Zero Shot Learning via Low-rank Embedded Semantic AutoEncoder

Yang Liu[1], **Quanxue Gao**[1*], **Jin Li**[2], **Jungong Han**[3], **Ling Shao**[4]
[1] State Key Laboratory of Integrated Services Networks, Xidian University, Xi’an, China
[2] Institute of Artificial Intelligence and Robotics, Xi’an Jiaotong University, Xi’an, China
[3] School of Computing and Communications, Lancaster University, United Kingdom
[4] Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates
{xidianliuyang, j.lixjtu, jungonghan77}@gmail.com, qxgao@xidian.edu.cn, ling.shao@ieee.org

Abstract

Zero-shot learning (ZSL) has been widely researched and get successful in machine learning. Most existing ZSL methods aim to accurately recognize objects of unseen classes by learning a shared mapping from the feature space to a semantic space. However, such methods did not investigate in-depth whether the mapping can precisely reconstruct the original visual feature. Motivated by the fact that the data have low intrinsic dimensionality e.g. low-dimensional subspace. In this paper, we formulate a novel framework named Low-rank Embedded Semantic AutoEncoder (LESAE) to jointly seek a low-rank mapping to link visual features with their semantic representations. Taking the encoder-decoder paradigm, the encoder part aims to learn a low-rank mapping from the visual feature to the semantic space, while decoder part manages to reconstruct the original data with the learned mapping. In addition, a non-greedy iterative algorithm is adopted to solve our model. Extensive experiments on six benchmark datasets demonstrate its superiority over several state-of-the-art algorithms.

1 Introduction

In recent years, along with the explosive growth of web data, there has been significant progress in large-scale classification with aid of conventional frameworks, e.g. Deep Neural Networks (DNN) [Krizhevsky *et al.*, 2012]. Conventional frameworks mainly depend on a large number of training samples to build robust models. However, as obtaining well-annotated training samples are labor-intensive and time-consuming, sufficient labeled training samples are usually unavailable in many real-world situations. Meanwhile, the number of newly defined classes is ever-growing, indicating that training a particular model for each of them is unattainable.

Zero-shot learning (ZSL) [Palatucci *et al.*, 2009] has been widely recognized as a feasible solution to deal with these problems. ZSL attempts to learn mechanism of human brain and recognize new classes which are not observed in the training stage. For instance, one can easily recognize a new species of objects after being told how it is similar to or different from other observed objects. The main reason is that humans can explore the relationship across different objects, and adapt the knowledge from seen classes to unseen ones. Likewise, ZSL aims to uncover the intrinsic relationship between seen and unseen classes. Specifically, the fundamental idea of ZSL is to learn a general mapping from the feature space to a semantic space using the labelled training samples consisting of seen classes only. This mapping is then used to project the visual representation of the unseen class images into the semantic space. Hereafter, the task of unseen class recognition becomes a typical classification problem which can be realized by a simple nearest neighbour (NN) search.

However, most existing ZSL methods neglect the importance of reconstruction. They put much attention on learning a projection only from the feature space to a semantic space instead of considering reconstructing the original visual feature representation. This can lead to a projection domain shift problem [Fu *et al.*, 2015a]. Sometimes this shift will adversely affect the final classification results. Recently, [Kodirov *et al.*, 2017] proposed a model named Semantic AutoEncoder (SAE) which imposes a new constraint in learning a projection from the visual space to the semantic space so that the projection must also preserve all the information contained in the original visual features. However, this constraint cannot guarantee the reconstructed data has a low-rank structure, which is important for undercomplete autoencoders [Xie *et al.*, 2016] like Principal Component Analysis (PCA) [Jolliffe, 1986]. In this paper, we proposed a model named Low-rank Embedded Semantic AutoEncoder (LESAE). We assume that the latent semantic space for unseen samples should share its majority with semantic space for the seen samples, which should be identified in the low-rank embedding space. Taking the encoder-decoder paradigm, the encoder part tries to learn a low-rank projection from the feature space to a semantic space as in the existing ZSL models. The decoder part aims to the learned mapping can reconstruct the original visual features precisely.

We summarize our main contributions as follows: (1) We

[*] Corresponding author: Q. Gao. (qxgao@xidian.edu.cn)

build a bridge between reconstruction and Low-rank representation to capture shared discriminative features across seen and unseen classes. A robust model is proposed for zero-shot learning (ZSL) and generalized zero-shot learning (GZSL). (2) We formalize ZSL as the problem of learning a low-rank semantic representation of input data that can also be used for data reconstruction. (3) An efficient iterative algorithm based on Sylvester equation is introduced to solve this model and leads to state-of-the-art recognition performance on six benchmark datasets for ZSL and GZSL.

2 Related Works

Low-rank embedding This technique aims to recover the low-rank clean data from the corruption data and has been successfully applied to many applications including image clustering [Chen *et al.*, 2018; Li *et al.*, 2018], image classification [Liao *et al.*, 2018; Liu *et al.*, 2018], object tracking [Zhang *et al.*, 2015; Yang *et al.*, 2017], hyperspectral image denoising [Fan *et al.*, 2017] and dynamic MRI [Nakarmi *et al.*, 2017]. Robust Principal Component Analysis (RPCA) [Candès *et al.*, 2011] is of the most representative methods. This model was demonstrated that PCA can be made robust to outliers if exactly recovering the low-rank representation by solving a simple convex problem. Similarly, Low-Rank Representation (LRR) [Liu *et al.*, 2013] aims to seek the lowest rank representation among all the candidates that can represent the data samples as linear combinations of the bases in a given dictionary. [Peng *et al.*, 2012] proposed RASL to seek an optimal set of image domain transformations such that the matrix of transformed images can be decomposed as the sum of a sparse matrix of errors and a low-rank matrix of recovered aligned images. Recently, [Nie *et al.*, 2017] proposed a novel model named Multi-view Learning with Adaptive Neighbours (MLAN). With the reasonable rank constrain, the obtained optimal graph can be partitioned into specific clusters directly. Inspired by these successful approaches, we apply low-rank embedding into zero-short learning in the paper.

Zero-short learning A variety of approaches for zero-shot learning have been recently proposed. To circumvent learning independent attributes, embedding-based ZSL frameworks are proposed to learn a projection that can map the visual space to semantic space at once. The class label is then determined in the semantic space using various relatedness measurements [Akata *et al.*, 2013; Socher *et al.*, 2013; Zhang and Saligrama, 2016]. In addition to directly using fixed semantic embedding, some work tries to map them into a different space by sparse coding [Zhang and Saligrama, 2015; Kodirov *et al.*, 2015] and CCA [Fu *et al.*, 2015a]. Recent work [Long and Shao, 2017; Ding *et al.*, 2017; Long *et al.*, 2017; Kodirov *et al.*, 2017] combines the embedding-inferring procedure into a unified framework and empirically demonstrates better performance.

In our evaluation we choose following representative methods for comparison on several benchmark datasets. DAP [Lampert *et al.*, 2014] and IAP [Lampert *et al.*, 2014] are two of the most fundamental methods in ZSL research. Such models utilize semantic attributes as intermediate clues. CONSE [Norouzi *et al.*, 2014] is one of the most widely used representatives of learning a mixture of class proportions. SSE [Zhang and Saligrama, 2015] uses the mixture of seen class proportions as the common space and leverages similar class relationships both in the visual space and the semantic space. SJE [Akata *et al.*, 2015] and ESZSL [Romera-Paredes and Torr, 2015] learn the bilinear compatibility function between the visual and the semantic space optimizing the structural SVM loss and square loss respectively. ALE [Akata *et al.*, 2016] and DEVISE [Frome *et al.*, 2013] learn the bilinear compatibility with similar ranking loss functions. LATEM [Xian *et al.*, 2016] gives non-linear extension to bilinear compatibility learning methods. CMT [Socher *et al.*, 2013] aims to learn a non-linear projection from visual space to semantic space by a neural network. SYNC [Changpinyo *et al.*, 2016] constructs classifiers for unseen classes by a linear combination of base classifiers. SAE [Kodirov *et al.*, 2017] learns a low dimensional semantic representation of input data that can be used for data reconstruction. SS-Voc [Fu and Sigal, 2016] utilizes open set semantic vocabulary to help train better classifiers in supervised learning. AMP [Fu *et al.*, 2015b] computes semantic manifold distance by a absorbing Markov chain process.

3 Approach

For ZSL task, we aim to classify the samples from unseen classes according to their class-level attributes, where both the samples and labels of unseen classes are totally independent from the training phase.

Suppose there are c seen classes with n labeled samples $\mathbf{S} = \{\mathbf{X}, \mathbf{A}, \mathbf{Y}\}$ and c_u unseen classes with n_u unlabeled samples $\mathbf{U} = \{\mathbf{X}_u, \mathbf{A}_u, \mathbf{Y}_u\}$, where $\mathbf{X} \in \mathbf{R}^{m\times n}$ and $\mathbf{X}_u \in \mathbf{R}^{m\times n_u}$ are m-dimensional visual features in the seen and unseen data, while their corresponding class labels are $\mathbf{Y}$ and $\mathbf{Y}_u$, respectively. The seen and unseen classes have no label overlap, i.e., $\mathbf{Y} \cap \mathbf{Y}_u = \emptyset$. $\mathbf{A} \in \mathbf{R}^{d\times n}$ and $\mathbf{A}_u \in \mathbf{R}^{d\times n_u}$ are d-dimensional semantic representations of instances in the seen and unseen datasets. In the zero-shot learning task, we aim to learn a classifier $f : \mathbf{X}_u \rightarrow \mathbf{Y}_u$, where the samples in $\mathbf{X}_u$ are completely unavailable during training.

The intuition behind ZSL is that the classifier would be able to capture the relationship between the visual space and the the semantic space. Inspired by the recent work [Kodirov *et al.*, 2017] considering the transpose of projection matrix as a decoder to reconstruct the original visual feature, we develop an effective Low-rank Embedded Semantic AutoEncoder (LESAE) that integrates the merits of both low-rank discriminative embedding and semantic representation learning. Specifically, LESAE tries to learn a Low-rank projection matrix $\mathbf{W} \in \mathbf{R}^{d\times m}$ $(d < m)$ from the feature space $\mathbf{X}$ to the semantic space $\mathbf{A}$. At the same time, the semantic space can be projected back to the feature space with $\mathbf{W}^T \in \mathbf{R}^{k\times d}$ to reconstruct the input data exactly. This can be achieved by optimising the following function:

$$\begin{aligned} &\min_{\mathbf{W}} \left\|\mathbf{X} - \mathbf{W}^T\mathbf{W}\mathbf{X}\right\|_F^2 + \beta rank(\mathbf{W}) \\ &s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{A} \end{aligned} \tag{1}$$

where $\|\cdot\|_*$ denotes Frobenius norm of a matrix, β is the balance parameter, $rank(\cdot)$ is the rank operator of a matrix.

Rank minimization in Eq. (1) is a well-known NP hard problem, and considerable approaches have been proposed. One of appealing strategies is to adopt trace norm $\|\mathbf{W}\|_*$ as a surrogate of the term $rank(\mathbf{W})$. Instead of adapting the traditional Singular Value Thresholding (SVT) [Cai *et al.*, 2010] to solve the objective, we exploit a regularization term that guarantees that the low-rank feature of optimized $\mathbf{W}$. Mathematically, we have following equation:

$$\|\mathbf{W}\|_* = tr((\mathbf{W}^T\mathbf{W})^{\frac{1}{2}}) = tr(\mathbf{W}^T(\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}\mathbf{W}) \quad (2)$$

According to Eq. (1) and Eq. (2), the new formula with low-rank constraint can be rewritten as:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{H}} \left\|\mathbf{X} - \mathbf{W}^T\mathbf{A}\right\|_F^2 + \beta tr(\mathbf{W}\mathbf{H}\mathbf{W}^T) \\ &s.t. \quad \mathbf{W}\mathbf{X} = \mathbf{A} \end{aligned} \quad (3)$$

where $\mathbf{H} = (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}$.

Solving the objective function (3) with a hard constraint such as $\mathbf{W}\mathbf{X} = \mathbf{A}$ is difficult. Therefore, we consider to relax the constraint into a soft one and rewrite the objective function (3) as:

$$\min_{\mathbf{W},\mathbf{H}} \left\|\mathbf{X} - \mathbf{W}^T\mathbf{A}\right\|_F^2 + \alpha \left\|\mathbf{W}\mathbf{X} - \mathbf{A}\right\|_F^2 + \beta tr(\mathbf{W}\mathbf{H}\mathbf{W}^T) \quad (4)$$

3.1 Optimization Algorithms

The objective function (4) has two unknown variables. It is difficult to directly solve the solution. An algorithm can be developed for alternatively updating $\mathbf{W}$ (while fixing $\mathbf{H}$) and $\mathbf{H}$ (while fixing $\mathbf{W}$). Here we develop an efficient algorithm to solve this problem.

Update $\mathbf{H}$: When $\mathbf{W}$ is fixed, $\mathbf{H}$ can be computed by $\mathbf{H} = (\mathbf{W}\mathbf{W}^T)^{-\frac{1}{2}}$.

Update $\mathbf{W}$: When $\mathbf{H}$ is fixed, taking the derivative of Eq. (4) with respective to $\mathbf{W}$ and setting it to zero, we have:

$$\mathbf{M}\mathbf{W} + \mathbf{W}\mathbf{N} = \mathbf{C} \quad (5)$$

where

$$\begin{aligned} &\mathbf{M} = \mathbf{A}\mathbf{A}^T \\ &\mathbf{N} = \alpha\mathbf{X}\mathbf{X}^T + \beta\mathbf{H} \\ &\mathbf{C} = (\alpha + 1)\mathbf{A}\mathbf{X}^T \end{aligned} \quad (6)$$

Before solving Eq. (5), we first introduce one definition and two theorems:

Definition 1 [Sylvester, 1884]: A Sylvester equation is a matrix equation of the form:

$$\mathbf{P}\mathbf{X} + \mathbf{X}\mathbf{Q} = \mathbf{D} \quad (7)$$

Theorem 1 [Lancaster and Tismenetsky, 1985]: The sufficient and unnecessary condition for Eq. (7) having a solution is that the matrices $\begin{bmatrix} \mathbf{P} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} \end{bmatrix}$ and $\begin{bmatrix} \mathbf{P} & \mathbf{D} \\ \mathbf{0} & -\mathbf{Q} \end{bmatrix}$ are similar.

Theorem 2 [Lancaster and Tismenetsky, 1985]: The sufficient and unnecessary condition for Eq. (7) having a unique solution is that the eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$ of $\mathbf{P}$ and $\eta_1, \eta_2, \cdots, \eta_k$ of $\mathbf{Q}$ satisfy $\lambda_i + \eta_j \neq 0$; $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, k)$.

Detailed explanation of Definition 1 and proofs of Theorem 1 and Theorem 2 can refer to [Lancaster and Tismenetsky, 1985].

According to Definition 1, it is easy to know Eq. (5) is a Sylvester equation. So Eq. (5) has a unique solution if it meets conditions of Theorem 1 and Theorem 2, which can be easily satisfied in the real-world zero-short learning.

Vectorizing the unknown matrix $\mathbf{W}$, Eq. (5) can be transformed to a linear equation:

$$(\mathbf{I}_m \otimes \mathbf{M} + \mathbf{N}^T \otimes \mathbf{I}_d)vec(\mathbf{W}) = vec(\mathbf{C}) \quad (8)$$

where $\otimes$ is the Kronecker product, $\mathbf{I}_m \in \mathbf{R}^{m\times m}$ and $\mathbf{I}_d \in \mathbf{R}^{d\times d}$ are identity matrices, $vec(\mathbf{C})$ is the vectorization of the matrix $\mathbf{C}$. Then, $\mathbf{W}$ can be obtained by following equation.

$$vec(\mathbf{W}) = (\mathbf{I}_m \otimes \mathbf{M} + \mathbf{N}^T \otimes \mathbf{I}_d)^+ vec(\mathbf{C}) \quad (9)$$

In MATLAB, it can be implemented with a single line of code: sylvester[1].

So far, we have build the optimization rules for two variables. Then, we iteratively update them until converge. For clarity, Algorithm 1 lists the pseudo code of solving our model (4). In Algorithm 1, η is a small positive parameter to ensure $\mathbf{H}$ has a solution. We terminate the algorithm 1 when the Frobenius norm of relative changes of $\mathbf{W}$ is below 10^{-6}.

```
Algorithm 1:
Input Training set $\{\mathbf{X}, \mathbf{A}\}$.
Initialize $\mathbf{W} = \mathbf{I}$
repeat
1. Update $\mathbf{H}$ by $\mathbf{H} = (\ \mathbf{W}(\mathbf{W})^T + \eta\mathbf{I})^{-\frac{1}{2}}$.
2. Update $\mathbf{W}$ by solving the Eq. (5) and Eq. (6).
until converge
Output projection matrix $\mathbf{W}$.
```

3.2 Zero-shot Recognition

Once we obtain the projection matrices $\mathbf{W}$, the visual features of unseen classes can be easily synthesized from their semantic attributes $\mathbf{A}_u$ by following equation:

$$\mathbf{X}_u = \mathbf{W}^T\mathbf{A}_u \quad (10)$$

It is noticeable that for image-level attributes, $\mathbf{X}_u$ contains as many instances as the test set. The zero-shot recognition task now becomes a typical classification problem. Thus, any existing supervised classifier can be applied. Since we focus on the quality of the synthesized features, we simply use Nearest Neighbour (NN) in the task.

4 Experiments

In this section, we will validate our proposed method on five small-scale datasets (SUN, CUB, AWA1, AWA2 and APY) and one large-scale dataset (ImageNet), compared with other state-of-the-art methods mentioned in related works.

[1] https://uk.mathworks.com/help/matlab/ref/sylvester.html

method	SUN	CUB	AWA1	AWA2	aPY
DAP	39.9	40.0	44.1	46.1	33.8
IAP	19.4	24.0	35.9	35.9	36.6
CONSE	38.8	34.3	45.6	44.5	26.9
CMT	39.9	34.6	39.5	37.9	28.0
SSE	51.5	43.9	60.1	61.0	34.0
LATEM	55.3	49.3	55.1	55.8	35.2
ALE	58.1	54.9	59.9	62.5	39.7
DEVISE	56.5	52.0	54.2	59.7	39.8
SJE	53.7	53.9	65.6	61.9	32.9
ESZSL	54.5	53.9	58.2	58.6	38.3
SYNC	56.3	**55.6**	54.0	46.6	23.9
SAE	59.7	53.6	65.4	66.2	34.5
LESAE	**60.0**	53.9	**66.1**	**68.4**	**40.8**

Table 1: Zero-shot learning (ZSL) results on SUN, CUB, AWA1, AWA2 and aPY using ResNet features. The results report Top-1 accuracy in %.

4.1 Datasets Descriptions

Five small-scale datasets: SUN Attribute (SUN) [Patterson *et al.*, 2014] is a fine-grained dataset, which contains 14,340 images coming from 717 types of scenes annotated with 102 attributes. Following [Lampert *et al.*, 2014], 645 out of 717 classes are used for training and rest 72 classes for testing.

CUB-200-2011 Birds (CUB) [Wah *et al.*, 2011] is a fine-grained and medium scale dataset with respect to both number of images and number of classes, i.e. 11,788 images from 200 different types of birds annotated with 312 attributes. [Akata *et al.*, 2016] introduces the first zero-shot split of CUB with 150 train classes and 50 test classes.

Animals with Attributes 1 (AWA1) [Lampert *et al.*, 2014] is a coarse-grained dataset, which has totally 30,475 images and 85-dim class-level attributes, in which 40 classes are used for training and 10 others for testing.

Animals with Attributes 2 (AWA2) [Xian *et al.*, 2017] uses the same 50 animal classes as AWA1 dataset, while 37,322 images are collected from the public open source. Compared to AWA1, AWA2 dataset contains more images, e.g. horse and dolphin among the test classes, antelope and cow among the training classes. Same as AWA1, 40 classes are used for training and 10 others for testing in AWA2 dataset.

A Pascal and Yahoo (aPY) [Farhadi *et al.*, 2009] is a small-scale coarse-grained dataset with 64 attributes. Among the total number of 32 classes, 20 Pascal classes (we randomly select 5 for validation) and 12 Yahoo classes are used for training and testing, respectively.

One large-scale dataset: ImageNet [Russakovsky *et al.*, 2015] has totally 218,000 images and 1000-dim class-level attributes. In this large-scale dataset, as in [Fu and Sigal, 2016], 1,000 classes of ILSVRC2012 are used as seen classes, while 360 classes of ILSVRC2010, which are not included in ILSVRC2012, for unseen classes.

4.2 Evaluations

For five small-scale datasets, we follow the settings of [Xian *et al.*, 2017] to make sure the absence of any image from test classes during training. Then the 2048-dim feature of each image is extracted from the 101-layered ResNet [He *et al.*, 2016]. As class embeddings, we use per-class attributes. All the features and attributes used in experiments are published for open access[2].

method	ZSL results	GZSL results
CONSE	15.5	-
AMP	13.1	-
DEVISE	12.8	-
SS-Voc	16.8	-
SAE	27.2	12.0
LESAE	**27.6**	**12.4**

Table 2: Zero-shot learning (ZSL) results and Generalized Zero-shot learning (GZSL) results on ImageNet dataset. The results report Top-5 accuracy in %.

To evaluate the performance of methods on five small-scale datasets, the Top-1 accuracy (the prediction is accurate when the predicted class is the correct one) is measured. Moreover, for GZSL task, the samples from seen/unseen classes are classified into all classes. In GZSL setting, the search space at evaluation time is not restricted to only test classes, but includes also the training classes. In our experiments, we computer the harmonic mean of seen and unseen classification accuracies by flowing function [Xian *et al.*, 2017]:

$$H = \frac{2 \times acc_s \times acc_u}{acc_s + acc_u} \tag{11}$$

where acc_s and acc_u represent the Top-1 accuracy of images from seen, and images from unseen classes respectively.

For fair comparison with published results, we follow the settings with [Kodirov *et al.*, 2017] in ILSVRC2012/ILSVRC2010 dataset. In detail, we use GoogleNet features [Szegedy *et al.*, 2015] which is the 1024-dim activation of the final pooling layer. As an image usually contains multiple objects, we measure the Top-5 accuracy in this large-scale dataset.

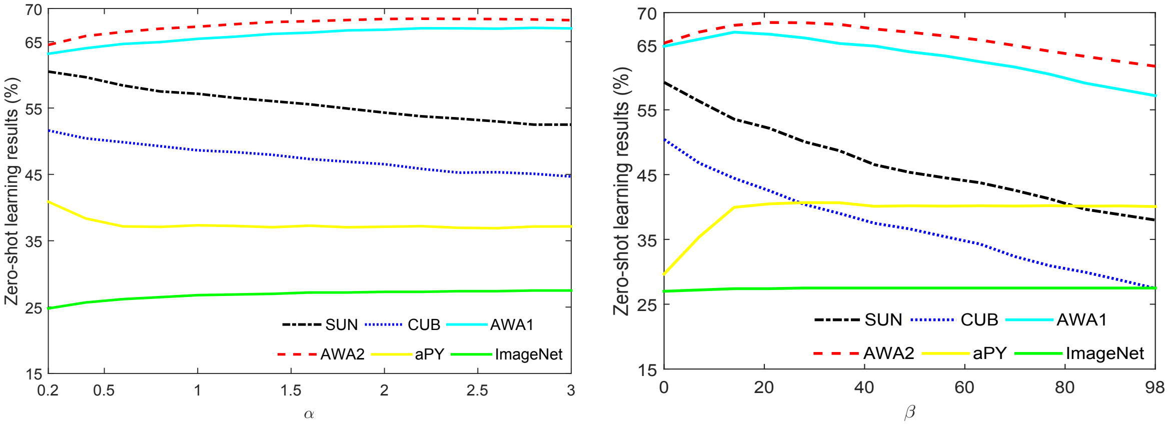
4.3 Parameter Settings

Our LESAE model has two free parameters: α and β (see Eq. (4)). Figure 1 shows that the value of α and β achieving the best performance in different datasets are concentrated in a small range. Specifically, from the parameter analysis on α (see Figure 1 (a)), our model can achieve better performance when the value of α approaches zero on SUN, CUB and aPY datasets, while performs better around $\alpha = 2.5$ on other three datasets. From the parameter analysis on β (see Figure 1 (b)), our model can achieve better performance when the value of α approaches zero on SUN and CUB datasets, while performs better around $\beta = 20$ on other four datasets. Empirically, α can be set to $0 < \alpha < 3$, while β varies from 0 to 40.

4.4 ZSL and GZSL Results

Table 1 presents the ZSL Top-1 accuracy on five small-scale datasets. Table 2 shows the ZSL and GZSL Top-5 accuracy

[2] The zero-shot learning benchmark can be found in the following link: http://www.mpi-inf.mpg.de/zsl-benchmark



(a) (b)

Figure 1: The accuracy of ZSL in six datasets influenced by super-parameter α (β), while β (α) fixed.

	SUN			CUB			AWA1			AWA2			aPY		
Method	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H	ts	tr	H
DAP	4.2	25.1	7.2	1.7	67.9	3.3	0.0	**88.7**	0.0	0.0	84.7	0.0	4.8	78.3	9.0
IAP	1.0	37.8	1.8	0.2	**72.8**	0.4	2.1	78.2	4.1	0.9	87.6	1.8	5.7	65.6	10.4
CONSE	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	**90.6**	1.0	0.0	**91.2**	0.0
CMT	8.1	21.8	11.8	7.2	49.8	12.6	0.9	87.6	1.8	0.5	90.0	1.0	1.4	85.2	2.8
CMT*	8.7	28.0	13.3	4.7	60.1	8.7	8.4	86.9	15.3	8.7	89.0	15.9	10.9	74.2	19.0
SSE	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	8.1	82.5	14.8	0.2	78.9	0.4
LATEM	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
ALE	21.8	33.1	26.3	23.7	62.8	**34.4**	16.8	76.1	27.5	14.0	81.8	23.9	4.6	73.7	8.7
DEVISE	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8	4.9	76.9	9.2
SJE	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	8.0	73.9	14.4	3.7	55.7	6.9
ESZSL	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
SYNC	7.9	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	10.0	90.5	18.0	7.4	66.3	13.3
SAE	17.8	32.0	22.8	18.8	58.5	28.5	14.2	81.2	24.1	16.7	82.5	27.8	9.9	74.7	17.5
LESAE	**21.9**	34.7	**26.9**	**24.3**	53.0	33.3	**19.1**	70.2	**30.0**	**21.8**	70.6	**33.3**	**12.7**	56.1	**20.1**

Table 3: Generalized Zero-Shot Learning (GZSL) results on SUN, CUB, AWA1, AWA2 and aPY using ResNet features. ts = acc_u , tr = acc_s, H = harmonic mean (CMT*: CMT with novelty detection). We measure Top-1 accuracy in %.

on ImageNet. Table 3 shows three evaluation indicators for the GZSL task on five small-scale datasets. Comparing the aforementioned experiments, we have several interesting observations:

(1) For zero-short learning, our model achieves the best results on all datasets except the CUB dataset although most of the compared methods apply complicated nonlinear models. Specifically, the accuracies on the AWA2 dataset increase 6% compared the strongest competitor. Our model and SAE have a similar result on SUN and CUB datasets. This is because our model achieve better performance when β approaches zero (see Figure 1 (b)), which means the low-rank constraint plays a small role on SUN and CUB datasets. On the large-scale dataset ImageNet, our model and SAE have a similar result, which increases 12.5% compared the existing best one, i.e. SS-Voc. However, all the methods perform poorly which indicates that there is a large room for improvement in this large-scale dataset.

(2) For generalized zero-short learning, our approach achieves the highest "ts" value in all small-scale datasets, which demonstrates that a low-rank projection would benefit the GZSL task. Moreover, it is easy to see that the "ts" value and the "H" value for those baselines with big "tr" value are generally very small. The main reason is that a very big "tr" value reflects the over-fitting training for the seen classes, i.e. the trained model in these methods cannot be generalized to new classes. On the large-scale dataset ImageNet, same as the ZSL result, our model increases 0.4% than SAE.

4.5 Complexity and Convergence Analysis

Figure 2 shows the algorithm 1 converges within only 5 steps. Moreover, the complexity of Eq. (5) depends on the size of feature dimension $O(m^3)$ instead of the number of samples n. These demonstrate our algorithm has a good practical application for its low complexity and good convergence.

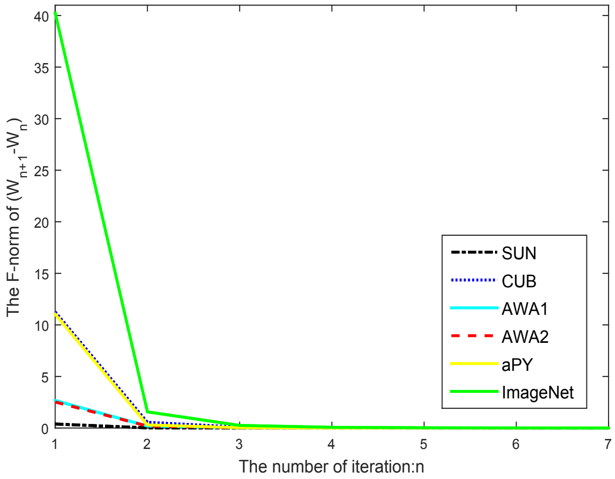


Figure 2: Convergence curve of LESAE on six datasets.

5 Conclusions

In this paper, a novel ZSL model named Low-rank Embedded Semantic AutoEncoder (LESAE) is proposed to learn a low-rank mapping to link visual features with their semantic representations. The encoder part tries to learn a low-rank projection from the feature space to a semantic space as in the existing ZSL models. The decoder part aims to the learned mapping can reconstruct the original visual features precisely. An optimization algorithm is also given to solve our model. Empirical results on five small-scale datasets and one large-scale dataset showed our method is significantly better than several well-established ZSL approaches.

Acknowledgements

This work is supported by National Natural Science Foundation of China under Grant 61773302, the 111 Project of China (B08038), and NSFC No. 91748208.

References

[Akata *et al.*, 2013] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826, 2013.

[Akata *et al.*, 2015] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, pages 2927–2936, 2015.

[Akata *et al.*, 2016] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2016.

[Cai *et al.*, 2010] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[Candès *et al.*, 2011] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[Changpinyo *et al.*, 2016] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, pages 5327–5336, 2016.

[Chen *et al.*, 2018] Yuanyuan Chen, Lei Zhang, and Zhang Yi. Subspace clustering using a low-rank constrained autoencoder. *Information Sciences*, 424:27–38, 2018.

[Ding *et al.*, 2017] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, pages 2050–2058, 2017.

[Fan *et al.*, 2017] Fan Fan, Yong Ma, Chang Li, Xiaoguang Mei, Jun Huang, and Jiayi Ma. Hyperspectral image denoising with superpixel segmentation and low-rank representation. *Information Sciences*, 397:48–68, 2017.

[Farhadi *et al.*, 2009] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, pages 1778–1785, 2009.

[Frome *et al.*, 2013] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, pages 2121–2129, 2013.

[Fu and Sigal, 2016] Yanwei Fu and Leonid Sigal. Semi-supervised vocabulary-informed learning. In *CVPR*, pages 5337–5346, 2016.

[Fu *et al.*, 2015a] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345, 2015.

[Fu *et al.*, 2015b] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *CVPR*, pages 2635–2644, 2015.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[Jolliffe, 1986] Ian T Jolliffe. Principal component analysis and factor analysis. In *Principal component analysis*, pages 115–128. Springer, 1986.

[Kodirov *et al.*, 2015] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, pages 2452–2460, 2015.

[Kodirov *et al.*, 2017] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, 2017.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

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

[Lancaster and Tismenetsky, 1985] Peter Lancaster and Miron Tismenetsky. *The theory of matrices: with applications*. Academic press, 1985.

[Li *et al.*, 2018] Bo Li, Risheng Liu, Junjie Cao, Jie Zhang, Yu-Kun Lai, and Xiuping Liu. Online low-rank representation learning for joint multi-subspace recovery and clustering. *IEEE Transactions on Image Processing*, 27(1):335–348, 2018.

[Liao *et al.*, 2018] Shuangli Liao, Jin Li, Yang Liu, Quanxue Gao, and Xinbo Gao. Robust formulation for pca: Avoiding mean calculation with l2p-norm maximization. In *AAAI*, pages 636–644, 2018.

[Liu *et al.*, 2013] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[Liu *et al.*, 2018] Yang Liu, Feiping Nie, and QuanXue Gao. Nuclear-norm based semi-supervised multiple labels learning. *Neurocomputing*, 275:940–947, 2018.

[Long and Shao, 2017] Yang Long and Ling Shao. Learning to recognise unseen classes by a few similes. In *ACM MM*, pages 636–644, 2017.

[Long *et al.*, 2017] Yang Long, Li Liu, Fumin Shen, Ling Shao, and Xuelong Li. Zero-shot learning using synthesised unseen visual data with diffusion regularisation. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[Nakarmi *et al.*, 2017] Ukash Nakarmi, Yanhua Wang, Jingyuan Lyu, Dong Liang, and Leslie Ying. A kernel-based low-rank (klr) model for low-dimensional manifold recovery in highly accelerated dynamic mri. *IEEE transactions on medical imaging*, 36(11):2297–2307, 2017.

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[Norouzi *et al.*, 2014] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zeroshot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, pages 1410–1418, 2009.

[Patterson *et al.*, 2014] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision*, 108(1-2):59–81, 2014.

[Peng *et al.*, 2012] Yigang Peng, Arvind Ganesh, John Wright, Wenli Xu, and Yi Ma. Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2233–2246, 2012.

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

[Russakovsky *et al.*, 2015] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, pages 935–943, 2013.

[Sylvester, 1884] James Joseph Sylvester. Sur l'équation en matrices px=xq. *CR Acad. Sci. Paris*, 99(2):67–71, 1884.

[Szegedy *et al.*, 2015] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *CVPR*, pages 1–9, 2015.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. In *California Institute of Technology*, 2011.

[Xian *et al.*, 2016] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *CVPR*, pages 69–77, 2016.

[Xian *et al.*, 2017] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*, 2017.

[Xie *et al.*, 2016] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis. In *ICML*, pages 478–487, 2016.

[Yang *et al.*, 2017] Yehui Yang, Wenrui Hu, Yuan Xie, Wensheng Zhang, and Tianzhu Zhang. Temporal restricted visual tracking via reverse-low-rank sparse learning. *IEEE transactions on cybernetics*, 47(2):485–498, 2017.

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

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

[Zhang *et al.*, 2015] Tianzhu Zhang, Si Liu, Narendra Ahuja, Ming-Hsuan Yang, and Bernard Ghanem. Robust visual tracking via consistent low-rank sparse learning. *International Journal of Computer Vision*, 111(2):171–190, 2015.