Joint Learning of Attended Zero-Shot Features and Visual-Semantic Mapping

Yanan Li
ynli@zju.edu.cn

Institute of Artificial Intelligence
Zhejiang Lab
Hangzhou, China

Donghui Wang
dhwang@zju.edu.cn

Institute of Artificial Intelligence
Zhejiang University
Hangzhou, China

Abstract

Zero-shot learning (ZSL) aims to recognize unseen categories by associating image features with semantic embeddings of class labels and its performance can be improved progressively through learning better features and more generalized visual-semantic mapping (V-S mapping) to unseen classes. Current methods typically learn feature extractors and V-S mapping independently. In this work, we propose a simple but effective joint learning framework with fused autoencoder (AE) paradigm, which can simultaneously learn features specific to ZSL task as well as V-S mapping inseparable to learning features. In particular, the encoder in AE can not only transfer semantic knowledge to the feature space, but also achieve semantics-guided attended feature learning. At the same time, the decoder in AE can be used as a V-S mapping, which further improves the generalization ability to unseen classes. Extensive experiments show that the proposed approach can achieve promising results.

1 Introduction

Driven by massive labeled dataset and deep learning networks, the boundaries of standard classification tasks have been pushed further. However, it is of great difficulty to collect and annotate all the concepts that beyond daily objects with high quality, thus failing to further extend conventional classifiers. Attempting to model people's ability to identify a new object by just reading a description of it, zero-shot learning (ZSL) can recognize previously unseen classes through **semantic embeddings** of class labels, typically in the form of attributes and word vectors, and is gaining increasing popularity in the community [1, 4, 12, 13, 17, 18, 30].

ZSL relies mainly on aligning the semantic relationship between image features and the corresponding semantic embeddings of class labels. The current existing ZSL approaches can be summarized in a two-stage framework: (i) *extracting image features*: aims to improve image representations by advanced CNN networks with a focus on the top layer [21] or object regions [6, 15]; (ii) *learning visual-semantic mapping* (**V-S mapping**): aims to improve the generalization ability to unseen classes by borrowing ideas from auto-encoders [12, 14, 23], manifold alignment [4], distribution consistency [23], etc. However, we believe that there are two possible shortcomings in this typical framework.

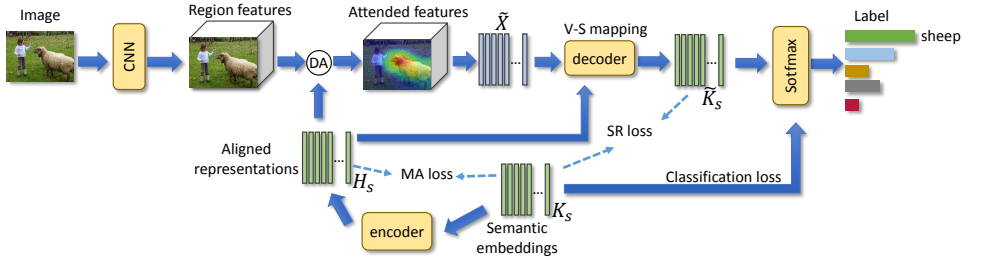

Figure 1: Illustration of the proposed method. We first encode the class embeddings K_s into the aligned visual representations H_s by enforcing a manifold alignment (MA) criterion. It guide in obtaining semantically relevant features $\tilde{X}$ for all classes simultaneously by using the dense attention (DA) mechanism. We then predict their semantic embeddings of all classes $\tilde{K}_s$ through the coupled V-S mapping, constrained to reconstruct all semantics (SR) during training. The input image is finally categorized as the nearest class with largest similarity score in the semantic embedding space.

First, image features acquired by compressing the entire picture into a fixed expression by a pre-trained network in advance, rather than being specially learned for each task in ZSL, may not be informative enough. Since they are originally trained for classification tasks, these wholistic features have the potential to underrate some local discriminative information, which is more useful for distinguishing different classes and is also relevant to its class embeddings at the same time. When there is a great difference between the pre-trained dataset (e.g. ImageNet) and testing dataset (e.g. CUB [32]), the unaligned semantics between image features and semantic embeddings can be a major factor affecting ZSL performances [2]. Therefore, it necessitates a powerful mechanism to steer image features to important semantics for the task at hand.

Second, image feature extraction and V-S mapping learning are treated independently and generally carried out in isolation, leading to the potential difficulty in further improving ZSL performance [15]. It is widely recognized that V-S mapping builds a bridge of knowledge transfer among mutually exclusive classes, whose transferability is closely related to the degree of semantic alignment between image features and semantic embeddings. In this light, these two line of research should be viewed as complementary technologies to be optimized together.

In this paper, we propose a simple but effective end-to-end approach based on autoencoder (AE) paradigm to achieve the purpose of semantic alignment in ZSL, shown in Fig.1. It can learn task-specific image features and more generalized V-S mapping on unseen classes in a joint and inseparable way. To be specific, we encode the class embeddings into aligned latent representations by enforcing a manifold alignment criterion and use these representations to guide the exclusive image features for each particular class by employing an attention mechanism. The important local information can be highlighted while other irrelevant information can be deemphasized. At the same time, we establish a coupling between V-S mapping and the decoder, which is designed to reconstruct the semantic embeddings of all classes during training. With this explicit reconstruction loss, V-S mapping is able to generalize better to unseen classes. In this way, joint interaction between these two stages can boost unseen prediction gradually. Finally, we extensively evaluate our model using four popular benchmarks in two different settings, where promising results can be obtained.

2 Related Work

Current ZSL methods mainly focus on two different lines of research to improve their performances: enhancing the generalization ability of V-S mapping and refining the semantics in either image representations or semantic embeddings.

V-S Mapping. When it comes to the cross-modal V-S mapping, we generally choose the conventional regression models [1, 13] or deep neural network regression models [8, 19]. Also, the reverse mapping from semantic embeddings to image feature could be chosen, so as to alleviate the hubness problem that generally suffered by nearest neighbor prediction in ZSL [12] or generate virtual samples for unseen classes [7, 14, 26, 28]. Modeling it as an indirect mapping is also a very popular practice, where both image features and the semantic embeddings are projected into a common latent space by using the one-stage bilinear models [1, 2, 22] or two-stage projections [29]. More recently, there arises many new ideas borrowed from such fields as manifold alignment [4, 28], manifold distribution consistency with variational AE [23]. In all these above works, the image features are either crafted manually or extracted from a pre-trained CNN model in advance. Learning V-S mapping is separated from learning image features.

Feature Enhancement. Another way to improve ZSL performance is to reduce the semantic gap between visual and textual domain by improving image representations or the semantic embeddings, with a boost to visual-semantic mapping as a byproduct. Since deep features are well structured in the high dimensional space, semantic embeddings are usually aligned with visual information [16, 21]. Instead of the global image features, some work suggest that using the relevant local regions (or region features) with class labels to represent an image can make a good alignment. For example, [15] proposed a cascaded zooming mechanism to automatically identify the most discriminative region with object as a focus in an image. [6] proposed to learn region-specific classifiers to connect text terms to their relevant regions and suppress connections to non-visual text terms without any part-text annotations. We also borrow ideas from this line of research in that we use an attention mechanism to automatically select relevant regions for each image.

Attention-based Models. Since introduced by [3], attention mechanism has recently become an integral part of compelling sequence modeling and transduction models in various tasks, ranging from tasks using single modality (e.g. machine translation [25]) to tasks using two or more modalities (e.g. image question answering [34]). It emulates one of the most curious facets of human visual system, i.e. the presence of attention, allowing salient image features to dynamically come to forefront as needed, which is especially important for image recognition when there is a lot of clutter in the background. In other words, the attention mechanism provides a way of information fusion and can selectively establish the cross-modal connection. However, the attention mechanism has not been well explored in ZSL, which is a typical cross-modal learning problem. We noticed a similar work in [11] which uses a kind of self attention to put larger weights on more relevant regions obtained by a specific part detection network for ZSL. In contrast, we adopt a dense attention mechanism to consider the interaction between any region implicitly obtained by the general CNN and any class for an input image. In the testing stage, we can directly apply the attention mechanism to each test data and obtain its class label by simple comparison, without having to undergo a second transformation of class embeddings.

3 Proposed Method

Existing ZSL methods generally adopt the visual features independent of the task at hand, which remains constant during the whole training process. *The main insight of our proposed method is that visual features should be specially learned for ZSL task, and the learning process should be guided by class embeddings to make them more semantically consistent with the knowledge domain*, so as to make the cross-modal V-S mapping more generalized over unseen classes. The key to our approach is the joint optimization of learning visual features and modeling V-S mapping. For the purpose, we introduce a solution consisting of coupling between these two factors and a novel dense attention mechanism. Fig.1 illustrates the overall architecture.

We begin by defining the problem of our interest. Let $\mathcal{L}_s = \{l_s^1, ..., l_s^{c_s}\}$ denotes a set of c_s seen class labels and $\mathcal{L}_u = \{l_u^1, ..., l_u^{c_u}\}$ a set of c_u unseen class labels with $\mathcal{L}_s \cap \mathcal{L}_u = \emptyset$. $\mathcal{K}_s = \{k_s^1, ..., k_s^{c_s}\}$ and $\mathcal{K}_u = \{k_u^1, ..., k_u^{c_u}\}$ are their corresponding semantic embeddings, respectively. Suppose we have a labeled training dataset $\mathcal{D}_s = \{(\mathcal{I}_i, y_i, k_i)\}_{i=1}^{n_s}$ of n_s samples for seen classes, where $\mathcal{I}_i$ is the i-th image, $y_i \in \mathcal{L}_s$ and $k_i = k_s^{y_i} \in \mathcal{K}_s$. Given the training dataset $\mathcal{D}_s$ and $\mathcal{K}_u$, the goal of ZSL is to learn a classifier $f_c : \mathcal{I} \rightarrow \mathcal{L}_u$ in the conventional setting or a classifier $f_g : \mathcal{I} \rightarrow \mathcal{L}_s \cup \mathcal{L}_u$ in the generalized setting [5].

3.1 Basic Idea of ZSL

Generally, the zero-shot classifier is a function taking the visual feature $v(\mathcal{I})$ of a testing image $\mathcal{I}$ and the class embeddings $\mathcal{K}_u$ and producing a class label with the largest similarity, which is defined as $f_c(\mathcal{I}) = \arg\max_{l \in \mathcal{L}_u} s(g(v(\mathcal{I};\phi);\theta), k_u^l)$. s denotes a measure of similarity between the predicted semantic embedding and class embedding, e.g. cosine similarity. $v(\mathcal{I};\phi)$ is typically an external feature extractor. $g(v;\theta)$ is the V-S mapping function from visual feature space to semantic embedding space with learnable parameter θ. The basic objective function is:

$$\mathcal{L} = \frac{1}{n_s} \sum_{i=1}^{n_s} l(g(v(\mathcal{I}_i;\phi);\theta), k_i) \tag{1}$$

where l is the loss function, e.g. mean square loss.

3.2 Joint Feature Learning with Semantic Enhancement for ZSL

In most ZSL methods, the feature extractor $v(\mathcal{I};\phi)$ is generally pre-trained with the off-the-shelf auxiliary data and keep unchanged during the whole training time, without consideration of the semantic gap between the visual features and the semantic embeddings. Our proposed model tries to adaptively learn image features with more consistent semantics under the guidance of class embeddings, so as to further benefit the V-S mapping. In specific, we base our model on local image features, since they can capture more local information that is relevant to the semantic embeddings and thus benefit the recognition at image level [9]. Then we adopt the attention mechanism to adaptively highlight the most discriminative and semantically relevant regions while ignore irrelevant or redundant regions for each class. In effect, the overall loss function becomes:

$$\mathcal{L} = \frac{1}{n_s} \sum_{i=1}^{n_s} l(g(v(\mathcal{I}_i, K;\phi);\theta), k_i) \tag{2}$$

where K contains all class embeddings and θ and ϕ are to be learned jointly.

We start with a compartment of convolutional nets responsible for learning local image features. Given an image $\mathcal{I}_i$, its local representations is $X_t^i = W * \mathcal{I}_i$, where W denotes the overall parameters of the deep network, $*$ is the set of operations on image $\mathcal{I}_i$ and $X_t^i \in \mathbb{R}^{T_1 \times T_2 \times d}$ is a tensor. We reshape this tensor along the third dimension and then transpose, yielding a $d \times T$ matrix $X_i = [x_i^1, x_i^2, ..., x_i^T], T = T_1 \times T_2$. It is our local representation of the input image, which stores the image feature at the t-th region in its t-th column vector of size d.

Learning Semantic-guided Features by Dense Attention Mechanism. We hypothesize that only a few semantically relevant regions in the entire image play an important role in ZSL. For example, we usually do not care much about the whole background information such as grass in the picture to recognize a 'sheep', verified by experimental results in Fig.2. Or several regions containing part of background maybe sufficient enough. To weight the regions that are highly relevant to the class embedding, we use dense attention in the sense that it correlates any region and any class for semantics alignment for ZSL. Given local representations X_i and seen class embeddings $K_s = [k_s^1, k_s^2, ..., k_s^{c_s}]$, the dense attention map is obtained with two steps. First, we feed K_s into the d-dimensional visual space through a simple encoder $h_e(k;\varphi) : \mathbb{R}^m \to \mathbb{R}^d$. Then, we compute the affinity matrix and normalize it in column-wise to derive the attention map, in which the class semantics are embedded. Last, we obtain the attended visual representations for each class by using multiplicative attention:

$$\begin{aligned} H_s &= h_e(K_s;\varphi); \\ A_i &= \mathrm{softmax}(X_i^{\mathrm{T}} H_s); \\ \tilde{X}_i &= X_i A_i, \end{aligned} \tag{3}$$

where the j-th column in $A_i \in \mathbb{R}^{T \times c_s}$ stores the semantic similarity of each region to the j-th class and the j-th column $\tilde{x}_i^j$ in $\tilde{X}_i \in \mathbb{R}^{d \times c_s}$ is the more informative image feature for the j-th class.

Naturally, if the input image is from j-th class, i.e. $y_i = l_s^j$, the attended image feature $\tilde{x}_i^j$ should contain most semantically relevant regions to the class embedding k_s^j. Accordingly, its predicted semantic embedding via the V-S mapping $g(\tilde{k}_i;\theta)$ and its ground-truth $k_s^{y_i}$ earns the highest score. A standard softmax loss can be used:

$$\mathcal{L}_{CLS} = -\frac{1}{n_s}\sum_{i=1}^{n_s} \log \frac{\exp(s(g(\tilde{x}_i^{y_i};\theta), k_s^{y_i}))}{\sum_{c=1}^{c_s} \exp(s(g(\tilde{x}_i^{c};\theta), k_s^{c}))} \tag{4}$$

Manifold Alignment (MA) Loss. In ZSL, it is important that the manifold structure of the semantic space and the visual space should be consistent [4, 16]. In order to ensure that the encoder h_e has the ability of manifold alignment, i.e. the latent visual representations of two nearby classes in semantic embedding space stay close as well, additional regularization term is necessary. We use a similarity graph to characterize the manifold structure [31] and match graphs across spaces to minimize the difference between two manifolds. Here, we use inner product similarity to construct each edge in the graph. The distance between H_s and K_s is given as:

$$\mathcal{L}_{MA} = ||H_s^T H_s - K_s^T K_s||_F^2 \tag{5}$$

where the subscript F denotes the Frobenius norm.

However, the above losses are still not efficient enough in ZSL, as evidenced by two reasons. First, as we all know, training dataset and testing dataset are totally disjoint in ZSL. Training without any semantic information about unseen classes will not well guarantee the generalization ability over them. Second, there would be a shift between these semantic-guided features and the real visual representation of each class, inducing the unexpected poor performance.

Semantics Reconstruction (SR) Loss. To the first problem, we employ the AE paradigm in the step of learning attended image features. We further impose a decoder $h_d(h;\psi):\mathbb{R}^d\rightarrow\mathbb{R}^m$ on the latent visual representation, whose goal is to faithfully reconstruct the semantic knowledges of all classes. To be specific, given class embeddings $K=[k_s^1,...,k_s^{c_s},k_u^1,...,k_u^{c_u}]$, the following loss should be minimized.

$$\mathcal{L}_{SR}=||K-h_d(h_e(K;\varphi))||_F^2 \tag{6}$$

The additional reconstruction loss imposes that h_e must also be able to preserve all the semantic information contained in the class embeddings. Although the semantics of different classes vary a lot, the demand for more truthful reconstruction of the semantic knowledge is generalizable.

Coupling between Decoder and the V-S Mapping. To the second problem, we propose to replace the V-S mapping g with the same decoder h_d, i.e. $g(\cdot)=h_d(\cdot)$. On one hand, the classification loss drives the latent visual representation of encoder be typical of the attended features of image in the same class, which drives the V-S mapping make a correct prediction. On the other hand, the semantic reconstruction constraint imposed on unseen classes ensures the generalization ability, to some extent.

Our final model combines all these losses, leading to the following objective:

$$\mathcal{L}=\mathcal{L}_{CLS}+\alpha\mathcal{L}_{MA}+\eta\mathcal{L}_{SR},s.t.\ g=h_d \tag{7}$$

where α and η are the weighting factors of the manifold alignment and semantic reconstruction loss, respectively. It can be optimized efficiently using standard back-propagation.

3.3 Unseen Class Prediction

Given a testing image $\mathcal{I}$ and a set of class embeddings $K=[k^1,...,k^c]$ (either from $\mathcal{K}_u$ in conventional ZSL or $\mathcal{K}_u\cup\mathcal{K}_s$ in generalized ZSL), we first learn the latent visual representation of K via h_e and compute the attended features $\tilde{X}=[\tilde{x}^1,...,\tilde{x}^c]$ by Eq.3. Then we compute its semantic embeddings for all these classes via h_d. After that, the classification of $\mathcal{I}$ is achieved by simply calculating its similarity to class embeddings in the semantic space, i.e.

$$y=\arg\max_c s(h_d(\tilde{x}^c;\theta),k^c) \tag{8}$$

Due to the semantic gap between seen and unseen classes, classifiers are inevitably biased toward seen classes. The prediction scores for the seen classes are often greater than that for unseen classes. Intuitively, we would like to mitigate the overconfidence to seen classes and increase the scores for unseen classes properly in generalized ZSL. Inspired from [5], we apply a calibration factor γ to balance these two conflicting forces during testing. The prediction function becomes:

$$y=\arg\max_c s(h_d(\tilde{x^c};\theta),k^c)+\gamma\mathbb{I}[c\in\mathcal{L}_u] \tag{9}$$

where γ is kind of the prior likelihood of a data coming from unseen classes.

4 Experiments

4.1 Experimental setup

Datasets. We conduct experiments on four popular benchmarks, including the Animals with Attributes (AwA1) [13], AwA2 [32], CUB-200-2011 (CUB) [27] and SUN attribute (SUN) [20], which contain 85, 85, 312 and 102 class-level attributes, respectively. We use the proposed data splits in [32] to make sure the absence of any testing image during training.

Competitors. We compare our method with representative ones published in the past few years and the state-of-the-art ones reported recently. They are: ESZSL [22], ALE [1], SJE [2], SSE [35], DeViSE [8], ConSE [19], SAE [12], LESAE [18], SynC [4], Relation Net [24], DCN [17], GAZSL [36] and f-CLSWGAN [33].

Experimental details We use representations from layer conv5 in the widely accepted ResNet-101 [10] with the size of $7 \times 7 \times 2048$. Then we reshape these feature maps to form the foundational image representation with size of 2048×49, i.e. $d = 2048$ and $T = 49$. In the AE paradigm, we use one hidden layer with sigmoid activation for encoder and a fully connected layer for decoder. Inference can be simply done in a single feedforward. We class-wise cross-validate hyper-parameters using the training data and set α and η to be 0.5 and 1, respectively. Although competitors and our methods are different in the way of extracting image features, they all explore the same deep architecture pre-trained on the same ImageNet, for fair comparison.

4.2 Conventional Zero-Shot Learning Results

We first evaluate the proposed method on conventional ZSL problem, where the testing data is from unseen classes by default. Tab.1 reports the published average top-1 accuracy of these comparing methods on the proposed split of these datasets.

Table 1: ZSL average accuracy (%) in the conventional setting on AwA1, AwA2, CUB and SUN. '-' denotes no results reported.

Method	AwA1	AwA2	CUB	SUN
ConSE	45.6	44.5	34.3	38.8
DeViSE	54.2	59.7	52.0	56.5
ESZSL	58.2	58.6	53.9	54.5
ALE	59.9	62.5	54.9	58.1
SJE	65.6	61.9	53.9	53.7
SSE	60.1	61.0	43.9	51.5
SynC	54.0	46.6	55.6	56.3
SAE	65.4	66.2	53.6	59.7
LESAE	66.1	68.4	53.9	60.0
Relation Net	68.2	64.2	55.6	-
DCN	65.2	-	56.2	**61.8**
GAZSL	68.2	-	55.8	61.3
f-CLSWGAN	68.2	-	57.3	60.8
Ours w/o SR	66.8	63.6	55.1	57.9
Ours	**69.5**	**70.4**	**59.4**	59.6

We can make the following observations: (1) The proposed method achieves the best or comparable results on all these datasets. It outperforms both shallow models using global image features and deep models using ResNet. For example, on the fine-grained CUB, the

performance improvement is 3.8% over the manifold alignment based SynC. This justifies the significance of adaptively extracting image features over using constant static features for the task at hand in ZSL, to some extent. However, it performs slightly worse on CUB than the similar work S^2GA [□], which is reasonable since the latter leverages the efficient part detection network SPDA-CNN instead of the general network to extract region features. Relatively speaking, ZSL performance on fine-grained dataset is still less than coarse-grained dataset. We conjecture that the much more similar class embeddings is one possible cause. (2) Our method further improves the accuracy from **Ours w/o SR**, the variant of the proposed method without using a decoder in obtaining the attention-based image features, i.e. $\eta = 0$ in Eq.7. On AwA2, the improvement achieves 6.8%, while the average improvement is about 3.4%. This validates the effectiveness of the semantics reconstruction constraint, i.e. the AE paradigm, for mitigating ZSL model from overfitting on seen classes and thus improving the generalization capability over unseen classes.

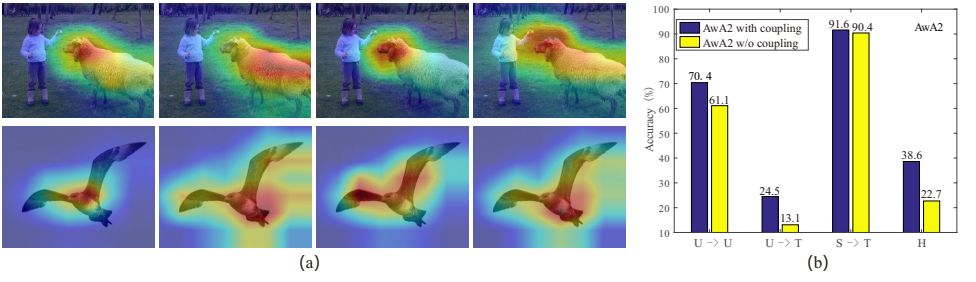

Figure 2: (a) Visualization of attention on AwA1 and CUB. Heatmaps in each row show the regions of interest obtained by the ground-truth class embeddings and other three negative class embeddings (best viewed in color). The first row goes sheep, seal, horse and blue whale. The second row goes pomarine jaeger, yellow headed blackbird, brandt cormorant and groove billed ani. (b) The importance of coupling between V-S mapping and the decoder on AwA2.

Visualization of Attention. To better comprehend the effectiveness of adaptively attention-based image features for ZSL, we further provide a qualitative visualization of the regions of interest generated by the attention maps of different class embeddings. To be specific, we first use linear interpolation to up-sample the attention map obtained by the class embeddings to the size of input image, i.e. from 7×7 to 224×224. Then, we multiply the input image with this attention map and show the heatmap in Fig.2 (a). For comparison, we illustrate the unseen image regions generated by the ground-truth label and other negative labels on AwA1 and CUB.

We can make the following observations. On one hand, the proposed method has the ability to highlight the main object area and ignore much of the noisy background. Even if there are multiple objects in the input image, it can successfully select the region where the target class is located. On the other hand, the ground-truth class focuses on much different regions from other negative classes, e.g. the neck area is most semantically related to 'sheep'. Features emphasizing other area would shift from the ground truth label in the semantic embedding space and thus results in the incorrect prediction. While without the manifold alignment constraint, i.e. $\alpha = 0$, these classes basically focus on almost the same object area without much difference. Tab.1 and Fig.2 (a) demonstrate both the feasibility of the proposed method and the necessity of automatically learning image features for ZSL.

4.3 Generalized Zero-Shot Learning Results

In generalized ZSL, where whether the testing instance is from seen or unseen classes is unknown in advance, we follow the same setting of [5]. We hold out 20% of data samples from seen classes and mix them with data samples from unseen classes for testing. We report the harmonic mean of seen and unseen class accuracies, i.e. $H = 2 \times (a_{\mathcal{U}\to\mathcal{T}} \times a_{\mathcal{S}\to\mathcal{T}})/(a_{\mathcal{U}\to\mathcal{T}} + a_{\mathcal{S}\to\mathcal{T}})$, where $a_{\mathcal{U}\to\mathcal{T}}$ ($a_{\mathcal{S}\to\mathcal{T}}$) is the top-1 accuracy of classifying samples from unseen (seen) classes into the joint label space. We set the calibration parameter γ in Eq.9 by means of cross validation on held-out seen classes in the new training dataset.

Table 2: ZSL average accuracy (%) in the generalized setting on AwA1, AwA2, CUB and SUN. '-' denotes no results reported.

Method	AwA1			AwA2			CUB			SUN		
	$a_{\mathcal{S}\to\mathcal{T}}$	$a_{\mathcal{U}\to\mathcal{T}}$	H	$a_{\mathcal{S}\to\mathcal{T}}$	$a_{\mathcal{U}\to\mathcal{T}}$	H	$a_{\mathcal{S}\to\mathcal{T}}$	$a_{\mathcal{U}\to\mathcal{T}}$	H	$a_{\mathcal{S}\to\mathcal{T}}$	$a_{\mathcal{U}\to\mathcal{T}}$	H
ESZSL	75.6	6.6	12.1	77.8	5.9	11.0	63.8	12.6	21.0	27.9	11.0	15.8
ALE	76.1	16.8	27.5	81.8	14.0	23.9	62.8	23.7	34.4	33.1	21.8	26.3
SJE	74.6	11.3	19.6	73.9	8.0	14.4	59.2	23.5	33.6	30.5	14.7	19.8
SSE	80.5	7.0	12.9	82.5	8.1	14.8	46.9	8.5	14.4	36.4	2.1	4.0
SynC	87.3	8.9	16.2	90.5	10.0	18.0	70.9	11.5	19.8	**43.3**	7.9	13.4
ConSE	88.6	0.4	0.8	90.6	0.5	1.0	**72.2**	1.6	3.1	39.9	6.8	11.6
DeViSE	68.7	13.4	22.4	74.7	17.1	27.8	53.0	23.8	32.8	27.4	16.9	20.9
SAE	77.1	1.8	3.5	82.2	1.1	2.2	54.0	7.8	13.6	18.0	8.8	11.8
LASAE	70.2	19.1	30.0	70.6	21.8	33.3	53.0	24.3	33.3	34.7	21.9	26.9
DCN	84.2	25.5	39.1	-	-	-	60.7	28.4	38.7	37.0	25.5	30.2
SE-GZSL	67.8	56.3	61.5	68.1	58.3	62.8	53.3	41.5	46.7	30.5	40.9	34.9
f-CLSWGAN	68.9	52.1	59.4	61.4	57.9	59.6	57.7	43.7	**49.7**	36.6	**42.6**	**39.4**
Relation Net	**91.3**	31.4	46.7	**93.4**	30.0	45.3	61.1	38.1	47.0	-	-	-
Ours w/o SR	88.4	34.5	49.6	89.5	19.7	32.3	53.6	33.2	41.0	35.9	20.4	26.0
Ours w/o C	88.8	37.3	52.5	91.6	24.5	38.3	59.8	36.6	45.4	36.6	23.2	28.4
Ours	73.5	**61.9**	**67.2**	73.8	**63.9**	**68.5**	45.1	**52.1**	48.3	34.5	28.4	31.2

Results in Tab.2 are much lower than conventional ZSL in Tab.1, especially unseen class accuracies. This is because when semantically related seen classes are included in the search space, ZSL models trained on these classes are more inclined to predict the testing image as a seen class. Competitors having a large performance discrepancy between seen and unseen classes often have a small H value. Our method achieves better results than most competitors on both unseen class recognition and the harmonic mean, which means it has a much better generalization ability. On CUB and SUN, it performs worse than GAN-based models, i.e. SE-GZSL [26] and f-CLSWGAN, which synthesize a great amount of examples for unseen classes to train classifiers. Since CUB and SUN have an average of only 59 and 20 images per class, too little data may be one reason for our poor performance. Additionally, it performs better than the proposed method without decoder, i.e. **Ours w/o SR**, highlighting the efficacy of the semantics reconstruction constraint. In contrast with **Ours w/o C**, the variant without calibration factor during testing, ours performs much better. It also easily outperforms the state-of-the-arts reported very recently, which validates the necessity of mitigating the bias toward seen classes. The harmonic mean actually provides a quantitative analysis of the inherent trade-off between recognizing seen classes and recognizing unseen classes, which is susceptible to minimal value. By varying γ in the proposed method, we also find that when the performance of one of the parties is too low, H will drop sharply even if the performance of the other party is high.

Ablation Study on the Coupling Mechanism. One key strength of our model comes from the coupling mechanism between the V-S mapping and the decoder in the AE paradigm. In order to evaluate its importance, we decouple this mechanism by using two different

functions. Results on AwA2 in Fig.2 (b) show that in both conventional and generalized ZSL, the coupling mechanism can make a difference.

5 Conclusion

We analyzed two possible shortcomings in current zero-shot learning approaches and proposed a novel end-to-end method based on the autoencoder paradigm, where image feature learning and visual-semantic mapping construction are carried out jointly. It automatically learns task-specific features more relevant to class semantics based on a dense attention map, which is computed by the aligned latent representations of class embeddings. At the same time, it couples the visual-semantic mapping with decoder to make it more generalizable through a semantics reconstruction loss over unseen classes. Experimental results on four benchmarks in two settings show the impressive effectiveness of the proposed method.

6 Acknowledgement

This work was supported by Zhejiang Lab AI project (No. 2018EC0ZX02, No. 2018EC0ZX01), the National Natural Science Foundation of China under grant 61473256 and the China Knowledge Center for Engineering Science and Technology.

References

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

[2] 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.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

[5] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *ECCV*, pages 52–68. Springer, 2016.

[6] Mohamed Elhoseiny, Yizhe Zhu, Han Zhang, and Ahmed M Elgammal. Link the head to the" beak": Zero shot learning from noisy text description at part precision. In *CVPR*, pages 6288–6297, 2017.

[7] Rafael Felix, Vijay BG Kumar, Ian Reid, and Gustavo Carneiro. Multi-modal cycle-consistent generalized zero-shot learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 21–37, 2018.

[8] 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.

[9] Jianlong Fu, Heliang Zheng, and Tao Mei. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *CVPR*, volume 2, page 3, 2017.

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

[11] Zhong Ji, Yanwei Fu, Jichang Guo, Yanwei Pang, Zhongfei Mark Zhang, et al. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *NIPS*, pages 5998–6007, 2018.

[12] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *CVPR*, pages 3174–3183, 2017.

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

[14] Jingjing Li, Mengmeng Jin, Ke Lu, Zhengming Ding, Lei Zhu, and Zi Huang. Leveraging the invariant side of generative zero-shot learning. *arXiv preprint arXiv:1904.04092*, 2019.

[15] Yan Li, Junge Zhang, Jianguo Zhang, and Kaiqi Huang. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, pages 7463–7471, 2018.

[16] Yanan Li, Donghui Wang, Huanhang Hu, Yuetan Lin, and Yueting Zhuang. Zero-shot recognition using dual visual-semantic mapping paths. In *CVPR*, pages 5207–5215, 2017.

[17] Shichen Liu, Mingsheng Long, Jianmin Wang, and Michael I Jordan. Generalized zero-shot learning with deep calibration network. In *NIPS*, pages 2009–2019, 2018.

[18] Yang Liu, Quanxue Gao, Jin Li, Jungong Han, and Ling Shao. Zero shot learning via low-rank embedded semantic autoencoder. In *IJCAI*, pages 2490–2496, 2018.

[19] 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.

[20] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2751–2758. IEEE, 2012.

[21] Scott Reed, Zeynep Akata, Honglak Lee, and Bernt Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, pages 49–58, 2016.

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

[23] Edgar Schönfeld, Sayna Ebrahimi, Samarth Sinha, Trevor Darrell, and Zeynep Akata. Generalized zero-and few-shot learning via aligned variational autoencoders. *arXiv preprint arXiv:1812.01784*, 2018.

[24] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. Learning to compare: Relation network for few-shot learning. In *CVPR*, pages 1199–1208, 2018.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[26] Vinay Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *CVPR*, 2018.

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

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

[29] Qian Wang and Ke Chen. Zero-shot visual recognition via bidirectional latent embedding. *IJCV*, 124(3):356–383, 2017.

[30] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *TIST*, 10(2):13, 2019.

[31] Xiaolong Wang, Yufei Ye, and Abhinav Gupta. Zero-shot recognition via semantic embeddings and knowledge graphs. In *CVPR*, pages 6857–6866, 2018.

[32] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, pages 4582–4591, 2017.

[33] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *CVPR*, 2018.

[34] Dongfei Yu, Jianlong Fu, Tao Mei, and Yong Rui. Multi-level attention networks for visual question answering. In *CVPR*, pages 4187–4195. IEEE, 2017.

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

[36] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1004–1013, 2018.