Structure Aligning Discriminative Latent Embedding for Zero-Shot Learning

Omkar Gune
guneomkar@ee.iitb.ac.in
Indian Institute of Technology Bombay, Mumbai, India

Biplab Banerjee
getbiplab@gmail.com
Indian Institute of Technology Bombay, Mumbai, India

Subhasis Chaudhuri
sc@ee.iitb.ac.in
Indian Institute of Technology Bombay, Mumbai, India

Abstract

We address the problem of zero-shot visual recognition in this paper and particularly focus on learning a discriminative latent embedding space where the visual image descriptors and the respective semantic class representations can be projected with coinciding alignment. While a supervised dimension reduction strategy which simultaneously optimizes the intra-class compactness and between-class separation is used to learn the latent space for the visual features, the semantic class prototypes are further projected onto this latent space via a multi-stage non-linear mapping function for re-alignment purposing. Furthermore, it is ensured that the visual and semantic class prototypes are likely to overlap in the latent space such that the overall similarity between samples from both the domains is maximized. Apart from remarkably reducing the so-called semantic gap, the discriminative property of the learned latent layer representations entails improved classification performance on both the standard zero-shot learning (ZSL) and the challenging generalized ZSL (GZSL) setups on three benchmark datasets (AWA, CUB, SUN) where the proposed method surpasses the state of the art results.

1 Introduction

The recent success of deep learning methods in the field of visual recognition can be attributed to availability of the large-scale labeled datasets such as ImageNet [33]. However from a different point of view, precisely annotating such colossal source of data has indeed posed a challenge since manual annotations can be time consuming and erroneous at the same time. It is therefore necessary that the improved performance for visual recognition be obtained while reducing the effort in annotating data.

Zero-shot learning (ZSL) [4, 7, 8, 18, 24, 32, 48] techniques provide an elegant solution in extending the capabilities of a classifier in recognizing *novel* (unseen during training) classes by deploying mid-level semantic representations at the class level. These semantic representations (also referred as class prototypes, class embeddings or attribute vectors) such as human-annotated attributes, word vector representations [22], and textual descriptions allow transfer of knowledge from *seen* (training) to *unseen* (test) classes. While ZSL training is devoted to learn some measure of compatibility between the labeled images and class embeddings by leveraging the complementarity of both the visual and semantic spaces, the test stage is based on evaluating the function for mapping the visual features of the unseen classes to the nearest class prototypes.

Early ZSL techniques are based on modeling a direct mapping from visual to semantic space using regressors. However, such methods lead to the loss of semantic structure of visual data after being projected onto the semantic space given the unbounded nature of the compatibility functions. Additionally, techniques based on direct mapping to semantic space suffers from hubness issue [6, 20, 35]. On the other hand, few ZSL methods such as [9, 14] learn the latent space where both visual and semantic features can be projected. Although the motivation is to learn a *better* space to improve the ZSL performance, such methods can suffer due to the fundamental difference in the inter-class structure in the semantic and visual space. As a single class prototype is available as opposed to different visual representations for the same class samples, intra-class variation poses a greater challenge in ZSL which is not addressed explicitly in most of the endeavors. The intra-class variation in visual samples come from occlusion, different lighting conditions, different viewing angles in which case some of the object attributes may not be readily attainable. Besides, simply finding a latent space based on standard least-square fitting may not help in case of ZSL for fine-grained visual categories. It is imperative that such visually similar categories remain discriminative in the learned latent space. A triplet-loss based metric learning for discriminating the visual categories in latent space has been studied in [29] but requires careful selection of triplets. To the best of our knowledge, the existing ZSL techniques do not take care of all these issues concerning the learning of discriminative visual and semantic embeddings in an unified framework. We, on the other hand, retrospect the necessity of discriminativeness in common latent space learning based ZSL and outline the main contributions below:

- We ensure discrimination of the visual features in the latent space in terms of a discriminative class-encoder model. This simultaneously minimizes the within-class variance by reconstructing one sample from another both sharing similar class labels and maximizes the between-class separation in terms of a softmax type classifier.

- The class prototypes are projected onto this latent space via an additional intermediate latent space, thus ensuring the learning of a prior abstract class embedding which is deemed to be more discriminative than the original class prototypes.

- Inter-class structures of the visual and semantic space are aligned in the latent space by minimizing the divergence between the class wise visual centroids (or visual prototypes) and semantic class prototypes when projected onto the latent space. It is also ensured that the overall similarity between the visual features and the class prototypes is maximized in the latent space. A trade-off between both these measures leads to better correspondence between the visual and semantic domains.

- Experimental results showcase that the proposed formulation performs better or comparable with the state-of-the-art for the challenging AWA [19], CUB [40], and SUN [28] datasets for standard ZSL considering both the attributes and word vector based semantic space representations. Likewise, in case of GZSL where training and test classes are not disjoint, our method delivers improved performance on AWA1 [19], CUB, and SUN, while showing consistent trends with the literature for AWA2 [42].

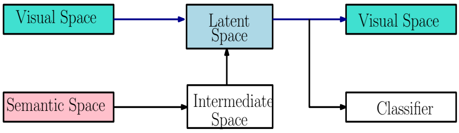

Figure 1: The architecture of a proposed model. During training, visual features are projected onto the common latent space and are reconstructed back at the output using a class-encoder. The seen class prototypes are projected onto common latent space via intermediate space. The classifier is fed with the latent representations to make them discriminative. During testing, ZSL is carried out in the latent space using nearest neighbor search after projecting test sample and unseen class prototypes.

2 Related Work

Semantic Space: Existing ZSL methods can be categorized based on the use of semantic space. [7, 27, 48] use attributes while word vector representations such as [22] have been used in [1, 37, 48]. Image sentence descriptions have been explored in [31, 46]. Crucially, attributes suffer from the drawback of manual annotations but are more effective than the word vector representations which are modeled in an unsupervised way in terms of neuronal probabilistic language models trained on large text corpus. In this regard, few study the attribute correlations problem [12, 13, 27]. Although, the use of word vectors looks more prominent practically, characterization of objects by the corresponding word vectors may be ambiguous thus resulting in lower performance. In contrary, [1, 2, 9, 39] appreciate the benefits of both attributes and word vectors for ZSL.

Embedding method: [8, 37] use direct visual to semantic projection while [1, 9, 32] learn the relation between visual and semantic space using intermediate latent space. Deep models [8, 30, 31, 37, 43, 46] are also used in ZSL. Few tackle the problem of ZSL by representing unseen class in term of seen classes [5, 25, 47, 49] while others check the compatibility of visual and latent space [3, 26, 32, 37]. The transductive ZSL in [9, 11, 16] use unlabeled unseen class samples along with seen samples and always outperform the standard or inductive ZSL models. More comprehensive survey on ZSL can be found in [10].

Simply mapping from visual/semantic to latent space may result in the different inter-class structure in the latent space than the inter-class structure in visual/semantic domain. To overcome this, neighborhood relation between different classes is learned in visual and semantic domain. The latent structure preserving methods have been used in [9, 14, 21, 36, 48]. Specifically, [48] uses similarity between seen classes while [9] uses Canonical Correlation Analysis for visual and multiple semantic domains in the transductive setting. Dictionary learning approach is used to learn discriminative latent attributes in [14] while [45] uses discriminative sparse non-negative matrix factorization to learn discriminative semantic representations.

The proposed technique differs from previous models in the following aspects. First, the class-encoder [34] in our framework makes the visual features to have lower intra-class variance. In principle, the decoder part of class-encoder is key to learn an abstract visual concepts at the class scale in the latent layer which can be better related to the abstract class prototypes. Second, the classifier in the latent space further makes latent features discriminative. Third, aligning projections of visual features with their corresponding class prototypes, such as in [8, 46], alone would not be sufficient as it may lose inter-class structure present in the original visual and semantic space. Identically, aligning semantic prototypes and visual prototypes (class means of visual features) as in [36] would have limited use in case of multi-modal distribution of classes. The trade-off between these two aspects is necessary to meaningfully align the embeddings of visual and semantic prototypes in the latent space. The explicit consideration of all these issues in our model demonstrates improved performance for ZSL as well as GZSL.

3 Proposed Model

3.1 Background on ZSL and Notations

Consider a visual dataset with s seen classes and u unseen classes. Let $D_S = \{x_i, y_i\}_{i=1}^{n_s} \in \mathcal{X} \times \mathcal{Y}$ denote the training data with $x_i \in \mathbb{R}^d$ as the visual feature and y_i as the label from $\mathcal{Y}_S = \{1, 2, \cdots, s\}$. On the other hand, let $D_U = \{x_j, y_j\}_{j=1}^{n_u}$ be the test data with $x_j \in \mathbb{R}^d$ being the j^{th} test sample and y_j is the corresponding label from $\mathcal{Y}_U = \{s+1, s+2, \cdots, s+u\}$ such that $\mathcal{Y}_S \cap \mathcal{Y}_U = \phi$. The semantic class prototype for the i^{th} class is $z_i \in \mathbb{R}^k$ which can be an attribute vector or a word vector representation. Therefore, there are $s+u$ class prototypes available such that $\mathcal{Z}_S = \{z_1, z_2, \cdots, z_s\}$ and $\mathcal{Z}_U = \{z_{s+1}, z_{s+2}, \cdots, z_{s+u}\}$. Finally, let $\theta(x)$ and $\phi(z)$ be the embedding functions which separately project the visual and semantic descriptors to the common latent space. Under the aforementioned setup, the standard ZSL problem aims at learning a compatibility function $\mathcal{F}(\theta(x), \phi(z))$ given D_S and $\mathcal{Z}_S$ such that $\mathcal{F}(\cdot)$ outputs a high value if a given x and z share an identical class label or otherwise produces low values. Given a test sample x_t coming from a random unseen class, its label $\widehat{y_t}$ is estimated as follows:

$$\widehat{y_t} = \arg\max_{y \in \mathcal{Y}_U} \mathcal{F}(\theta(x_t), \phi(z_y)) \tag{1}$$

In case of GZSL task, samples with labels from $\mathcal{Y}_S \bigcup \mathcal{Y}_U$ are present during testing. Nonetheless similar to standard ZSL, $\mathcal{F}(\cdot)$ is trained here solely based on D_S and $\mathcal{Z}_S$.

3.2 Discriminative and Structure Aligning Embedding

Figure 1 outlines the proposed ZSL model which contains one sub-network each for the visual and semantic space. We aim to learn $\theta(\cdot)$ and $\phi(\cdot)$ in order to project the visual and semantic data onto the common latent space where $\mathcal{F}(\cdot)$ can be applied to assess the compatibility between the two representations. Essentially, the proposed network depicts an encoder-decoder model with a class-encoder [34] for the visual features and a two-layer encoder for the semantic prototypes.

We initiate our discussion with the standard auto-encoder (AE) model and subsequently define our loss function. An AE in its simplest form is a three layer neural network which aims to reconstruct the input at the output using an encoder-decoder framework. Consider an input to AE be $X = [x_1 x_2 \cdots x_{n_s}] \in \mathbb{R}^{d \times n_s}$ which contains n_s samples from a d-dimension feature space. The encoder part of AE projects X onto l-dimension latent space ($l \ll d$) using $W \in \mathbb{R}^{l \times d}$ while decoder of AE aims to reconstruct the input as $\hat{X} \in \mathbb{R}^{d \times n_s}$ from these projected samples using $W^T \in \mathbb{R}^{d \times l}$. The encoding and decoding functions, $\theta(\cdot)$ and $\psi(\cdot)$, respectively, can be realized along with non-linearity $s_f(\cdot)$ and $s_g(\cdot)$ as follows:

$$h_i = \theta(x_i) = s_f(Wx_i) \qquad \text{and} \qquad \hat{x}_i = \psi(h_i) = g_f(W^T h_i). \tag{2}$$

The following loss is minimized while learning the AE parameters W:

$$\mathcal{L}_{AE} = \sum_{i=1}^{n_s} \|\hat{x}_i - x_i\|_F^2 \tag{3}$$

AE allows to learn useful latent space for compact representation of original features in an unsupervised way. However, more useful representation in terms of lower intra-class variance can be obtained using a class-encoder [34]. A class-encoder is a variant of AE which reconstructs the output from the different samples belonging to the same class. Let C_x be the class label of two randomly picked samples x and $\tilde{x}$. Then the loss function to be minimized for class-encoder is:

$$\mathcal{L}_{CE} = \sum_{x \in X} \sum_{\tilde{x} \in C_x} \|\hat{x} - \tilde{x}\|_F^2. \tag{4}$$

Ideally, different images of the same class may vary significantly in the visual feature space owing to aforementioned transformations. However, in semantic space a single unique representation is available for each class. Therefore in order to draw the correspondence between two different domain representations of the same class, intra-class variance must be lowered. Apparently, the visual - latent - visual path of Figure 1 acts like a typical AE model. As stated before, instead of standard AE we use a class-encoder to reduce the intra-class variance of the visual features in the latent space.

From a different point of view, learning $\mathcal{F}(\cdot)$ in the latent space is a regression problem which may lose the discriminative property of features in the original visual and semantic spaces. The problem is even severe while dealing with fine-grained visual categories having overlapping feature descriptors. To overcome this difficulty, we simultaneously posit the use of a classifier which is explicitly trained on latent visual features in order to enforce discriminativeness. Note that the learning of a classifier requires minimization of a typical cross-entropy loss $\mathcal{L}_{CLFR}$ for s seen classes. The gradients due to $\mathcal{L}_{CLFR}$ affects the learning of an embedding function during back propagation making the latent representations discriminative. The cumulative loss for the visual - latent - (visual, classifier) branch now is put forward as follows:

$$\mathcal{L}_{VIS} = \mathcal{L}_{CE} + \mathcal{L}_{CLFR}. \tag{5}$$

Further, the class prototypes are projected onto the latent space through an intermediate layer. Let the respective non-linear embeddings be $\phi_1(\cdot)$ and $\phi_2(\cdot)$, with $\phi(\cdot) = \phi_2(\phi_1(\cdot))$. Ideally, given the original $\mathcal{Z}$, $\phi(z)$ learns abstract semantic class representations which can better be associated with the latent visual concepts than the original class prototypes. Since we seek to minimize the pairwise divergence between the two embeddings in the latent space, the respective loss ($\mathcal{L}_{LE}$) is defined as:

$$\mathcal{L}_{LE} = \sum_{i=1}^{n_s} \|h_i - \phi(z_{y_i})\|_F^2. \tag{6}$$

Visual space and semantic space have inherently different inter-class structures. In other words, relation between visual features of different classes are likely to be different than the relation between the respective semantic class prototypes. As class-encoder is used for visual feature reconstruction, a lower intra-class variance signifies that the latent visual representations are centered around the mean vectors of the classes separately. Given that, we

Parameters	**AWA1**	**AWA2**	**CUB**	**SUN**
# of seen classes/# of unseen classes	40/10	40/10	150/50	645/72
# of instances	30475	37322	11786	14340
Attribute dimension	85	85	312	102

Table 1: Descriptions of different datasets of ZSL.

seek to match the structures of visual and semantic features in the latent space by aligning these mean vectors with the projection of semantic prototypes of the corresponding classes. Since the latent space is discriminative as well, thanks to $\mathcal{L}_{CLFR}$, such an alignment further ensures that the semantic prototypes do not get projected within close vicinity, a problem frequently encountered in ZSL. Let μ^{y_i} denote the mean of the latent features h_i of class y_i from D_S, we define the structure alignment loss ($\mathcal{L}_{SA}$) as:

$$\mathcal{L}_{SA} = \sum_{i=1}^{s} \|\mu^{y_i} - \phi(z_{y_i})\|_F^2. \tag{7}$$

The overall loss to be minimized along with the standard ℓ_2 regularization $\mathcal{R}$ on the model parameters to avoid a trivial solution is expressed as

$$\mathcal{L} = \alpha_1 \mathcal{L}_{CE} + \alpha_2 \mathcal{L}_{LE} + \alpha_3 \mathcal{L}_{SA} + \alpha_4 \mathcal{L}_{CLFR} + \beta \mathcal{R}. \tag{8}$$

where $\alpha_1, \alpha_2, \alpha_3, \alpha_4$, and β are scalars to appropriately weight the different losses.
Training and inference: $\mathcal{L}$ is a non-convex function given that $\phi(\cdot)$ and $\theta(\cdot)$ are non-linear mappings. Following the same, $\mathcal{L}$ is minimized based on the standard mini-batch gradient descent optimization strategy. For $\mathcal{L}_{VIS}$, input-output pairs are selected randomly from each of the classes in $\mathcal{Y}_S$ in each iteration of the training. We do not observe any convergence related issue during training. During testing, the visual samples and class prototypes of the unseen classes are separately projected onto the latent space using $\theta(\cdot)$ and $\phi(\cdot)$, respectively. $\mathcal{F}(\cdot)$ is evaluated to assign the class labels to the visual features using Eqn.(1).

4 Experiments

4.1 Datasets

We consider three standard datasets for the evaluation of the proposed inductive ZSL model: Animals with Attributes (AWA) [19], Caltech Birds 200-2011 (CUB) [40] and SUN Attributes (SUN) [28]. The details of these datasets are given in Table 1. We use $d = 1024$-dimension GoogleNet [38] features for all the datasets as visual embeddings while experimenting separately with the manually annotated attributes and distributed word vectors based representations. We use 1000-dimension word vectors for AWA provided by [46] while 500-dimension word vectors for CUB provided by [39] is deployed (both obtained using the word2vec model trained on the *wikipedia* corpus). For GZSL, we experiment on AWA1 [19], AWA2 [42], CUB, and SUN datasets, as suggested in [42] and utilize $d = 2048$-dimension ResNet features [46].

4.2 Model Architecture

The proposed model shown in Figure 1 is simple in architecture yet achieves the state of the art performance in most of the experimental setups. Specifically, we implement all the

Method	Visual Feature	AWA		CUB		SUN
		Attribute	word2vec	Attribute	word2vec	Attribute
ConSE [24]	F_G	59.0	53.2	33.6	28.8	49.6
SSE [48]	F_V	76.3	-	30.4	-	-
ESZSL [32]	F_G	76.3	-	47.2	-	59.2
SPLE [36]	F_G	78.4	66.5	56.7	35.2^*	69.3
SYNC [5]	F_G	72.9	-	54.5	-	62.8
RKT [39]	F_G	71.6	59.1	33.5	23.2	-
ALE [41]	F_G	71.9	61.1	45.5	$31.8^‡$	63.7
LAD [14]	F_V	82.5	-	56.6	-	-
SAE [17]	F_G	84.7	-	61.4	-	65.2
DZSL [46]	F_G	86.7	78.8	58.3	-	-
DSR [45]	F_V	**87.2**	-	57.1	-	-
Ours-AE	F_G	83.7	79.8	61.2	28.7	**69.6**
Ours-CE	F_G	85.0	**80.7**	**62.2**	**31.0**	68.1

Table 2: ZSL classification accuracy (%) comparison on different datasets. Ours-AE: Proposed model with standard AE, Ours-CE: proposed model with class-encoder. F_V: VGG features, F_G: GoogleNet features. Kernelized prototypes are used in $*$. For CUB we use 500D CBoW word2vec from [39] while ‡ uses 400D word2vec.

functional mappings explained in section 3 using fully-connected (*fc*) neural network layers. For the visual - latent - visual branch (Figure 1), d-dimension *fc* input and output layers are considered while the latent layer dimension is set to 1000. For the semantic-intermediate-latent path, we have k-dimension *fc* input layer, one *fc* intermediate layer of 750-dimensions followed by the *fc* latent layer of 1000-dimensions where k represents the dimension of the semantic representations. In addition, we use ReLu [23] non-linearity at the latent and intermediate layers. The classifier is modeled as s-dimension *fc* softmax layer whose inputs are the latent visual representations. Note that we use the same model specifications for all the datasets while separately tuning the relative weights of the individual loss terms ($\alpha_1 - \alpha_4, \beta$). Furthermore the weights of *fc* layers are randomly initialized during training. We use a learning rate of 0.0001 and train the model using Adam [15] with batch of 64. We also consider the following two different scenarios while fusing both the attribute and word vector based class prototypes: (1) We concatenate both the representations in order to form the new class embeddings; (2) We non-linearly ($\tanh(\cdot)$) project both the representations separately onto the common space, which is subsequently projected onto the visual latent space after performing a weighted combination of both the intermediate representations [46].

4.3 Performance on ZSL and GZSL

Evaluation protocols: We compare the classification accuracy on unseen classes of our model with others on the standard ZSL in Table 2. For GZSL, we use the same separately on seen classes (S), unseen classes (U), and also report the harmonic mean (H) as defined in [42]. Table 3 shows the performance comparison [1] with the recent literature for GZSL. *Only discriminative models are used while comparing the ZSL and GZSL performance.*

ZSL performance: It is clear that the ZSL performance is better using attributes than the word vectors. The primary reason for the same is that carefully and manually designed at-

[1] https://github.com/lzrobots/DeepEmbeddingModel_ZSL

Method	AWA1			AWA2			CUB			SUN		
	S	U	H	S	U	H	S	U	H	S	U	H
DeViSE[8]	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9
SYNC [5]	**87.3**	8.9	16.2	**90.5**	10.0	18.0	**70.9**	11.5	19.8	**43.3**	7.9	13.4
SJE [1]	74.6	11.3	19.6	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
ALE [41]	76.1	16.8	27.5	81.8	14.0	23.9	62.8	23.7	34.4	33.1	**21.8**	26.3
SAE [17]	77.1	1.8	3.5	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8
DZSL [46]	84.7	32.8	47.3	86.4	**30.5**	**45.1**	57.9	19.6	29.2	34.3	20.5	25.6
PSR [44]	-	-	-	73.8	20.7	32.3	54.3	24.6	33.9	37.2	20.8	26.7
Ours-CE	85.5	**34.7**	**49.4**	87.1	30.1	44.7	60.7	**30.2**	**40.3**	41.1	21.2	**27.9**

Table 3: Performance comparison of GZSL on different datasets. Ours-CE: proposed model with class-encoder. S: Accuracy on seen classes, U: Accuracy on unseen classes, H: Harmonic mean (Settings followed from [42].)

tributes provide holistic representations for the classes as compared to word vectors which are learned in an unsupervised way. Despite this fact, for AWA our model outperforms all the recent methods and achieves a relative improvement of 2.41% using word vectors over the best performing model [46] so far. In addition, our shallow model attains an impressive result of 85.0% with attributes for AWA, a score marginally inferior to the deep model of [46] which takes the advantage of end-to-end training of the visual features, but superior to all the ad-hoc ZSL techniques. CUB is a fine grained dataset where our best result of 62.2% beats all the recent ZSL models. This excellent performance can be partly associated with the classifier in the latent space which makes the classes discriminative. We achieve the best performance using class-encoder instead of standard AE structure in the visual-latent-visual path of our model. As fewer training samples are available for CUB, class-encoder allows to capture better intra-class variance (by reconstructing each sample from every other sample of the same class) as compared to the standard AE. For SUN, we report the state of the art performance of 69.6% using attributes. We also experiment by fusing the attribute and word vectors together and report the results in Table 5. Simple concatenation of attribute and word vectors gives the best accuracy of 63.1% on CUB while the other more principled fusion strategy (section 4.2) achieves superior performance on AWA with 87.4%.

GZSL performance: For GZSL we experiment using attributes. Our model achieves the best performance on AWA1 while reports a slightly lower harmonic mean on AWA2 with respect to [46]. Differently, we report a relative improvement of 17.1% in term of harmonic mean for CUB which is the best performance for CUB on GZSL to date. We again outperform recent methods on SUN with harmonic mean of 27.9. It is to be noted that we perform relatively better for unseen classes for all the datasets as compared to the others. This trend is indeed encouraging as GZSL is tenacious than standard ZSL due to a bias towards seen classes while identifying the unseen data.

4.4 Ablation study

Table 4 shows the results of more controlled experimental cases by purposefully overlooking the effects of individual loss measures in Eqn.(8) on the ZSL performance. We are interested in assessing the effects of: (1) the visual class-encoder ($\mathcal{L}_{CE}$), (2) the softmax classifier on the visual latent features ($\mathcal{L}_{CLFR}$), and (3) the structure alignment in terms of the difference between latent visual and semantic prototypes ($\mathcal{L}_{SA}$). Clearly, the classifier plays a significant role in improving the overall performance, specifically in case of fine-grained datasets CUB

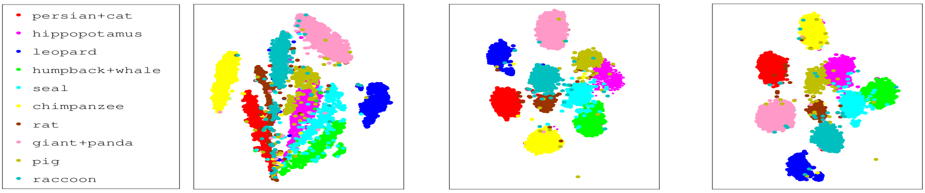

Figure 2: t-SNE plots of embeddings of unseen class visual samples onto latent space for AWA. Latent space is learned using following loss measures: (Left) $\mathcal{L}_{\mathcal{CE}}$,$\mathcal{L}_{\mathcal{LE}}$,$\mathcal{L}_{SA}$; (Middle)$\mathcal{L}_{\mathcal{AE}}, \mathcal{L}_{\mathcal{LE}}, \mathcal{L}_{\mathcal{CLFR}}, \mathcal{L}_{SA}$; (Right)$\mathcal{L}_{\mathcal{CE}}, \mathcal{L}_{\mathcal{LE}}, \mathcal{L}_{\mathcal{CLFR}}, \mathcal{L}_{SA}$. Best viewed in color.

Loss measures	AWA	CUB	SUN
$\mathcal{L}_{\mathcal{CE}}, \mathcal{L}_{\mathcal{LE}}$	72.5	20.8	38.5
$\mathcal{L}_{\mathcal{CE}}, \mathcal{L}_{\mathcal{LE}}, \mathcal{L}_{\mathcal{SA}}$	78.1	37.2	34.1
$\mathcal{L}_{\mathcal{CE}}, \mathcal{L}_{\mathcal{LE}}, \mathcal{L}_{\mathcal{CLFR}}$	80.3	55.2	66.7
$\mathcal{L}_{\mathcal{CE}}, \mathcal{L}_{\mathcal{LE}}, \mathcal{L}_{\mathcal{CLFR}}, \mathcal{L}_{SA}$	85.0	62.2	69.6

Table 4: Ablation study for standard ZSL using different loss measures using attributes.

Method	AWA	CUB
AMP [11]	66.0	-
SJE [1]	73.9	51.7
DZSL [46]	**88.1**	59.0†
Ours-CE(CT)	83.7	**63.1**
Ours-CE(FS)	87.4	62.7

Table 5: ZSL performance accuracy (%) using both attributes and word2vec. CT: Feature concatenation, FS: Feature fusion. Sentence description is used in †.

and large dataset SUN (in terms of number of classes). Precisely, the classifier is distinctly responsible for enhancing the ZSL accuracy from 72.5% to 80.3%, from 20.8% to 55.2%, and from 38.5% to 66.7% for AWA, CUB, and SUN respectively. The dissimilar inter-class structures of visual and semantic domain are aligned in the latent space by minimizing $\mathcal{L}_{SA}$ which further boosts the ZSL accuracy by 4.7%, 7.0%, and 2.9% for AWA, CUB, and SUN, respectively. Additionally from t-SNE plots in Figure 2, it is clearly perceptible that the inclusion of the classifier in the latent space benefits the ZSL performance. Moreover, a sensitivity analysis on the ZSL performance with respect to the dimension of the latent layer is shown in Figure 3. It can be observed that the ZSL performance is relatively unaffected by latent feature dimension.

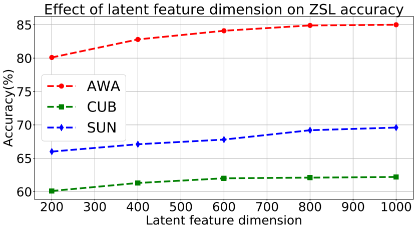

Figure 3: Effect of latent feature dimension on the ZSL accuracy using attributes on different datasets. Best viewed in color.

5 Conclusions

We propose a novel discriminative and inter-class structure preserving latent space learning based ZSL model in this paper. The prime goal of our work is to make the visual latent representations discriminative while aligning the visual and semantic prototypes in the latent space simultaneously. Using extensive experiments, we thoroughly evaluate the efficacy of our approach both on the standard and generalized ZSL settings for three challenging datasets where overall superior classification performance with respect to the literature can be observed. We are currently interested in extending our model for the transductive ZSL where unlabeled test visual features can be used along with the seen data during training.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 2927–2936. IEEE, 2015.

[2] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 59–68, 2016.

[3] 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.

[4] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, pages 730–746. Springer, 2016.

[5] Soravit Changpinyo, Wei-Lun Chao, Boqing Gong, and Fei Sha. Synthesized classifiers for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5327–5336, 2016.

[6] Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*, 2014.

[7] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

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

[9] Yanwei Fu, Timothy M Hospedales, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *European Conference on Computer Vision*, pages 584–599. Springer, 2014.

[10] Yanwei Fu, Tao Xiang, Yu-Gang Jiang, Xiangyang Xue, Leonid Sigal, and Shaogang Gong. Recent advances in zero-shot recognition: Toward data-efficient understanding of visual content. *IEEE Signal Processing Magazine*, 35(1):112–125, 2018.

[11] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2644, 2015.

[12] Dinesh Jayaraman and Kristen Grauman. Zero-shot recognition with unreliable attributes. In *Advances in neural information processing systems*, pages 3464–3472, 2014.

[13] Dinesh Jayaraman, Fei Sha, and Kristen Grauman. Decorrelating semantic visual attributes by resisting the urge to share. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1629–1636, 2014.

[14] Huajie Jiang, Ruiping Wang, Shiguang Shan, Yi Yang, and Xilin Chen. Learning discriminative latent attributes for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4223–4232, 2017.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2452–2460, 2015.

[17] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. *arXiv preprint arXiv:1704.08345*, 2017.

[18] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[19] 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.

[20] Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 270–280, 2015.

[21] Yang Long, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis. 2017.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[23] Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[24] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[25] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[26] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418, 2009.

[27] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[28] 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.

[29] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. Visually aligned word embeddings for improving zero-shot learning. *arXiv preprint arXiv:1707.05427*, 2017.

[30] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[31] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–58, 2016.

[32] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, pages 2152–2161, 2015.

[33] 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.

[34] Hailin Shi, Xiangyu Zhu, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning discriminative features with class encoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 46–52, 2016.

[35] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 135–151. Springer, 2015.

[36] Yi-Ren Yeh Shih-Yen Tao, Yao-Hung Hubert Tsai and Yu-Chiang Frank Wang. Semantics-preserving locality embedding for zero-shot learning. In *BMVC*, 2017.

[37] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943, 2013.

[38] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, et al. Going deeper with convolutions. Cvpr, 2015.

[39] Donghui Wang, Yanan Li, Yuetan Lin, and Yueting Zhuang. Relational knowledge transfer for zero-shot learning. In *AAAI*, volume 2, page 7, 2016.

[40] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

[41] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[42] 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.

[43] Yongxin Yang and Timothy M Hospedales. A unified perspective on multi-domain and multi-task learning. *arXiv preprint arXiv:1412.7489*, 2014.

[44] Soma Biswas Yashas Annadani. Preserving semantic relations for zero-shot learning. In *CVPR*, 2018.

[45] Meng Ye and Yuhong Guo. Zero-shot classification with discriminative semantic representation learning.

[46] Li Zhang, Tao Xiang, Shaogang Gong, et al. Learning a deep embedding model for zero-shot learning. 2017.

[47] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4166–4174, 2015.

[48] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via semantic similarity embedding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4166–4174, 2015.

[49] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6034–6042, 2016.