Semantics-Preserving Locality Embedding for Zero-Shot Learning

Shih-Yen Tao
b01901055@gmail.com
Language Technology Institute, Carnegie Mellon University, USA

Yao-Hung Hubert Tsai
yaohungt@andrew.cmu.com
Machine Learning Department, Carnegie Mellon University, USA

Yi-Ren Yeh
yryeh@nknu.edu.tw
Department of Mathematics, National Kaohsiung Normal University, Taiwan

Yu-Chiang Frank Wang
ycwang@citi.sinica.edu.tw
Research Center for IT Innovation, Academia Sinica, Taiwan

Abstract

Zero-shot learning (ZSL) aims at recognizing data as an unseen category, using information learned from the training data of predefined (seen) labels or attributes. In this paper, we propose an effective learning model for solving ZSL, which focuses on relating image and semantic domains with classification guarantees. In particular, we introduce semantics-preserving locality embedding when associating the above cross-domain data. We show that our ZSL model can be extended from inductive and transductive ZSL settings, if unlabeled data of unseen categories are presented during training. In the experiments, we show that our proposed method would perform favorably against baseline and state-of-the-art approaches on multiple benchmark datasets.

1 Introduction

Typically, in order to train image classifiers for object recognition, one would require a sufficient amount of training samples per category. When it comes to recognizing *novel* (or *unseen*) categories (i.e., no training data of that class available), such classifiers cannot be properly generalized. This is known as the problem of *zero-shot learning* (ZSL).

A common practice for ZSL is to utilize information from the *semantic domain*. That is, each class is represented by a vector representation in a semantic space, which is either based on *human-annotated* attributes (e.g., "long tail", "white fur", etc.) or in terms of *unsupervised* word embedding (e.g., word2vec [24]). During the training stage of ZSL, labeled images and their semantic vectors are jointly used to leverage the information across the corresponding visual and semantic spaces. In the inference stage, semantic vectors of unseen categories are observed, and thus the test images can be classified accordingly.

Previous methods such as Direct Attribute Prediction (DAP), Indirect Attribute Prediction (IAP) [15], and Semantic Output Code (SOC) [28] divide the ZSL problem into multiple independent attribute prediction tasks. To observe information across attributes, methods aiming at learning the mapping between visual and semantic spaces have also been proposed [9, 27, 35]. Nevertheless, since only the mapping from visual to semantic spaces

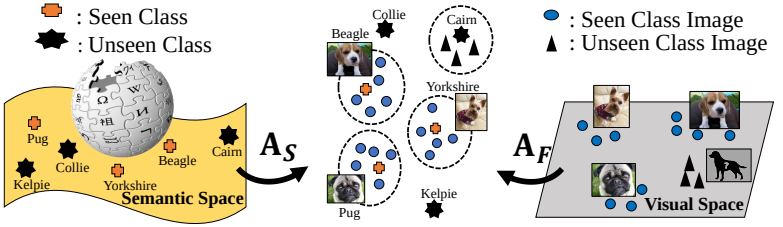

Figure 1: Illustration of semantics-preserving locality embedding for zero-shot learning. Note that $\mathbf{A}_S$ and $\mathbf{A}_F$ are the transforms for the semantic and feature spaces, respectively.

is performed, inter-class similarities measured in one domain might not be consistent with those observed in the other domain. Recent approaches like [2, 10, 32, 38, 41, 42] choose to share information across domains by learning the associated similarity metrics.

In this paper, we propose a novel learning framework for ZSL. We introduce the idea of semantics-preserving locality embedding, which performs *concept matching* between visual and semantic domains by preserving the locality of within-class data when learning our ZSL model (see Figure 1 for illustration). As a result, image and semantic data can be better associated for recognizing data of unseen categories. We note that, our proposed method can be extended from the standard *inductive* to *transductive* settings, in which unlabeled data of unseen classes can be observed during training [10, 12, 14, 31]. The main contributions of this paper are highlighted as follows:

- We propose semantics-preserving locality embedding for ZSL. Instead of relying on data observed from either image or semantic information domain, we perform subspace learning via matching cross-domain concepts for improved performance.
- Our proposed semantics-preserving locality embedding exploits the locality of within-class image data, with the semantics jointly embedded in the derived subspace.
- Our proposed method would result in a closed-form solution via eigen-decomposition, which is easy to implement and to solve.
- Depending on the availability of unlabeled data of unseen categories during learning, our method can be robustly performed in both inductive and transductive settings.

2 Related Work

Existing ZSL approaches typically consider semantic embedding of class labels for relating data of seen and unseen categories. The manually-defined attributes space was the first to be adopted [8, 15, 17, 22, 28, 37, 40]. However, such methods might not generalize to practical application, since one would expect a large number of classes for determining the manually-labeled semantic vectors for each class. As the result, recent works [2, 3, 9, 11, 27] chose to extract *unsupervised* semantic vectors mined from large text corpus on the Internet (e.g., Word2Vec [24], Glove [30]). Nevertheless, this strategy typically results in noisy attribute spaces. To alleviate this problem, works like [2, 3, 10] combined multiple semantic spaces and utilized additional strong supervision of visual data for improved performances.

We note that, ZSL methods based on learning feature transformation mainly seek for mapping data from visual to semantic spaces. Take DAP and IAP [15] for example, the former treated the learning of each attribute as an independent binary classification problem, while the latter utilized the classification outputs of seen classes for relating unseen visual and semantic data. [22, 37] further improved such models by considering the correlation between attributes. SOC [28] used multiple output linear regression to learn the mapping efficiently. Both Deep Visual-Semantic Embedding Model (DeViSE) [9] and [35] learned projection from image to word vectors using deep neural network (DNN). Convex Combination of Semantic Embeddings (ConSE) [27] chose to map images into the semantic space via a convex combination of the seen class label embedding vectors. [6] learned a metric function for relating visual and semantic data, and Multi-Task Embedding (MTE) [39] applied multi-task learning to observe the embedding of each attribute.

Recent ZSL approaches also seek for common representations across domains. For example, methods like [1, 2, 32] learned a bilinear function between the two spaces, and Latent Embedding Model (LatEm) [38] derived the piece-wise linear ones instead. Several methods based on deriving common feature spaces also exist. For example, [10, 21] learned a latent space using Canonical Correlation Analysis (CCA), and [41, 42] proposed to transform the observed features via semantic similarity embedding. These methods become preferable when domain discrepancy between visual and semantic domains is expected.

Another category of ZSL would be the *classifier-based* methods. In short, these methods create a new classifier for unseen classes by combining or adapting existing classifiers for seen classes. [7] design their classifiers by utilizing a domain transfer function and a probabilistic regressor. Co-Occurrence Statistics (COSTA) [23] builds a classifier for the novel class as a weighted combination of classifiers derived from seen classes, using co-occurrence statistics obtained from the Internet. More recently, [6] chooses to synthesize the classifier for unseen classes from the existing classifiers using similarity observed between the semantic vectors from the seen classes.

For transductive ZSL, Propagated Semantic Transfer (PST) [31] performs *label propagation* by exploiting the manifold structure for novel classes. Transductive Multi-View Zero-Shot Learning (TMV) [10] and [14] utilize CCA and regularized sparse coding for relating cross-domain data. Shared Model Space (SMS) [12] extended [32] for the transductive setting and reported promising performance.

3 Our Approach

3.1 Problem Settings and Notations

Let $\mathcal{D} = \{\mathbf{X}, Y\} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$ denote training data in the visual domain, where the ith column $\mathbf{x}_i$ of $\mathbf{X} \in \mathbb{R}^{d_F \times N}$ represents the ith instance, and y_i is its corresponding label from $\mathcal{L} = \{1, 2, .., C\}$. For ZSL, we have $\mathcal{D}^U = \{\mathbf{X}^U, Y^U\} = \{\mathbf{x}_i^U, y_i^U\}_{i=1}^{N^U}$ as test data, where $\mathbf{x}_i^U \in \mathbb{R}^{d_F}$ denotes the ith test instance and y_i^U is the associated label from the *unseen* label set $\mathcal{L}^U = \{1^U, 2^U, ..., C^U\}$. Note that the label sets $\mathcal{L}$ and $\mathcal{L}^U$ from training and testing data are *disjoint* (i.e., $\mathcal{L} \cap \mathcal{L}^U = \varnothing$). Note that each class is associated with a semantic vector in a d_S dimensional space. Thus, we have $\mathcal{S} = \{\mathbf{s}_i \in \mathbb{R}^{d_S}\}_{i=1}^{C}$ and $\mathcal{S}^U = \{\mathbf{s}_i^U \in \mathbb{R}^{d_S}\}_{i=1}^{C^U}$ as semantic vectors of training data and test data, respectively. Our goal is to predict unseen labels $\{y_i^U\}_{i=1}^{N_U}$ of test data by leveraging visual and semantic-domain data.

3.2 Our Proposed ZSL Framework

Inspired by [32, 35], we cast ZSL as a *subspace learning* problem. We aim at finding distinct feature transformations $\mathbf{A}_F \in \mathbb{R}^{d_F \times d_k}$ and $\mathbf{A}_S \in \mathbb{R}^{d_S \times d_k}$ for visual feature and semantic spaces, respectively. Once the transformations are learned, data from both domains can be projected into a d_k dimensional latent space for ZSL classification.

In particular, we perform *concept matching* between semantic and visual domains, in which semantic vectors would be aligned with visual exemplars of each class with the proposed *semantics-preserving locality embedding*. We note that such an embedding would enforce within-class data locality to be preserved in the derived manifold. We now detail our proposed formulations below.

To obtain $\mathbf{A}_F$ and $\mathbf{A}_S$, we aim to solve the optimization problem below:

$$\begin{aligned} \min_{\mathbf{A}_S, \mathbf{A}_F} \; & E_C(\mathbf{A}_S, \mathbf{A}_F) + \lambda_1 E_S(\mathbf{A}_F) + \lambda_2 \Omega(\mathbf{A}_F, \mathbf{A}_S) \\ \text{s.t.} \; & \mathbf{Z}\mathbf{H}\mathbf{Z}^\top = \mathbf{I}. \end{aligned} \tag{1}$$

where $\mathbf{Z} = [\mathbf{A}_F^\top \mathbf{X}, \mathbf{A}_S^\top \mathbf{S}] \in \mathbb{R}^{d_k \times (N+C)}$ indicates the concentrated projected data matrix, $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_C] \in \mathbb{R}^{d_s \times C}$, and $\mathbf{I}$ is the identity matrix.

In the above formulation, E_C measures the difference between visual and semantic concepts, while E_S performs semantics-preserving locality embedding across domains. The regularization term $\Omega(\mathbf{A}_F, \mathbf{A}_S) = \|\mathbf{A}_F\|_F^2 + \|\mathbf{A}_S\|_F^2$ is penalized by λ_2 to prevent overfitting. As suggested in [18], we enforce the constraint of $\mathbf{Z}\mathbf{H}\mathbf{Z}^\top = \mathbf{I}$, where the center matrix $\mathbf{H} = \mathbf{I}_{N+C} - \frac{1}{N+C}\mathbf{1}_{N+C}$, and $\mathbf{1}_{N+C}$ is the matrix with all elements equal to one). This constraint is to prevent trivial solutions and to maximize the variance of the projected data. Further details can be found in [18].

3.3 Bridging Image and Semantic Domains

As noted in [9, 27, 35], nearest-neighbor classification in the semantic space would produce satisfactory performance, if the projected visual data can be properly identified. In other words, semantic vectors typically contain the concept of a particular class, and thus exhibit discriminative abilities. Based on the above observation, we aim to match the concepts between semantic and visual domains in the derived latent space. With the idea that each semantic vector represents the concept of a particular class, we take the class means to represent the concepts of each seen class in the visual domain. Thus, given data points $\mathbf{x}_i^j$ in class j, we have $\mathbf{A}_S^\top \mathbf{s}_j \approx \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{A}_F^\top \mathbf{x}_i^j$ in the latent space. As a result, to match the concepts between visual and semantic domains, the first term of (1) can be expressed as follows:

$$E_C(\mathbf{A}_S, \mathbf{A}_F) = \sum_{j=1}^{C} \|\mathbf{A}_S^\top \mathbf{s}_j - \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{A}_F^\top \mathbf{x}_i^j\|^2. \tag{2}$$

3.4 Semantics-Preserving Locality Embedding

Locality-embedded projection has been utilized in computer vision tasks [19, 20, 26, 33]. From a geometric perspective, locality preserving projection aims to maintains the neighborhoods of data samples across transformations, and thus the locality information can be transferred to the resulting subspace. Typically, locality-preserving projections can be learned via

imposing a *manifold regularizer* [4]. To be more specific, one can introduce a regularizer term $\frac{1}{2}\sum_{i,j}\|\mathbf{A}^\top\mathbf{x}_i-\mathbf{A}^\top\mathbf{x}_j\|^2\mathbf{W}_{i,j}$ into the formulation of subspace learning, where $\mathbf{W}$ is the affinity matrix that encodes the local structure of the data observed in the original space. Based on such regularization, transformation $\mathbf{A}$ with locality guarantees can be obtained.

Different from existing locality-embedding techniques, we propose to enforce semantics-preserving locality regularization and integrate the associated term into the framework of subspace learning. This is to derive a ZSL subspace in which semantic and feature information would be jointly embedded (instead of relying on information from either domain which would limit the ZSL performances).

In our work, a semantics-preserving locality embedding term is proposed into our learning framework of (1). This is achieved by constructing a simple *homogeneous* graph which only connects samples from the *same* class, i.e. $\mathbf{W}_{ij}=1$ if $\mathbf{x}_i,\mathbf{x}_j$ are in the same class, and $\mathbf{W}_{ij}=0$ otherwise. As the result, the second term of (1) is determined as follows:

$$E_S(\mathbf{A}_F)=\frac{1}{2}\sum_{j=1}^{C}\{\frac{1}{N_j^2}\sum_{i=1}^{N_j}\sum_{k=1}^{N_j}\|\mathbf{A}_F^\top\mathbf{x}_i^j-\mathbf{A}_F^\top\mathbf{x}_k^j\|^2\}, \tag{3}$$

where $\frac{1}{N_j^2}$ is for normalization.

We now explain why the introduction of semantics-preserving locality embedding is preferable for relating data across visual and semantic domains, and thus improved ZSL performance can be expected. It can be seen that, by minimizing (3), the constructed homogeneous affinity graph in the resulting subspace only enforces projected image data of the same class to be close to each other. This implies that the local structure of this graph which corresponds to a particular semantic label would be more compact after solving the above optimization problem. As a result, improved separation between projected images of different semantic labels can be expected. It is worth emphasizing that, the above idea is very different from previous methods that relate cross-domain data by directly projecting visual data to the semantic domain [9, 16, 35].

3.5 Zero-Shot Classification

Once the learning of (1) is complete, the above transformations can be obtained. In other words, one can project test image data and semantic vectors of unseen classes onto the resulting subspace. Thus, prediction of unseen labels for test data can be performed accordingly.

To be more precise, with the learned transformations $\mathbf{A}_S$ and $\mathbf{A}_F$ for semantic and visual feature spaces, $\mathbf{X}^U$ and $\mathcal{S}^U$ are projected onto the latent space accordingly. Then, for each test image, it can be easily classified using the cosine similarity between all semantic vectors:

$$\arg\max_j \frac{\langle\mathbf{A}_F^\top\mathbf{x}^U,\mathbf{A}_S^\top\mathbf{s}_j^U\rangle}{\|\mathbf{A}_F^\top\mathbf{x}^U\|\|\mathbf{A}_S^\top\mathbf{s}_j^U\|}. \tag{4}$$

The complete ZSL process of our proposed method is summarized in Algorithm 1. Please see the Supplementary for detailed derivations.

4 From Inductive to Transductive ZSL

We now extend our ZSL formulation to a transductive version. In other words, we now deal with the semi-supervised setting in which the semantic vectors $\mathbf{s}_i^U$ and test instance $\mathbf{x}_i^U$

Algorithm 1 Our ZSL with Semantics-Preserving Locality Embedding

Input: $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, $\{\mathbf{x}_i^U\}_{i=1}^{N^U}$, $\mathcal{S}$, $\mathcal{S}^U$, and latent space dimension d_k
1: Solve $\mathbf{A}_S, \mathbf{A}_F$ in (1) using $\mathcal{D}$ and $\mathcal{S}$
2: Project $\mathcal{D}^U$ and $\mathcal{S}^U$ on the latent space
3: Classify $\mathcal{D}^U$ using the projected $\mathcal{S}^U$
Output: $\mathbf{A}_S, \mathbf{A}_F$, and $\{y_i^U\}_{i=1}^{N^U}$

Algorithm 2 From Inductive to Transductive ZSL

Input: $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, $\{\mathbf{x}_i^U\}_{i=1}^{N^U}$, $\mathcal{S}$, $\mathcal{S}^U$, latent space dimension d_k
1: Initialize pseudo labels of $\mathcal{D}^U$ using the inductive version
2: **while** not converge **do**
3: Solve (5) for updating $\mathbf{A}_S, \mathbf{A}_F$
4: Update the pseudo labels of $\mathcal{D}^U$
5: **end while**
Output: $\mathbf{A}_S, \mathbf{A}_F$, and $\{y_i^U\}_{i=1}^{N^U}$

are both available and presented during the training stage. While recent works such as [10, 12, 14] particularly proposed transductive ZSL for achieving promising performances, our formulation can be easily adapted from inductive to transductive settings as follows:

$$\begin{aligned} \min_{\mathbf{A}_S, \mathbf{A}_F} \; & E_C(\mathbf{A}_S, \mathbf{A}_F) + E_C^U(\mathbf{A}_S, \mathbf{A}_F) + \lambda_1 [E_S(\mathbf{A}_F) + E_S^U(\mathbf{A}_F)] + \lambda_2 \Omega(\mathbf{A}_F, \mathbf{A}_S) \\ \text{s.t.} \; & \hat{\mathbf{Z}}\hat{\mathbf{H}}\hat{\mathbf{Z}}^\top = \mathbf{I}, \end{aligned} \tag{5}$$

where $\hat{\mathbf{Z}} = [\mathbf{A}_F^\top \mathbf{X}, \mathbf{A}_F^\top \mathbf{X}^\mathbf{U}, \mathbf{A}_S^\top \mathbf{S}, \mathbf{A}_S^\top \mathbf{S}^U] \in \mathbb{R}^{d_k \times (N+N^U+C^U+C)}$ indicates the projected data matrix for all visual instance and semantic vectors, and $\hat{\mathbf{H}} = \mathbf{I}_{\hat{N}+\hat{C}} - \frac{1}{\hat{N}+\hat{C}}\mathbf{1}_{\hat{N}+\hat{C}}$, where $\hat{N} = N + N^U$ and $\hat{C} = C + C^U$ denotes the total number of visual sample and semantic vector, respectively.

Similar to the inductive version in (2), the term $E_C^U(\mathbf{A}_S, \mathbf{A}_F)$ is defined as follows:

$$E_C^U(\mathbf{A}_S, \mathbf{A}_F) = \sum_{j=1}^{C^U} \|\mathbf{A}_S^\top \mathbf{s}_j^U - \frac{1}{N_j^U} \sum_{i=1}^{N_j^U} \mathbf{A}_F^\top \mathbf{x}_i^{U,j}\|^2, \tag{6}$$

where $\mathbf{x}_i^{U,j}$ is the ith instance of the unseen class j, and N_j^U is the total number of instances of unseen class j. Note that, since no label information is available for test instances, the labels of $\mathbf{x}_i^{U,j}$ can only be *estimated* during the subspace learning process. Thus, we adopt the *self-taught* strategy. That is, we view the predicted labels as pseudo labels of the visual data of unseen classes. As for $E_S^U(\mathbf{A}_F)$, we simply apply (3) and replace $\mathbf{x}$ by $\mathbf{x}^U$.

To learn our transductive ZSL model, we start with learning the inductive version to predict the pseudo labels for $\mathbf{x}_i^U$. Then, we update transformations $\mathbf{A}_F$ and $\mathbf{A}_S$ by (5), and predict the pseudo labels y_i^U in an iterative fashion until convergence (or the maximum iteration number is reached). Algorithm 2 summarizes the learning of the transductive version of our ZSL model.

Table 1: Descriptions of the datasets.

	CUB	DOG	AWA	SUN
# of seen classes	150	85	40	645/646
# of unseen classes	50	28	10	72/71
# of images	11786	19499	30473	14340
Dim of Attribute	312	-	85	102
Dim of Word2Vec	400	400	400	-
Dim of Glove	400	200	400	-
Dim of Wordnet	-	163	-	-

Table 2: Performance comparisons of inductive ZSL. Ours* denotes our method without the proposed embedding term, while Ours† (with embedding) directly applies $\mathbf{s}_i$ instead of $g(\mathbf{s}_i)$ for representing semantic data. Note that, MTE [6] only considers 10 unseen classes in **SUN** so its improved result is expected.

Methods	CUB			AWA			DOG			SUN
	Attribute	Word2Vec	Glove	Attribute	Word2Vec	Glove	Word2Vec	Glove	Wordnet	Attribute
SOC [23]	34.7	30.9	30.6	58.6	50.8	68.0	24.6	17.8	17.3	50.4
DeViSE [8]	42.3	28.5	24.2	77.0	48.6	45.7	22.2	18.9	27.8	61.1
ConSE [22]	33.6	28.8	30.8	59.0	53.2	49.8	18.1	17.0	19.6	49.6
SSE [42]	31.8	27.9	25.4	63.8	58.6	65.8	25.3	17.8	28.9	51.2
SJE [3]	50.1	28.4	24.2	66.7	52.1	58.8	19.6	17.8	24.3	63.3
ESZSL [29]	50.3	33.4	34.1	76.8	62.2	67.7	27.5	17.1	27.2	59.2
JLSE [43]	33.7	28.0	27.1	71.8	64.0	68.0	26.5	18.8	29.3	49.5
LatEm [33]	45.5	31.8	32.5	71.9	61.1	62.9	22.6	20.9	25.2	63.7
Sync [6]	48.7	31.2	32.8	72.9	62.0	67.0	28.0	20.4	**30.7**	62.8
MTE [6]	43.3	-	-	77.3	-	-	-	-	-	84.1
Ours*	52.3	29.0	31.9	71.9	62.9	**70.5**	26.7	18.3	22.6	64.0
Ours†	54.9	32.7	33.2	70.9	61.3	68.8	29.7	25.3	27.1	**69.3**
Ours	**56.7**	**35.2**	**36.9**	**78.4**	**66.5**	68.6	**29.9**	**26.1**	27.2	66.2

5 Experiments

5.1 Datasets and Settings

We consider four benchmark datasets for evaluation: CUB-200-2011 Birds (**CUB**) [36], Stanford Dogs (**DOG**) [13], Animal with Attributes (**AWA**) [16], and SUN Attribute (**SUN**) [29]. We followed the seen/unseen class data split of [38] for **AWA**, **CUB**, and **DOG**. As for **SUN**, we use the 10 pre-specified data splits as suggested in [6]. All the visual features are extracted by CNN via GoogLeNet (i.e., the top-layer pooling units). For the semantic domain, we consider four different types of semantic features: Attribute [16], Word2Vec [24], Glove [30] and Wordnet Vector [25]. The first is annotated manually for each class, while Word2Vec and Glove are viewed as unsupervised ones extracted from Wikipedia corpus using deep neural networks. Finally, WordNet is a hierarchy-based word representation (see [6, 38] for detailed information). Table 1 summarizes the datasets considered in our experiments.

To process the data, we normalize each visual feature vector by z-score normalization. Each semantic vector $\mathbf{s}_i$ is normalized to unit ℓ_2 norm, and is transformed to a kernel representation $g(\mathbf{s}_i) = [d(\mathbf{s}_i,\mathbf{s}_1), d(\mathbf{s}_i,\mathbf{s}_2), \ldots, d(\mathbf{s}_i,\mathbf{s}_{\hat{C}})]$ where $d(\mathbf{s}_i,\mathbf{s}_j) = \exp(-\gamma\|\mathbf{s}_i - \mathbf{s}_j\|^2)$ with $\gamma = 1$. Later we will verify our use of kernelized semantic features $g(\mathbf{s}_i)$. $\hat{C}$ denotes the number of classes available during training. For all experiments, we select the parameters λ_1 and λ_2 from $\{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ via cross-validation. For simplicity, we fix the dimension d_k of the derived latent space as the number of classes available during training.

Table 3: Performance comparisons of transductive ZSL.

Methods	CUB			AWA			DOG			SUN
	Attribute	Word2Vec	Glove	Attribute	Word2Vec	Glove	Word2Vec	Glove	Wordnet	Attribute
UDA [23]	39.5	-	-	73.2	-	-	-	-	-	-
TMV [11]	51.2	32.5	38.9	89.0	69.0	88.7	24.4	15.2	25.6	61.4
SMS_{ESZSL} [3]	52.3	34.7	32.3	89.6	78.0	82.9	29.7	17.9	29.8	60.5
SMS_{Ours} [3]	59.2	36.0	36.1	**91.1**	88.2	81.3	**37.2**	27.3	31.3	60.5
Ours	**67.3**	**41.5**	**39.5**	90.4	**94.5**	**95.6**	33.5	**30.4**	**39.7**	**70.1**

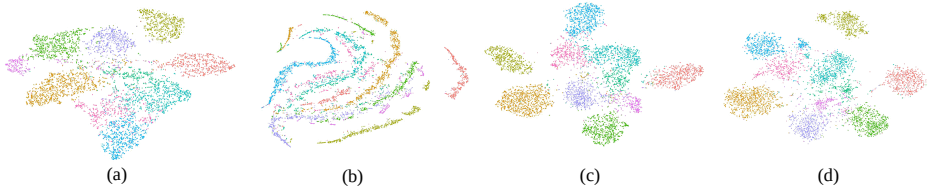

Figure 2: t-SNE visualization on **AWA** with Attribute semantic on (a) SOC [28], (b) SSE [41], (c) JLSE [42], and (d) ours. Comparing (c) and (d), although both JLSE and our method result in well-separated clusters for each class, each cluster in (d) is more compact and thus improved performance can be expected (see Table 2).

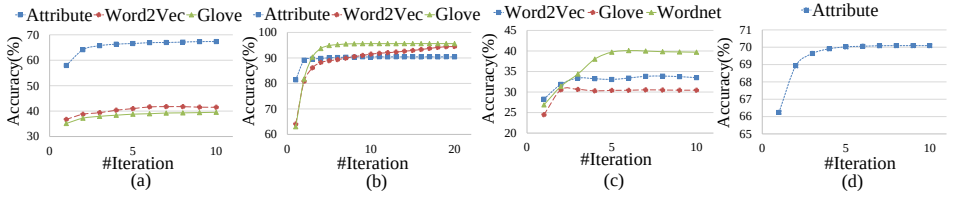

Figure 3: Convergence analyses on (a) **CUB**, (b) **AWA**, (c) **DOG**, and (d) **SUN**.

5.2 Inductive ZSL

To compare our approach with baseline and recent ZSL methods, we consider SOC [28], DeViSE [9], ConSE [27], SJE [2], ESZSL [32], LatEm [38], SSE [41], JLSE [42], Sync [6] and MTE [8]. We note that, VGG features are applied in [8] for describing the images. We do not take DAP and IAP [15] as the baselines, since they can only be applied to handle *binary* attributes (recent ZSL methods generally perform in continuous semantic spaces).

Table 2 now lists and compares the performances of different methods. In addition to the comparisons with recent ZSL methods, we further conduct controlled experiments for our approach. That is, we have Ours* in Table 2 denote our method without the proposed semantics-preserving locality embedding term, while Ours† only applies $\mathbf{s}_i$ instead of $g(\mathbf{s}_i)$ for representing semantic data (as noted in Section 5.1). In addition, we apply t-SNE to produce embedding visualization in Figure 2. Comparing to the visualization outputs of [28, 41, 42], it can be seen that our method not only resulted in improved separation between different classes, the local structure of the projected data for each class is also more compact compared to others.

From the above quantitative and qualitative results, we see that our method performed favorably against state-of-the-art ZSL approaches, while the use of the proposed embedding term with kernelized semantic representation can be successfully verified.

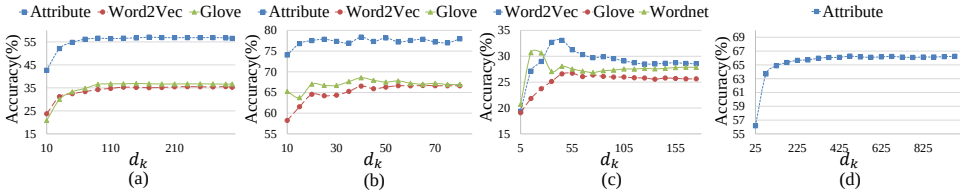

Figure 4: Performance sensitivity w.r.t. the subspace dimension d_k on (a)**CUB**, (b)**AWA**, (c)**DOG**, and (d)**SUN**.

5.3 Transductive ZSL

To compare with recent transductive ZSL methods, we have: UDA [14], TMV [10] and approach of SMS [12]. Note that UDA utilizes CNN OverFeat features [34] as image representations, and SOC is applied to initialize the transformation for TMV. As for the initialization for SMS, we apply two different strategies for comparisons: ESZSL and the inductive version of our ZSL model (denoted as SMS_{ESZSL} and SMS_{Ours}, respectively).

The results of different transductive ZSL methods are listed in Table 3. Compared to Table 2, it is clear that transductive ZSL generally achieved improved accuracy. This suggests that the exploration of information extracted from data of unseen class wound be preferable for ZSL problems. Nevertheless, from the results shown in this table, our method still achieved comparable or improved performances when comparing to state-of-the-art transductive ZSL methods. It is worth repeating that, these recent ZSL approaches are particularly designed for solving transductive learning problems, while ours can be generalized to both inductive and transductive modes.

5.4 Convergence and Parameter Sensitivity

While our inductive ZSL produces closed-form solutions for transformations $\mathbf{A}$, the transductive version of our method applies an iterative scheme for deriving the solutions. To verify the convergence during the learning of our ZSL model, we show the results on different datasets using semantic vectors in Figure 3. From this figure, it is clear that our method generally converged within 10 iterations and achieved satisfactory performance.

For the parameter sensitivity analysis, we first discuss the dimension d_k of the derived subspace. In Figure 4, we show the results with varying d_k. Based on the results shown in this figure, we can confirm that our choice of $d_k = C$ would be reasonable (i.e., the results were generally not sensitive to this choice except for **DOG**).

Finally, as noted in Section 5.1, we now assess the role of γ for the Gaussian kernelized semantic vectors $g(\mathbf{s}_i)$. In our experiments, we fix $\gamma = 1$ which results in satisfactory performance across all datasets. We show an example in Figure 5. In this figure, we compare the performance of varying γ on **CUB**, and we successfully verify that our choice of $\gamma = 1$ for representing Gaussian kernelized semantic vectors would be preferable.

6 Conclusion

We proposed a zero-shot learning approach via semantics-preserving locality embedding,

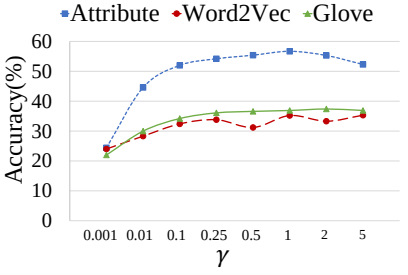

Figure 5: Performance on **CUB** with varying γ.

which aims at deriving a subspace for relating visual and semantic space via concept matching. We show that, by preserving the locality of within-class data in the derived subspace, improved separation between semantic data can also be achieved. In our experiments, we showed that our model can be applied to both inductive and transductive settings, and performed favorably against state-of-the-art approaches.

Acknowledgment

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST105-2221-E-002-236-MY2, MOST105-2218-E-001-006, and MOST105-2221-E-017-009-MY3 .

References

[1] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013.

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

[3] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *CVPR*, 2016.

[4] Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006.

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

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

[7] Mohamed Elhoseiny, Babak Saleh, and Ahmed Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *ICCV*, 2013.

[8] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009.

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

[10] Yanwei Fu, Timothy M Hospedales, Tao Xiang, and Shaogang Gong. Transductive multi-view zero-shot learning. *TPAMI*, 2015.

[11] Zhenyong Fu et al. Zero-shot object recognition by semantic manifold distance. In *CVPR*, 2015.

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

[13] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *CVPR*, 2011.

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

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

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

[17] Jingen Liu, Benjamin Kuipers, and Silvio Savarese. Recognizing human actions by attributes. In *CVPR*, 2011.

[18] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S Yu. Transfer feature learning with joint distribution adaptation. In *ICCV*, 2013.

[19] Mingsheng Long, Jianmin Wang, Guiguang Ding, Dou Shen, and Qiang Yang. Transfer learning with graph co-regularization. *TKDE*, 2014.

[20] Jiwen Lu, Gang Wang, Weihong Deng, Pierre Moulin, and Jie Zhou. Multi-manifold deep metric learning for image set classification. In *CVPR*, 2015.

[21] Yao Lu. Unsupervised learning on neural network outputs: with application in zero-shot learning. In *IJCAI*, 2016.

[22] Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011.

[23] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[24] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *ICLR*, 2013.

[25] George A Miller. Wordnet: a lexical database for english. *ACM*, 1995.

[26] X Niyogi. Locality preserving projections. In *NIPS*, 2004.

[27] Mohammad Norouzi and Mikolov et al. Zero-shot learning by convex combination of semantic embeddings. In *ICLR*, 2014.

[28] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[29] Genevieve Patterson, Chen Xu, Hang Su, and James Hays. The sun attribute database: Beyond categories for deeper scene understanding. *IJCV*, 2014.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[31] Marcus Rohrbach, Sandra Ebert, and Bernt Schiele. Transfer learning in a transductive setting. In *NIPS*, 2013.

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

[33] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.

[34] Pierre Sermanet and Eigen at al. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

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

[36] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[37] Yang Wang and Greg Mori. A discriminative latent model of object classes and attributes. In *ECCV*, 2010.

[38] Yongqin Xian and Akata et al. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[39] Timothy Hospedales Xun Xu and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation, 2016.

[40] Felix X Yu, Liangliang Cao, Rogerio S Feris, John R Smith, and Shih-Fu Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013.

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

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