserved ones on AwA than on aP&aY. For CUB, there are 200 bird species and some birds are very similar. Therefore, CUB is a very challenging dataset for ZSL.

Table 3. Performance of our approach with various size of seen/unseen categories on the CUB dataset.

	$C(C_u) = 50$	$C(C_u) = 100$	$C(C_u) = 150$
$C_u = 50$	40.6	51.2	55.9
$C = 50$	40.2	38.2	29.3

Moreover, we further visualize the zero-shot classification results of the proposed approach in term of the confusion matrices (Figure 2), where we experiments on aP&aY and AwA using VGG-VeryDeep-19 features. In each confusion matrix, the column denotes the ground truth and the row represents the predicted results. Seen from the confusion matrix for aP&aY, we notice that our model presents appealing results on certain classes, e.g., donkey (58.54%) and centaur (60.18%). While for AwA, we observe from the confusion matrix that our algorithm achieves over 80% accuracy for some animal classes, e.g., leopard (84.21%) and rat (83.08%). Considering the fact that we have no data from these test classes to train our model, it strongly supports the superiority of our proposed approach for effective zero-shot learning.

4.3. Qualitative Results

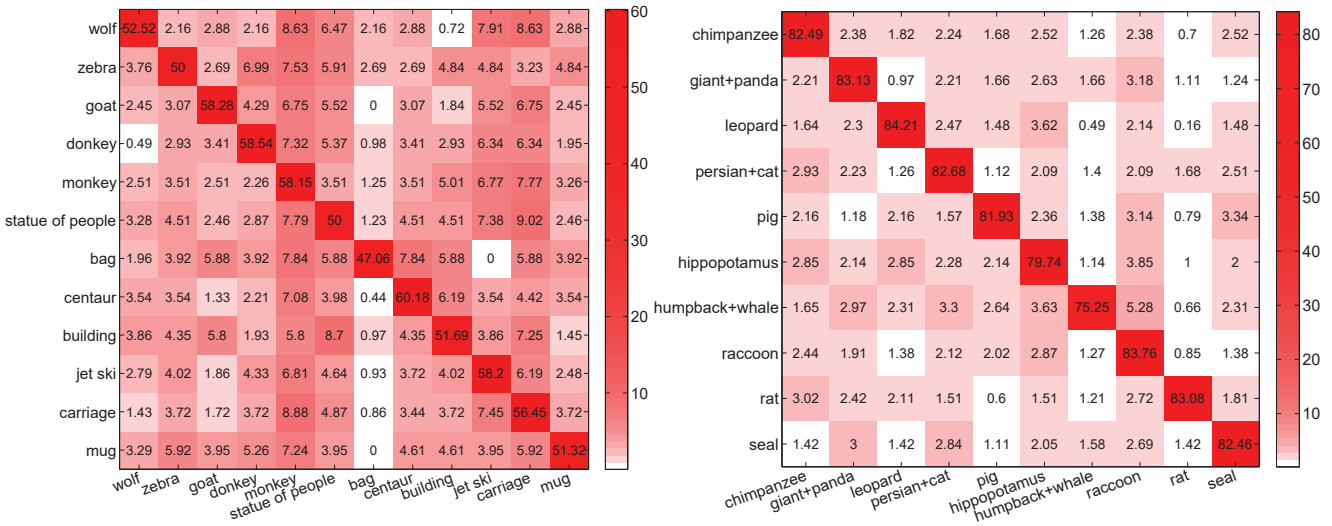
We further provide some qualitative analysis for our proposed algorithm. Specifically, we show what kind of visual information the model captures for unseen categories.

Figure 3 and 4 present 10 categories of the unseen test data from CUB and SUN, where we report the Top-5 samples classified into each category using GoogLeNet features. From the top retrieved images, we can witness that our model can reasonably capture discriminative visual information for each unseen category. Furthermore, we notice that the misclassified images have the similar appearance to that of predicted category which even humans are unable to easily distinguish them.

4.4. Evaluation on the Size of Seen/Unseen Classes

In this part, we evaluate our model under different number of seen/unseen classes on the CUB dataset (GoogLeNet features).

First of all, we testify the performance of zero-shot learning with various sizes of unseen classes (e.g., 50, 100, 150) by fixing the number of unseen categories as 50. We randomly select 10 times of seen/unseen categories. Interestingly, we notice that increasing the size of seen classes during the training stage results in better accuracy shown in Table 3.



(a) aP&aY (b) AwA

Figure 2. Confusion matrices of the classification accuracy on unobserved categories for our approach on (a) aP&aY and (b) AwA, where diagonal position indicates the classification accuracy. Column means the ground truth and row denotes the predicted results.

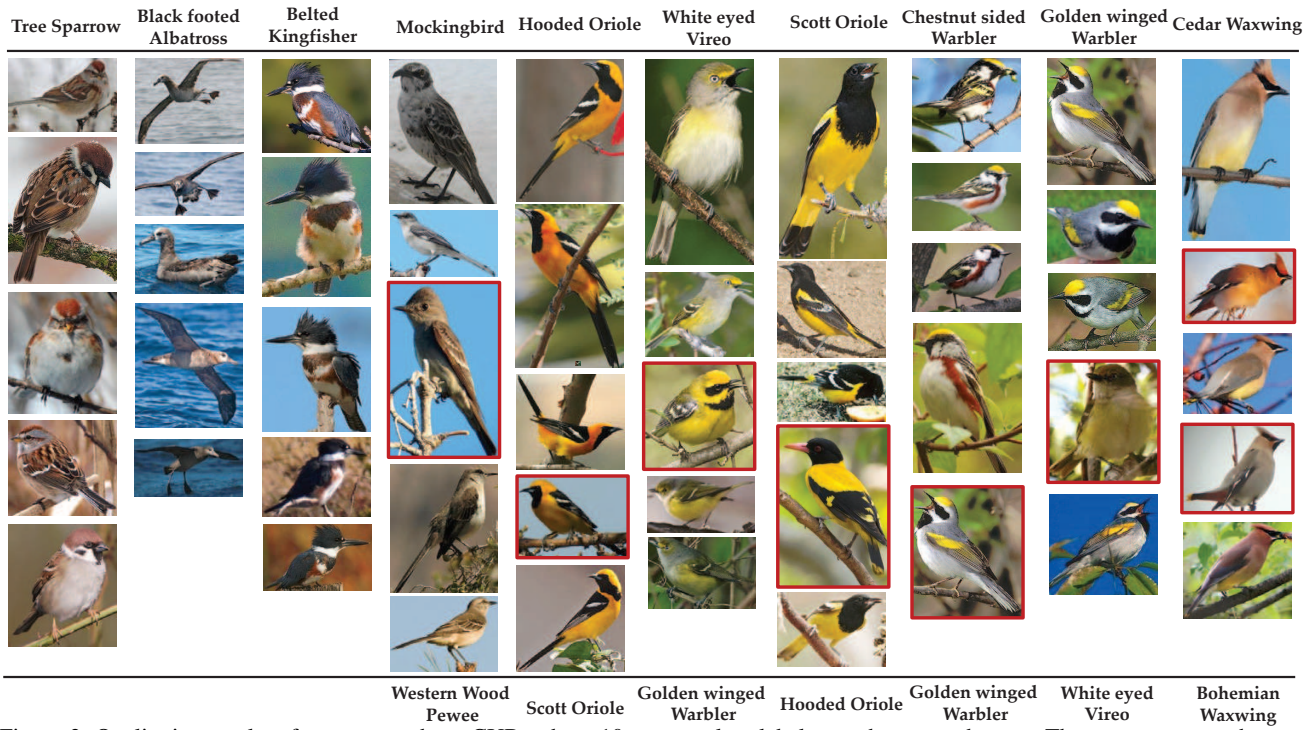


Figure 3. Qualitative results of our approach on CUB, where 10 unseen class labels are shown on the top. Then we represent the top-5 images recognized in each class in the middle, where misclassified images are marked with red bounding boxes. The misclassified class labels are listed in the bottom part.

Secondly, we show the effectiveness of our proposed algorithm under various sizes of unseen classes (e.g., 50, 100, 150), with the size of seen categories fixed as 50 during model learning. We also repeat 10 times to build the seen and unseen categories. The performance in terms of average accuracy are reported in Table 3, where we notice the performance decreases with more unseen categories involved.

4.5. Empirical Analysis

We further testify some properties of our proposed model on four datasets with VGG-VeryDeep-19 features. We also testify the defectiveness of our low-rank projection, by removing rank constraint with Frobenius norm on W.

From the parameter analysis on α (Figure 5 (a)), we ob-

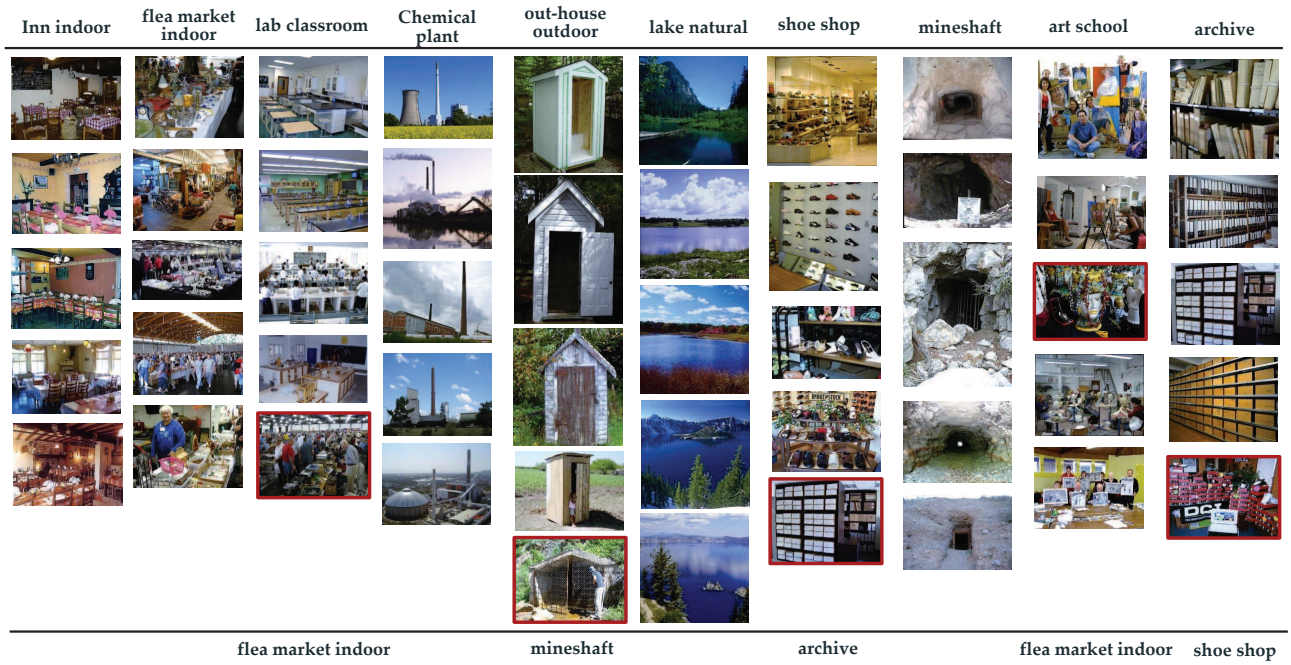
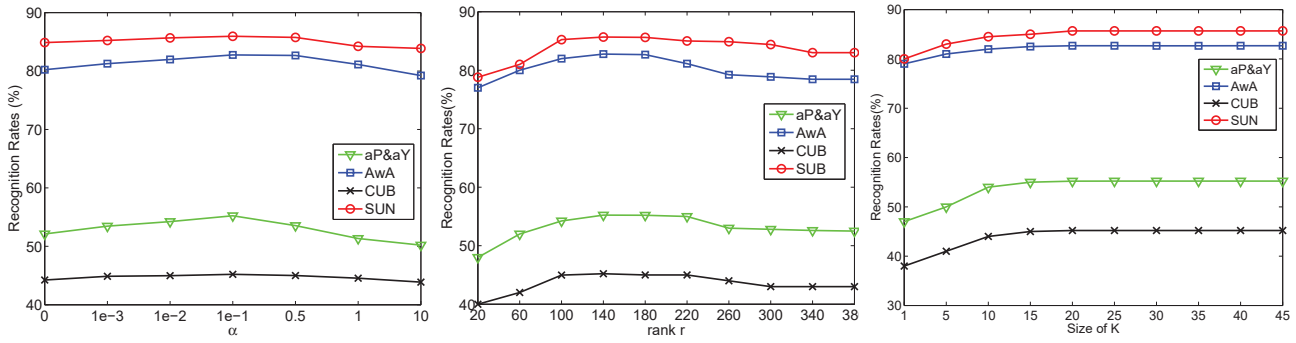


Figure 4. Qualitative results of our approach on SUN, where 10 unseen class labels are shown on the top. Then we represent the top-5 images recognized in each class in the middle, where misclassified samples are marked with red bounding boxes. The misclassified class labels are listed in the bottom part.



(a) parameter analysis on α (b) rank r (c) size K

Figure 5. (a) parameter α analysis, (b) rank r analysis and (c) evaluation of sampling size K on four benchmarks with VGG-VeryDeep-19.

serve that our model can achieve better performance around $\alpha = 0.1$ on four datasets. We further evaluate on $\alpha = 0$, meaning we remove rank constraint on W and replace with Frobenius norm. It verifies the effectiveness of rank constraint term.

From the analysis on rank r, we notice that when r is set around 140 to 180, the classification accuracy tends to better (Figure 5 (b)). While r is set too large or too small, the classification performance would both degrade. This demonstrates that a low-rank projection would benefit the zero-shot learning.

Moreover, we evaluate the impact of sampling size K. Figure 5 (c) shows the performance would increase when enlarging K. Specifically, $K = 1$ denotes we only learn one semantic dictionary from seen classes. Clearly, one semantic dictionary is not able to well capture the latent semantic dictionary for unseen classes. Generally, $K = 15$ is good enough to sample the space of the latent semantic dictionary based on our experiments.

5. Conclusion

In this paper, we proposed a novel low-rank embedded semantic dictionary learning through ensemble strategy for zero-shot learning challenges. Specifically, we developed an effective model for knowledge transfer by integrating low-rank embedding and semantic dictionary learning into a unified framework. In this way, the semantic gap across visual features and semantic representations would be mitigated. Moreover, ensemble strategy was exploited to build multiple semantic dictionaries to constitute the latent basis for the unseen classes. Experiments on four ZSL benchmarks verified the effectiveness of our designed approach.

References

[1] 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.

[2] R. H. Bartels and G. Stewart. Solution of the matrix equation ax+ xb= c [f4]. *Communications of the ACM*, 15(9):820–826, 1972.

[3] M. Bucher, S. Herbin, and F. Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *ECCV*, pages 730–746. Springer, 2016.

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, June 2016.

[5] Z. Ding and Y. Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119. IEEE, 2014.

[6] Z. Ding, M. Shao, and Y. Fu. Latent low-rank transfer subspace learning for missing modality recognition. In *AAAI*, 2014.

[7] Z. Ding, M. Shao, and Y. Fu. Deep robust encoder through locality preserving low-rank dictionary. In *ECCV*, pages 567–582. Springer, 2016.

[8] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785. IEEE, 2009.

[9] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE TPAMI*, 37(11):2332–2345, 2015.

[10] Z. Fu, T. Xiang, E. Kodirov, and S. Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015.

[11] C. Gan, T. Yang, and B. Gong. Learning attributes equals multi-source domain generalization. In *CVPR*, pages 87–97, June 2016.

[12] Y. Hu, D. Zhang, J. Ye, X. Li, and X. He. Fast and accurate matrix completion via truncated nuclear norm regularization. *IEEE TPAMI*, 35(9):2117–2130, 2013.

[13] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, pages 2452–2460, 2015.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

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

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 36(3):453–465, 2014.

[17] X. Li, Y. Guo, and D. Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, pages 4211–4219, 2015.

[18] H. Liu, M. Shao, S. Li, and Y. Fu. Infinite ensemble for image clustering. In *KDD*, pages 1745–1754. ACM, 2016.

[19] T. Mensink, E. Gavves, and C. G. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, pages 2441–2448, 2014.

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

[21] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, pages 503–510. IEEE, 2011.

[22] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758. IEEE, 2012.

[23] P. Peng, Y. Tian, T. Xiang, Y. Wang, and T. Huang. Joint learning of semantic and latent attributes. In *ECCV*, pages 336–353. Springer, 2016.

[24] G.-J. Qi, W. Liu, C. Aggarwal, and T. S. Huang. Joint intermodal and intramodal label transfers for extremely rare or unseen classes. *IEEE TPAMI*, 2016.

[25] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, pages 2249–2257, 2016.

[26] Y. Quan, Y. Xu, Y. Sun, Y. Huang, and H. Ji. Sparse coding for classification via discrimination ensemble. In *CVPR*, pages 5839–5847, 2016.

[27] M. Radovanović, A. Nanopoulos, and M. Ivanović. Hubs in space: Popular nearest neighbors in high-dimensional data. *JMLR*, 11(Sep):2487–2531, 2010.

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

[29] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[31] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[32] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[33] X. Xu, T. M. Hospedales, and S. Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *ECCV*, pages 343–359. Springer, 2016.

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

[35] W. Zhang, A. Surve, X. Fern, and T. Dietterich. Learning non-redundant codebooks for classifying complex objects. In *ICML*, pages 1241–1248. ACM, 2009.

[36] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, pages 4166–4174, 2015.

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

[38] N. Zhou, Y. Shen, J. Peng, and J. Fan. Learning inter-related visual dictionary for object recognition. In *CVPR*, pages 3490–3497. IEEE, 2012.