The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

Deep Semantic Structural Constraints for Zero-Shot Learning

Yan Li,[*][1,2] **Zhen Jia,**[*][1,2] **Junge Zhang,**[1,2] **Kaiqi Huang,**[1,2,3] **Tieniu Tan**[1,2,3]
[1] CRIPAC & NLPR, Institute of Automation, Chinese Academy of Sciences
[2] University of Chinese Academy of Sciences
[3] CAS Center for Excellence in Brain Science and Intelligence Technology
yan.li@cripac.ia.ac.cn, {zhen.jia, jgzhang, kqhuang, tnt}@nlpr.ia.ac.cn

Abstract

Zero-shot learning aims to classify unseen image categories by learning a visual-semantic embedding space. In most cases, the traditional methods adopt a separated two-step pipeline that extracts image features from pre-trained CNN models. Then the fixed image features are utilized to learn the embedding space. It leads to the lack of specific structural semantic information of image features for zero-shot learning task. In this paper, we propose an end-to-end trainable Deep Semantic Structural Constraints model to address this issue. The proposed model contains the Image Feature Structure constraint and the Semantic Embedding Structure constraint, which aim to learn structure-preserving image features and endue the learned embedding space with stronger generalization ability respectively. With the assistance of semantic structural information, the model gains more auxiliary clues for zero-shot learning. The state-of-the-art performance certifies the effectiveness of our proposed method.

Introduction

As one of the most basic problems in the computer vision area, image classification methods gain huge progress in recent years with the impressive development of deep learning. Although ResNet (He et al. 2016), an outstanding representation of the Convolutional Neural Network (CNN) classification models, gets the top-5 error rate as low as 3.57% on ImageNet classification task, its classification ability is still limited to the image categories in the training dataset. The limitation that models can only classify image categories within the training set, restricts them to become more intelligent as human beings. For a simple example, human beings are able to classify different kinds of animals by just reading their descriptions rather than seeing them. More and more researchers try to break through this limitation by introduce Zero-Shot Learning (ZSL) into image classification (Lampert, Nickisch, and Harmeling 2009; Frome et al. 2013; Norouzi et al. 2013; Socher et al. 2013; Fu et al. 2015; Akata et al. 2015; Romera-Paredes and Torr 2015; Bucher, Herbin, and Jurie 2016; Akata et al. 2016; Huang, Loy, and Tang 2016; Changpinyo et al. 2016; Xian et al. 2017; Morgado and Vasconcelos 2017).

Zero-shot learning seeks to make image classification models able to classify image categories which never appear in the training dataset. In the zero-shot learning task, we refer to the image categories in the training set as *seen classes* and those in the test set as *unseen classes*. The category characteristic of unseen classes are learned from the *side information*, *i.e.*, the semantic features of the images. The commonly used side information could be the human annotated attribute features of images (Lampert, Nickisch, and Harmeling 2009; Akata et al. 2016), the text descriptions of the image categories (Reed et al. 2016), word vectors of the category labels (Frome et al. 2013; Norouzi et al. 2013) and so on.

A large number of previous state-of-the-art methods focus on building a common space where image features and semantic features are embedded (Frome et al. 2013; Socher et al. 2013; Akata et al. 2015; Romera-Paredes and Torr 2015; Akata et al. 2016). The embedding space is built on the correspondence between the seen images and their semantic features. Then in the test stage, unseen image features will be mapped to the embedding space where the classification method, such as nearest neighbour (NN) search, can be operated easily. Most of these methods adopt a separated two-step pipeline, *i.e.*, extracting image features from pre-trained CNN models and using fixed image features to learn the embedding space.

However, we argue that separating the image feature extraction and the embedding space construction harms the ZSL models severely. The separation leads to the result that models cannot regulate the image features for the specific ZSL task during training. What's more, the image features extracted from a fixed pre-trained CNN model will not capture the plentiful semantic information in the side information. The semantic information of human annotated attributes, text descriptions or word vectors constructs the semantic structure of a specific category. We believe that combining the learning of image features and embedding space in an end-to-end manner, meanwhile, incorporating the structural information into the whole learning process would contribute to much better zero-shot performance.

In this paper, we come up with a new Deep Semantic Structural Constraints (DSSC) model for zero-shot learning looking forward to training the model in an end-to-end style and using the semantic structural information to supervise

[*]The first two authors contributed equally to this work.

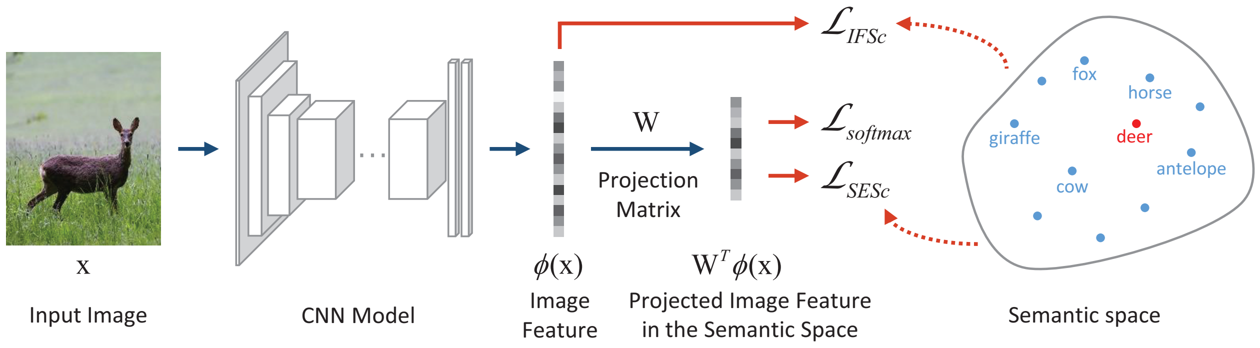

Figure 1: The framework of Deep Semantic Structural Constraints (DSSC) model. The Image Feature Structure constraint (IFSc) and the Semantic Embedding Structure constraint (SESc) use the structural information of semantic space to supervise the model's image feature extraction and embedding space construction respectively.

the image feature extraction and embedding space construction. The contributions of this work are as follows:

- Distinguished from most of the current work, the DSSC model is end-to-end trainable. Benefit from the end-to-end training, the image feature extraction can be brought into the training stage. We can add specific supervision to train the model and make the features more suitable for the ZSL task.
- In DSSC model, we set the Image Feature Structure constraint (IFSc) specially in the image feature extraction stage. The constraint makes the model automatically learn structure-preserving image representations. The learned image features rather than the fixed features extracted from pre-trained CNN models are beneficial to the construction of semantic embedding space. To the best of our knowledge, this is the first attempt to add specific image feature extraction constraint in ZSL models.
- For the construction of embedding space, the DSSC model contains the Semantic Embedding Structure constraint (SESc) to learn a better embedding space for the unseen classes by relaxing the commonly used strict softmax loss. The embedding space directly learned on seen classes will not generalize the unseen classes well due to the domain shift problem (Fu et al. 2015). The semantic embedding structure constraint can endue the model with much better generalization capability for the ZSL task.
- As shown by the experiments, the DSSC model gets the state-of-the-art performance on the representative datasets of zero-shot learning.

Related Work

Zero-shot learning aims to classify images of unseen classes which is an impossible challenge for traditional image classification models.

In recent years, researchers try to build a common embedding space for image and their semantic features by learning a compatibility function between them. The DeViSE model (Frome et al. 2013) and the ALE model (Akata et al. 2016) learn linear transformation matrices by solving a hinge ranking loss function. R. Socher et al. operate a two-layer neural network to map unseen images to the semantic feature space in (Socher et al. 2013). The ESZSL model (Romera-Paredes and Torr 2015) adopts a principled designed Frobenius norm regularizer making itself simple but efficient. The SJE model (Akata et al. 2015) combines several compatibility function linearly to form a joint embedding which captures the non-redundant information from different aspects of side information. Most recently, the SAE model (Kodirov, Xiang, and Gong 2017) draws successful lessons from linear auto-encoder. Besides the encoder mapping images to the semantic space, it employs a decoder to make the model able to reconstruct the original visual feature from the mapped semantic feature. Similar to DeViSE and ALE, N. Karessli et al. optimize a hinge ranking loss in (Karessli et al. 2017), but they introduce the gaze information as side information to the ZSL task innovatively.

In addition to learning the compatibility function directly, hybrid models (Norouzi et al. 2013; Changpinyo et al. 2016) bring us another way to think about zero-shot learning. The basic idea of hybrid models is to use the composition of seen classes to classify the unseen images. The ConSE model (Norouzi et al. 2013) convexly combines the image classification probabilities of seen classes to determine the unseen classification result. The SynC model (Changpinyo et al. 2016) synthesizes the classifiers of agent classes as the classifier of the unseen class. Almost all these methods separate the image feature extraction and embedding space construction. They use fixed image features when building the embedding space. This trend leads to the image features' lack of the semantic information in ZSL task.

Similar to our work, the SCoRe model (Morgado and Vasconcelos 2017) also puts forward an end-to-end training pipeline for ZSL. The authors add two semantic constraints at the end of a CNN classification model to supervise the attribute prediction and the category classification respectively. Although the SCoRe model is end-to-end trainable and deployed with specific constraints for ZSL task, it does not consider any semantic structural constraints in the image feature extraction stage. It remains the standard CNN image feature extraction pipeline in the model.

To address the shortcomings of previous works, we bring the image feature extraction into training by deploying an end-to-end model. What's more, we incorporate the semantic structural information into the complete learning process. Concretely, we propose a Deep Semantic Structural Constraints (DSSC) model which contains the Image Feature Structure constraint (IFSc) and the Semantic Embedding Structure constraint (SESc). The IFSc aims to preserve semantic structural information in image feature extraction. While the SESc supervises embedding space construction in order to strengthen the generalization ability for unseen classes. As shown in the experiments, the proposed two constraints contribute to the improvement of zero-shot classification.

Methodology

In this section, we formalize the zero-shot learning problem at first. Then we introduce the proposed Deep Semantic Structural Constraints model in detail.

Problem Statement

In the zero-shot learning task, we define the training set as $\mathcal{S} \equiv \{(x_s, y_s, a_s)^n, n = 1, \ldots, N\}$, where $x_s \in \mathcal{X}_\mathcal{S}$ is the image of the seen class and $y_s \in \mathcal{Y}_\mathcal{S}$ is the corresponding image label. Similarly, the test set is defined as $\mathcal{U} \equiv \{(x_u, y_u, a_u)^{n'}, n' = 1, \ldots, N'\}$, where $x_u \in \mathcal{X}_\mathcal{U}$ is the unseen image and $y_u \in \mathcal{Y}_\mathcal{U}$ is the label of it. As to the side information, $a_s, a_u \in \mathcal{A}$ are the semantic features of each category or image. According to the definition of zero-shot learning, $\mathcal{Y}_\mathcal{S} \cap \mathcal{Y}_\mathcal{U} = \emptyset$, *i.e.*, only the seen class images are available during the training stage. The goal of ZSL is to learn a classifier $f : \mathcal{X}_\mathcal{U} \to \mathcal{Y}_\mathcal{U}$ on training set $\mathcal{S}$ with the assistance of side information $\mathcal{A}$.

Baseline Model

Our baseline ZSL model aims to learn compatibility function to associate visual and auxiliary semantic information. Formally,

$$F(x, y; \mathbf{W}) = \phi(x)^T \mathbf{W} \psi(y) \tag{1}$$

where $\phi(x)$ is the image representations typically extracted by pre-trained CNN models and $\psi(y)$ is the semantic feature of category y, *i.e.*, $\psi(y) \in \mathcal{A}$.

Similar to most of previous ZSL methods, we use the semantic space of $\mathcal{A}$ as the common embedding space and define the compatibility score by the inner product. Formally,

$$\mathbf{s} = \langle \mathbf{W}^T \phi(x), \psi(y) \rangle \tag{2}$$

where $\mathbf{W}$ is the weight to learn in a fully connected layer. It can be viewed as a linear projection matrix that maps image representations $\phi(x)$ to the semantic space $\mathcal{A}$.

Like the classification scores in traditional object recognition task, the compatibility scores are used to measure the matching degree between an image and the semantic representations of classes. Similar to object recognition task, we use a standard softmax loss to train our ZSL model:

$$\mathcal{L}_{softmax} = -\frac{1}{N} \sum_i^n \log \frac{\exp(\mathbf{s})}{\sum_c \exp(\mathbf{s}_c)}, c \in \mathcal{Y}_\mathcal{S} \tag{3}$$

At the test stage, the classification result of an unseen image x_u can be achieved by simply selecting the most matched category from unseen classes. Formally,

$$y^* = \underset{c \in \mathcal{Y}_\mathcal{U}}{argmax} \, (\mathbf{s}_c) = \underset{c \in \mathcal{Y}_\mathcal{U}}{argmax} \, \langle \mathbf{W}^T \phi(x_u), \psi(y_c) \rangle \tag{4}$$

It is noted that the proposed model is end-to-end trainable and the image representations $\phi(x)$ are also learnable during the training process. Most of previous ZSL methods adopt the fixed $\phi(x)$ and only focus on learning the projection matrix $\mathbf{W}$ with additional regularization terms. However, we argue that it cannot regulate the image representations for specific ZSL tasks using fixed image features $\phi(x)$. On the contrary, making the image features compatible with side information of ZSL task in an end-to-end framework is necessary and will contribute to better performance. In the experiments section, we show that, without any other regularization terms, the performance of our baseline model is comparable to previous state-of-the-art results.

Image Feature Structure Constraint (IFSc)

The baseline model utilizes softmax loss in Equation 3 to project image instances to its corresponding semantic representations. In most of cases, the semantic representations, for example human-annotated attributes or text descriptions, contain more plentiful expert knowledge than the simple visual appearances. They have sufficient capacity of discriminating different classes (*i.e.*, the intra-class distance is small while the inter-class distance is large in semantic space).

On the contrary, the image features only imply the visual cues, such as colors, shapes, textures and so on. When only considering visual representations, the distribution of different categories may be amorphous and unstructured. 1) The inter-class distance between some categories with similar visual appearances are very small. For example, both *conference center* and *court room* have multiple chairs in their images while the two categories can be distinguished in semantic space, as the *conference center* may have the attribute of *conducting business* which will never appear in *court room*. 2) On the other hand, the intra-class distance for some categories may be too large. For example, all images of the *assembly line* share the same *working* and *using the tool* attributes but the visual elements in *assembly line* images change drastically with different products and various implements. It is difficult to directly project such unstructured visual representations to discriminative semantic space using a projection matrix $\mathbf{W}$. To address this issue, we propose the Image Feature Structure constraint (IFSc) aiming to use a triplet loss to regulate the intra-class and inter-class distances in the image space. Formally,

$$\begin{aligned} &\mathcal{L}_{IFSc} \\ &= \max(0, m_{IFSc} + d(\phi(x_i), \phi(x_k)) - d(\phi(x_i), \phi(x_j))) \end{aligned} \tag{5}$$

where x_i denotes the anchor image, x_k is the image from the same class ($y_i = y_k$). x_j is from a different class ($y_i \neq y_j$). $d(x, y)$ is used to calculate the squared Euclidean distance between x and y, *i.e.*, $d(x, y) = \|x - y\|_2^2$. m_{IFSc} is the margin of the IFSc and is set to 1.0 for all experiments. Under the guidance of semantic information, the IFSc aims to learn structure-preserving image representations. It assists the baseline model to learn the visual-semantic embedding space more easily with such structured image features.

Semantic Embedding Structure Constraint (SESc)

In the baseline model, maximizing the compatibility score with softmax loss is similar to minimizing the Euclidean distance between the image and semantic representations, considering the semantic representations $\psi(y)$ are typically L2-normalized in ZSL tasks. In other words, the baseline model aims to exactly project the image instances to corresponding semantic representations. Such strict projection may cause domain-shift problem between seen and unseen classes at the test stage. To address this issue, we need to relax the strict softmax loss to endue the model with stronger generalization ability. Meanwhile, even with the relaxed projection, the projected image features are still required to be compatible with the semantic features. Based on the above considerations, we propose a Semantic Embedding Structure constraint (SESc) as follows:

$$\begin{aligned}\mathcal{L}_{SESc} = \max(0, m_{SESc} + d(\mathbf{W}^T\phi(x_i), \mathbf{W}^T\phi(x_k)) \\ - d(\mathbf{W}^T\phi(x_i), \mathbf{W}^T\phi(x_j)))\end{aligned} \tag{6}$$

Instead of exactly projecting image features to the point of semantic representations, the SESc aims to learn a relaxed manifold. Meanwhile the structure of learned manifold still align the semantic space properly. Using SESc as an extra relaxation to train the baseline model, as shown in experiments, it makes the learned model generalize better to unseen classes.

Self-adaptive margin As the common practice in triplet loss, the margin m_{IFSc} is set to 1.0 for all categories in IFSc. However, in SESc, we adaptively set the margin m_{SESc} according to the Euclidean distance between each pair of the semantic representations, *i.e.*, $m_{SESC(y_i,y_j)} = \|\psi(y_i) - \psi(y_j)\|_2^2$.

It is noted that our goal is still to map the image features $\phi(x_i)$ to its corresponding semantic representations $\psi(y_i)$. Thus, in the triplet formulation, the intra-class and inter-class distances of the projected image features should be compatible with the distance between the corresponding semantic representations.

Let's consider a case with two categories i and j. The semantic distance between the two categories is 0.5, *i.e.*, $\|\psi(y_i) - \psi(y_j)\|_2^2 = 0.5$. If we set margin in SESc to 1.0, the model has to push the projected image features $\mathbf{W}^T\phi(x_i)$ and $\mathbf{W}^T\phi(x_j)$ away from their semantic representations $\psi(y_i)$ and $\psi(y_j)$, which is obvious inappropriate. Finally, it will lead to an undesired inferior embedding space.

Table 1: Benchmarks of ZSL.

Dataset	Instances	SS	SS-D	Classes
AwA	30,475	Attribute	85	40/10
CUB	11,788	Attribute	312	150/50
SUN	14,340	Attribute	102	645/72

'SS' denotes semantic type.
'SS-D' indicates the dimension of the semantic space.

Deep Semantic Structure Constraints Model

Combining the mentioned three constraints, we obtain our deep semantic structure constraints model (DSSC), as illustrated in Figure 1. Formally,

$$\mathcal{L}_{DSSC} = \mathcal{L}_{softmax} + \lambda\mathcal{L}_{IFSc} + \beta\mathcal{L}_{SESc} \tag{7}$$

where λ and β are trivially set to 1.0 for all the experiments.

Experiments

Datasets

We evaluate the DSSC model and compare with existing state-of-the-art approaches on three standard zero-shot learning benchmark datasets. Table 1 summarizes their statistics.

- **Animals with Attributes (AwA)** (Lampert, Nickisch, and Harmeling 2014) includes 30,475 images from 50 animals categories. We adopt the class-level continuous 85-dim attributes as the semantic representations and use the standard 40/10 zero-shot split.
- **Caltech-UCSD Birds 200-2011 (CUB)** (Wah et al. 2011) is a fine-grained bird dataset with 200 different species of birds and 11,788 images. Each image is annotated with a 312-dim binary attribute vector and the class-level continuous attributes are also provided. We follow SynC (Changpinyo et al. 2016) to use 150/50 zero-shot split setting and utilize the class-level attributes as the semantic representations.
- **SUN-Attribute (SUN)** (Patterson et al. 2014) contains 14,340 images coming from 717 fine-grained scenes. Each category includes 20 images. In SUN, each sample is paired with a binary 102-dim attribute vector. We compute class-level continuous attributes as our semantic representations by averaging the image-level attributes for each class. Following SynC, we use 645 classes of SUN for training and 72 classes for test.

Image Representations

We use GoogLeNet (Szegedy et al. 2015) (layer pool5 with 1024 units) pre-trained on ImageNet (Russakovsky et al. 2015) to implement image representation, $\phi(x)$, for all three datasets. During the training stage, all images are resized to 256×256 pixels. When dealing with CUB, the SynC model

(Changpinyo et al. 2016) crops all images with the provided bounding boxes. However in our experiments, we do not utilize the additional bounding box annotations and directly process the entire image for training.

Additionally, for SUN, the CNN model pre-trained on the MIT Places dataset (Zhou et al. 2014) is used to obtain better performance on scene classification tasks in (Changpinyo et al. 2016). Explored by (Xian et al. 2017), the MIT Places dataset has intersected categories with SUN for both the seen and unseen classes. Using the CNN model pre-trained on MIT Places dataset will violate the setting of zero-shot learning. But for fair comparison, we also provide the experiment result with the MIT Places pre-trained CNN model.

It should be noted that all the CNN models are not fixed but trainable in our experiments.

Evaluation Criteria

We use the multi-way classification accuracy (MCA) for all three datasets as the same with previous works. Comparing with per-image accuracy, MCA is more suitable to measure the classification performance when the dataset is class-imbalanced.

Experimental Results

Effectiveness of the trainable image feature extraction and semantic structural constraints In the DSSC model, we adopt two semantic structural constraints, IFEc and SESc, which are respectively used to improve the image feature learning and embedding space construction. In this part, we first compare our DSSC model to **four** variant ZSL models to verify the effectiveness of the trainable image feature extraction and two semantic constraints.

1) **Baseline**: The baseline model sets λ=0 and β=0 in Equation 7.
2) **Baseline-fixed**: The baseline-fixed model fixes the parameters of the image feature extraction CNN model during training.
3) **Deep-IFSc**: The Deep-IFSc model sets β=0 in Equation 7.
4) **Deep-SESc**: The Deep-SSRc model sets λ=0 in Equation 7.

All these four models and our DSSC model are trained end-to-end and evaluated on all three datasets. The comparison results are shown in the second component of Table 2.

We first notice that the baseline model has already achieved comparable performance with previous state-of-the-art methods without any additional regularization terms. Different from previous ZSL methods learning compatibility function with fixed image features, our baseline model is trained end-to-end. The image features $\phi(x)$ are also learned to be compatible with the semantic information for different ZSL datasets. In order to verify the advantage of the trainable image feature extraction and end-to-end framework, we fix the parameters of the image feature extraction CNN during training of the baseline model. As shown in Table 2, the performance of the baseline model considerably surpasses the baseline-fixed model.

Table 2: Comparative ZSL classification accuracy (%).

Method	AwA	CUB	SUN
DAP (2014)	60.1	-	44.5
ESZSL (2015)	75.3	48.7	18.7
SJE (2015)	73.9	50.1	56.1
SynC (2016)	72.9	54.4	59.2
SynC-Places (2016)	-	-	62.7
SAE (2017)	84.7	61.4	65.2
SAE* (2017)	80.7	33.4	42.4
MFMR (2017)	76.6	46.2	-
Low-Rank (2017)	76.6	56.2	-
SCoRe (2017)	78.3	58.4	-
Baseline (Ours)	75.19	58.26	58.75
Baseline-fixed (Ours)	73.70	50.31	54.63
Deep-IFSc (Ours)	77.76	58.80	60.21
Deep-SESc (Ours)	76.89	64.54	59.65
DSSC (Ours)	**78.51**	**65.99**	60.76
DSSC-Places (Ours)	-	-	**67.22**

SAE* denotes the reproducing results with ResNet-101 ImageNet features reported by (Xian et al. 2017).
For the SUN dataset, SynC-Places and DSSC-Places denote the results using pre-trained MIT Places model.

When comparing the Deep-IFEc and Deep-SESc with the baseline method, both the two models obtain improvements on the three datasets. When combining the two constraints, the ZSL performance of our DSSC model can be further improved. We also observe that, on CUB, Deep-IFSc only obtain slight improvements compared with the baseline (58.80% *vs*. 58.23%). We believe the reason is that the attribute annotations of CUB dataset is all about visual cues, such as colors or shapes. Thus, the image features $\phi(x)$, to some extent, have already captured the semantic information contained in these attributes. In such cases, the original image features $\phi(x)$ are "structured" and we can directly learn the embedding space without IFSc. On SUN dataset where annotated attributes have little connection with visual elements, the improvement of Deep-IFSc is more solid.

Comparisons with state-of-the-art methods Then we compare our method with previous ZSL approaches. Among the comparison methods, the published results of SAE (Kodirov, Xiang, and Gong 2017) are the best performance on three datasets. Recently, (Xian, Schiele, and Akata 2017; Xian et al. 2017) have reproduced SAE and other ZSL methods with ResNet-101 ImageNet features. In their reproducing results, the performance of SAE on AwA declines from 84.7% to 80.7%. The authors of (Xian et al. 2017) confirm that SAE reports per-image accuracy instead of MCA and improves GoogLeNet by adding Batch Normalization. However, on other two datasets, the performance of SAE declines more severely (from 61.4% to 33.4% on CUB and from 65.2% to 42.4% on SUN). For the Low-Rank model (Ding, Shao, and Fu 2017), MFMR (Xu et al. 2017) and SCoRe (Morgado and Vasconcelos 2017) that report results with

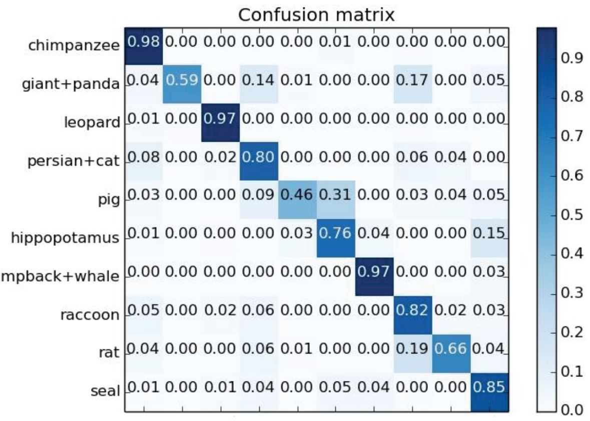

Figure 2: Confusion matrix of the classification accuracy on unseen categories for our method on AwA.

multiple CNN architectures, we only compare with their results using GoogLeNet. Because our DSSC model is learned from pre-trained GoogLeNet model.

From Table 2, we can observe that the DSSC model outperforms previous state-of-the-art methods. For AwA, our method attains 78.51%, which is slightly higher than the state-of-the-art result reported by SCoRe (78.3%). For CUB, the DSSC model has achieved impressive gains over the state-of-the-art SCoRe (from 58.4% to 65.99%). For SUN, it is noted that SynC extracts image features from CNN model pre-trained on MIT Places dataset and obtains the state-of-the-art result (62.7%). As we have mentioned in the Image Representations subsection, using MIT Places pre-trained CNN model on SUN violates the setting of ZSL. When SynC utilizes image features extracted from the CNN model pre-trained on ImageNet, the performance is 59.1% (reproducing result with ResNet-101 ImageNet features reported by (Xian et al. 2017)). In Table 2, all our experiments on SUN use ImageNet pre-trained CNN model besides DSSC-Places for fair comparison with SynC-Places. In both settings, our method clearly outperforms SynC (67.22% with CNN model pre-trained on MIT Places and 60.76% with CNN model pre-trained on ImageNet).

Moreover, we further visualize the zero-shot classification results of our DSSC method in term of the confusion matrix on AwA as shown in Figure 2. In the confusion matrix, the row represents the ground truth and the column corresponds to the predicted label. The diagonal position indicates the classification accuracy for each category. From the confusion matrix, we observe that our method obtains superb classification accuracy on *chimpanzee* (98%), *leopard* (97%) and *whale* (97%). For categories with lower accuracy, such as *pig* and *rat*, the errors mainly come from the semantically similar categories, for example *pig vs. hippopotamus* and *rat vs. raccon*.

Ablation Study

When utilizing SESc, we adaptively set the margin in triplet loss according to the Euclidean distance between each pair of the categories' semantic features. In this part, we arrange another experiment to verify the necessity of adopting such adaptable margin in SESc. We propose an alternative baseline, Deep-SESc-fixed-margin, which also adds SESc to the baseline ZSL model but the margin used in SESc is set to fixed value, *i.e.*, 1.0, for all categories. The comparisons of the three models (Baseline, Deep-SESc and Deep-SESc-fixed-margin) on AwA and CUB are shown in Table 4.

Table 3: Comparative results with self-adaptive margin (%).

Method	AwA	CUB
Baseline	75.19	58.26
Deep-SESc	**76.89**	**64.54**
Deep-SESc-fixed-margin	73.64	62.29

It is shown that the performance of Deep-SESc-fixed-margin is obvious lower than our Deep-SESc model on both datasets. On AwA, its performance is even lower than the baseline method. As we have mentioned before, setting a fixed margin for all categories may cause conflicts between the projected image feature $\mathbf{W}^T\phi(x)$ and the semantic representations $\psi(y)$ and it will lead to a worse embedding space and inferior performance.

Qualitative Results

In this section, we present qualitative results of the proposed DSSC model on CUB dataset and SUN dataset in Figure 3. We sample a subset of both datasets with 10 unseen classes and list the top-5 images which are classified into each category by the DSSC model. From these top retrieved images, it can be seen that our DSSC model is capable of capturing discriminative visual properties of each unseen class based on the semantic structural information. For example, the predicted *indigo bunting* images all share *blue wings* and *cone bill* attributes. We can also observe that the misclassified images have similar appearances to the images of predicted class. Such as the two misclassified *American crow* images (column 3 in CUB results) have the same *black wings* with the predicted *Groove billed Ani* category, which makes it difficult to distinguish between the two bird categories even for human beings. For SUN dataset, we observe another type of misclassification caused by inadequate attribute annotations. For example, in column 1 of SUN results, among the top-5 retrieved *athletic field* scene category, the three images actually belong to the *golf course* scene specific. We can easily distinguish the two scenes as the *golf course* images usually contain people playing golf. However, such distinction can not be described by designed attributes of SUN dataset. Addressing this issue may require incorporating the learning of undefined, latent but discriminative attributes from visual elements into the ZSL models.

Conclusion

In this paper, we focus on dealing with the defects of the two-step pipeline models of zero-shot learning. These models pay close attention to the formulation of the semantic

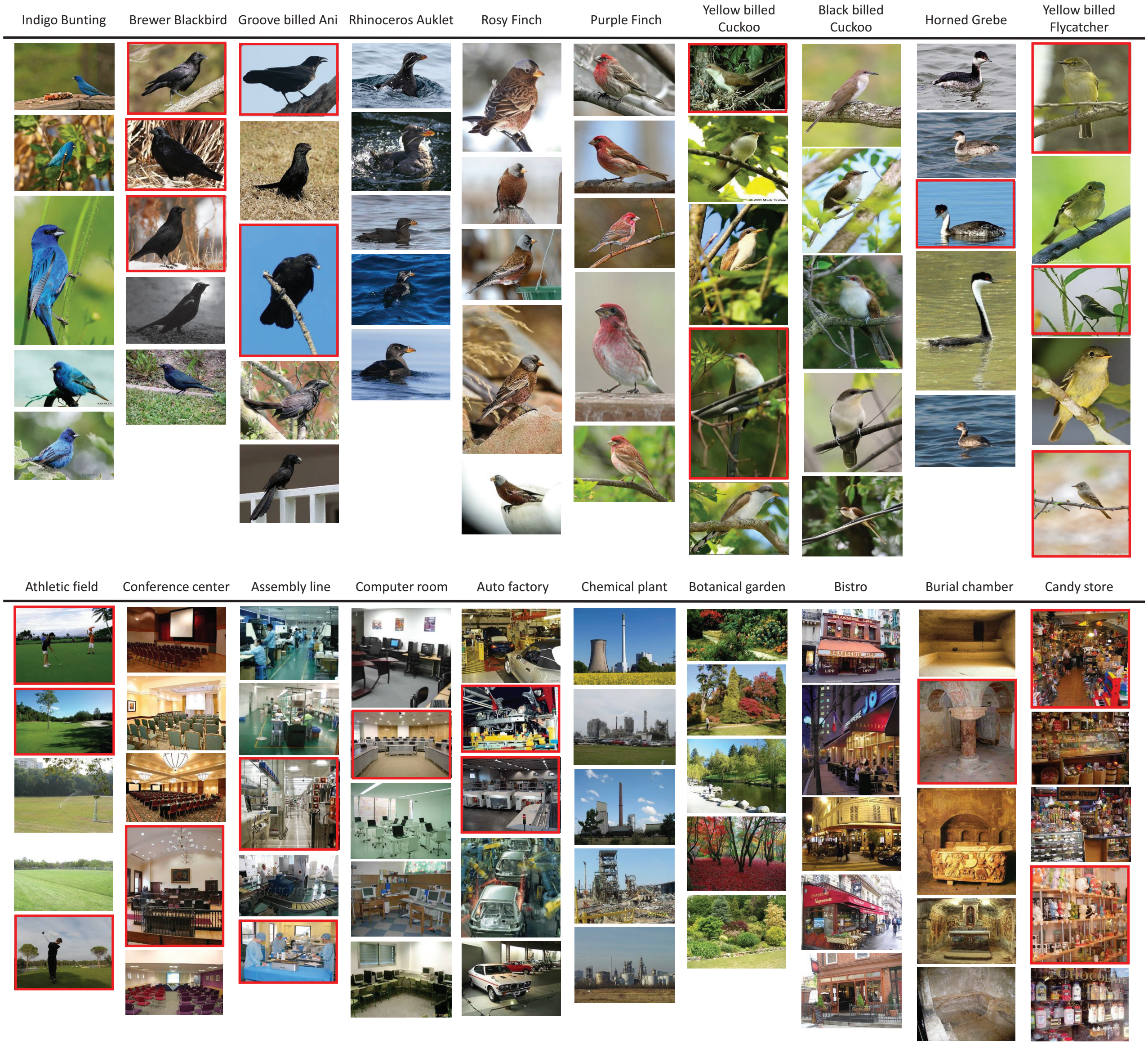

Figure 3: Qualitative results of the DSSC model on CUB (upper panel) and SUN (down panel). A subset with 10 categories of unseen class labels for both two datasets are listed. For each class, we visualize the top-5 images classified to it. Misclassified images are marked with red bounding boxes.

embedding space using fixed image features. The two-step pipeline tendency leads to the lack of semantic structural information for specific ZSL task. To address this issue, we propose an Image Feature Structure constraint to supervise image features' learning of zero-shot semantic information. Meanwhile, a Semantic Embedding Structure constraint is proposed to learn a generalized embedding space which aligns with the semantic space properly. The two constraints together with a softmax classification loss build up the Deep Semantic Structural Constraints model. The experimental results show that the model achieves the state-of-the-art performance on zero-shot learning task.

Acknowledgement

This work is funded by the National Natural Science Foundation of China (Grant 61403387, Grant 61673375 and Grant 61403383), the National Key Research and Development Program of China (Grant 2016YFB1001004 and Grant 2016YFB1001005) and the Projects of Chinese Academy of Sciences (Grant QYZDB-SSW-JSC006 and Grant 173211KYSB20160008).

References

Akata, Z.; Reed, S.; Walter, D.; Lee, H.; and Schiele, B. 2015. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2927–2936.

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2016. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 38(7):1425–1438.

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

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

Ding, Z.; Shao, M.; and Fu, Y. 2017. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*.

Frome, A.; Corrado, G. S.; Shlens, J.; Bengio, S.; Dean, J.; Mikolov, T.; et al. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*, 2121–2129.

Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive multi-view zero-shot learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 37(11):2332–2345.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Huang, C.; Loy, C. C.; and Tang, X. 2016. Local similarity-aware deep feature embedding. In *NIPS*, 1262–1270.

Karessli, N.; Akata, Z.; Bulling, A.; and Schiele, B. 2017. Gaze embeddings for zero-shot image classification. In *CVPR*.

Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. In *CVPR*.

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

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 36(3):453–465.

Morgado, P., and Vasconcelos, N. 2017. Semantically consistent regularization for zero-shot recognition. In *CVPR*.

Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*.

Patterson, G.; Xu, C.; Su, H.; and Hays, J. 2014. The sun attribute database: Beyond categories for deeper scene understanding. *International Journal of Computer Vision (IJCV)* 108(1-2):59–81.

Reed, S.; Akata, Z.; Lee, H.; and Schiele, B. 2016. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 49–58.

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

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)* 115(3):211–252.

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

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

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*.

Xian, Y.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-the good, the bad and the ugly. In *CVPR*.

Xu, X.; Shen, F.; Yang, Y.; Zhang, D.; Shen, H. T.; and Song, J. 2017. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *CVPR*.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *NIPS*, 487–495.