SAN: Sampling Adversarial Networks for Zero-Shot Learning *

Chenwei Tang[1], Yangzhu Kuang[2], Jiancheng Lv[1], and Jinglu Hu[2]

[1] Sichuan University, Chengdu 610065, P. R. China.
[2] Waseda University, Fukuoka 8080135, Japan.
lvjiancheng@scu.edu.cn

Abstract. In this paper, we propose a Sampling Adversarial Networks (SAN) framework to improve Zero-Shot Learning (ZSL) by mitigating the hubness and semantic gap problem. The SAN framework incorporates a sampling model and a discriminating model, and corresponds them to the minimax two-player game. Specifically, given the semantic embedding, the sampling model samples the visual features from the training set to approach the discriminator's decision boundary. Then, the discriminator distinguishes the matching visual-semantic pairs from the sampled data. On the one hand, by the measurement of the matching degree of visual-semantic pairs and the adversarial training way, the visual-semantic embedding built by the proposed SAN decreases the intra-class distance and increases the inter-class separation. Then, the reduction of universal neighbours in the visual-semantic embedding subspace alleviates the hubness problem. On the other, the sampled rather than directly generated visual features maintain the same manifold as the real data, mitigating the semantic gap problem. Experiments show that the sampler and discriminator of the SAN framework outperform state-of-the-art methods both in conventional and generalized ZSL settings.

Keywords: Zero-Shot Learning · Sampling Adversarial Networks · Hubness Problem · Semantic Gap

Image classification tasks have achieved great success due to the prosperous progress of deep learning [8]. However, most deep learning methods require labeling extensive training data, which is both labor-intensive and unscalable [19]. To tackle this limitation, Zero-Shot Learning (ZSL) is proposed to recognize new categories that have never seen during training, i.e., the categories in the training and test set are disjoint [20,22]. According to the categories included in the test, two ZSL settings are defined: conventional ZSL and Generalized ZSL (GZSL). Specifically, only the unseen classes are used to evaluation in the conventional ZSL setting. The GZSL provides a more practical point of view, where both seen and unseen categories are involved for testing.

* Thank Yangzhu Kuang for his contribution to this article. This paper is supported by the National Natural Science Fund for Distinguished Young Scholar under Grant No. 61625204 and the Key Program of National Science Foundation of China under Grant No. 61836006.

There are two fundamental challenges in ZSL: visual-semantic embedding [20] and domain adaption [5]. The knowledge can be transferred from the seen domain to the unseen domain by building a visual-semantic embedding. However, since the seen and unseen classes are different and potentially unrelated, the domain shift problem is triggered when the visual-semantic embedding is directly applied to the unseen data [5]. Thus, compared with the fully-supervised image classification tasks, the performance of ZSL is still far from perfect [3].

Most previous cross-modal embedding methods solved ZSL in two steps. First, project both visual features and semantic features to the embedding space [9]. Then, utilize nearest neighbour search in the embedding space to match the projection of visual or semantic feature vector against that of an unseen instance [29]. However, [17] proposed that there are many 'universal' neighbours, namely hubs, when performing nearest neighbour search in a high-dimensional space. They also showed that the hubness is an inherent property of data distributions in the high-dimensional vector space. Therefore, the cross-modal embedding methods always lead to the well-known hubness problem [3]. That is, a few unseen class prototypes will become the nearest neighbours of many hubs.

Recently, a new branch of methods target to ZSL by generative models [27]. They directly generate the unseen features from random noises which are conditioned by the semantic descriptions. With the generated unseen samples, zero-shot learning can be transformed to a supervised image classification task. However, since both the true and generated visual features contain intrinsic manifold structures, the manifold alignment is very challenge, especially in the high-dimension [18]. The semantic gap problem, i.e., the manifold of samples in the visual feature space is inconsistent with that of categories in the semantic space, often leads to model collapse, especially for the approaches [14] based on Generative Adversarial Networks (GAN) [6].

In this paper, we propose a novel Sampling Adversarial Networks (SAN) framework to improve ZSL by mitigating the hubness and semantic gap problem. The SAN framework proposes a new perspective for tackling ZSL by combining a sampling model and a discriminative model. The sampler aims at picking matching visual features given a semantic input. The discriminator focuses on measuring relevancy given a visual-semantic pair. These two models correspond to the minimax two-player game. On one hand, the discriminator guides the sampler to fit the underlying relevance distribution over visual features given the semantic presentation. On the other hand, the sampler tries to select visual features closing to the discriminator's decision boundary to confuse the discriminator. Our main contributes of this paper are summarized as follows:

- For the hubness problem, we construct a visual-semantic embedding by the adversarial training way. We utilize the discriminator to measure the matching degree of the visual-semantic pairs, rather than distinguish whether the visual features are real or fake. The proposed method considers the intra-class consistency and inter-class diversity, which alleviates the hubness problem.
- For the semantic gap problem, we propose to utilize the encoded attributes to sample visual features of seen classes, instead of directly generating visu-

al features. Then, the sampled visual features space is not affected by the semantic gap between attribute and visual feature space.
– Extensive experiments demonstrate that both the sampler and discriminator of the proposed SAN framework outperform state-of-the-art methods both in the conventional and generalized ZSL setting.

1 Related work

ZSL aims to classify images of new classes that have never been seen before, i.e., the training and test classes are disjoint. With the shared attributes annotated on class level, the ZSL is achieved by building the visual-semantic embedding to transfer the knowledge from seen classes to unseen classes [23]. GZSL is a more realistic setting, where the same information as ZSL is available at training phase, but both seen and unseen classes are classified during testing [26,2]. With the development of deep learning, many effective methods have been proposed to target to ZSL.

The cross-modal embedding models usually project either visual features or semantic features from one space to the other, or project both features into an intermediate space. Then, the compatibility function between visual and semantic features vectors is learned by using the ranking loss. ESZSL [18] learns a bilinear compatibility function between visual features, semantic features, and class labels with the square loss. LATEM [25] directly maps the visual feature to semantic space, and learns a bilinear compatibility function. SYNC [1] embeds both the visual and semantic features into another common space, and also learns a bilinear compatibility. SAE [9], following the Auto-Encoder, reconstructs the visual features in the semantic space. RethinkZSL [12] reformulates ZSL as a conditioned visual classification problem, i.e., classifying visual features based on the classifiers learned from the semantic descriptions.

The generative models reformulate ZSL as a standard fully-supervised classification task. GAZSL [30] takes noisy text descriptions about an unseen class as the input of generative model, and generates synthesized visual features for this class. f-CLSWGAN [28] synthesizes visual features conditioned on class-level semantic information, and pairs a Wasserstein GAN with a classification loss. LisGAN [24] trains a conditional Wasserstein GANs to directly generate the unseen features from random noises which are conditioned by the semantic descriptions.

Benefit from the synthesized missing features for unseen classes, the generative models achieve better results for unseen classes both in ZSL and GZSL. However, the manifold of samples in the visual feature space is inconsistent with that of categories in the semantic space. The semantic gap results in the disturbance of the generated unseen visual features to the original seen visual space. Therefore, the generative models seem kind of "confused" for the accuracy of seen classes in the GZSL setting. Building on ideas from these many previous works, we develop a simple and effective SAN framework incorporating sampler and discriminator, and corresponding them to the minimax two-player game.

2 Sampling Adversarial Networks

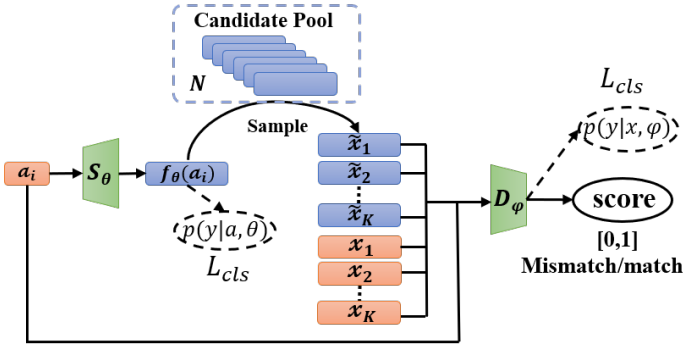


Fig. 1. The framework of the proposed SAN. a_i and x_i denote the attribute and the corresponding visual feature. $\tilde{x}_1$, $\tilde{x}_2$,..., $\tilde{x}_k$ denote the sampled visual features from all training visual instances $\mathcal{X}$. S_θ and D_φ are the sampler and discriminator, respectively.

Here we firstly introduce some notations and the problem definition. Let $\mathcal{S} = \{(x, y, a)|x \in \mathcal{X}_s, y \in \mathcal{Y}_s, a \in \mathcal{A}_s\}$ and $\mathcal{U} = \{(x, y, a)|x \in \mathcal{X}_u, y \in \mathcal{Y}_u, a \in \mathcal{A}_u\}$ where $\mathcal{S}$ and $\mathcal{U}$ denote training data of seen classes and testing data of unseen classes, respectively. $\mathcal{X}$ and $\mathcal{A}$ are the visual features and the semantic information in the form of attributes. $\mathcal{Y}_s$ and $\mathcal{Y}_u$ are the corresponding class labels. There is no overlap between seen and unseen classes, i.e., $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$. The goal of ZSL is to transfer the visual-semantic embedding learned in $\mathcal{S}$ to $\mathcal{U}$, and learn a classifier $f : \mathcal{X}_u \rightarrow \mathcal{Y}_u$. As for GZSL, we learn the classifier $f : \mathcal{X}_s, \mathcal{X}_u \rightarrow \mathcal{Y}_s \cup \mathcal{Y}_u$. Fig. 1 shows the overview of our method. We construct the sampling model S_θ and discriminative model D_φ as follow:

Sampling model. The sampler $\mathcal{S}_\theta : \mathcal{A} \rightarrow \mathcal{X}$ is a multi-layer perceptron. The sampling model $p_\theta(x|a)$ tries to approximate the true relevance distribution over visual features as much as possible, and confusing the discriminator's training next round. First, we utilize the sampler $\mathcal{S}_\theta$ to encode the input attribute a_i as an index vector $f_\theta(a_i)$ with the same dimension as the visual feature. Then, the sampling process of the sampling model is selecting K visual features $\{\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_K\}$ according to cosine similarity over discrete candidate pool. The candidate pool of each input semantic information is composed of N visual features randomly sampled from the whole trained seen visual features.

Discriminative model. The discriminative model aims at distinguishing the well-matched semantic-visual features from the sampled negative data. For each attribute, K visual features of the same categories as the attribute and K fake visual features sampled by the sampler are selected. Then we can get K pairs matching semantic-visual features as the positive samples and K pairs mismatching semantic-visual features as the negative samples. In order to calculate

the cosine similarity between semantic and visual information, we first encode the attribute a_i by the discriminator $\mathcal{D}_\varphi$. The discriminator $\mathcal{D}_\varphi : \mathcal{X}\times\mathcal{A} \to [0,1]$ is also a multi-layer perceptron, where the last layer is the cosine similarity between visual and semantic features. The $\mathcal{D}_\varphi$ is designed for estimating the probability of visual feature x being relevant to the semantic information a, i.e., they belong to the same category. The discriminative score $\mathcal{D}(x,a)$ can be defined as follow:

$$\mathcal{D}(x,a) = \cos(f_\varphi(a), x). \tag{1}$$

Objective function. Overall, as with the training procedure of GAN, the sampler $\mathcal{S}_\theta$ and discriminator $\mathcal{D}_\varphi$ of the proposed SAN framework play the two-player minimax game with the following objective function $V(\mathcal{S},\mathcal{D})$:

$$\min_\theta \max_\varphi V(\mathcal{S},\mathcal{D}) = \mathbb{E}_{x\sim p_{data}(x)}[\log \mathcal{D}(x,a)] + \mathbb{E}_{x\sim p_\theta(x|a)}[\log(1-\mathcal{D}(x,a))], \tag{2}$$

where a is the attribute vector, $x \sim p_{data}(x)$ is the visual feature of the same categories as the attribute a, and $x \sim p_\theta(x|a)$ is the visual feature sampled by the sampler $\mathcal{S}_\theta$. The discriminator $\mathcal{D}_\varphi$ tries to maximize the loss, and the sampler $\mathcal{S}_\theta$ tries to minimizes it. Following a replacing trick in [6], the optimal φ^* and θ^* can be obtained as follow:

$$\varphi^* = \arg\max_\varphi (\mathbb{E}_{x\sim p_{data}(x)}[\log \mathcal{D}(x,a)] + \mathbb{E}_{x\sim p_\theta(x|a)}[\log(1-\mathcal{D}(x,a))]). \tag{3}$$

$$\begin{aligned}\theta^* &= \arg\min_\theta (\mathbb{E}_{x\sim p_{data}(x)}[\log \mathcal{D}(x,a)] + \mathbb{E}_{x\sim p_\theta(x|a)}[\log(1-\mathcal{D}(x,a))]) \\ &= \arg\max_\theta \underbrace{\mathbb{E}_{x\sim p_\theta(x|a)}[-\log(1-\mathcal{D}(x,a))]}_{\text{denoted as } \mathcal{J}^{\mathcal{S}}(a)}.\end{aligned} \tag{4}$$

It is worth mentioning that the sampler is utilized to directly select known visual features from the candidate pool. Thus, the sampling of the visual features is discrete, which means that we cannot directly optimise the sampling model by gradient descent. Following [7], we use policy gradient [21] based on reinforcement learning to derive the gradient of $\mathcal{J}^{\mathcal{S}}(a)$. Given a query attribute a, the sampler is modeled as a reinforcement learning policy to sample a candidate visual feature x_n at the state, and is trained via policy gradients. The gradient of $\mathcal{J}^{\mathcal{S}}(a)$ can be derived as follows:

$$\begin{aligned}\nabla_\theta \mathcal{J}^{\mathcal{S}}(a) &= \nabla_\theta \mathbb{E}_{x\sim p_\theta(x|a)}[-\log(1-\mathcal{D}(x_n,a))] \\ &= \sum_{n=1}^{N} \nabla_\theta p_\theta(x_n|a)[-\log(1-\mathcal{D}(x,a))] \\ &= \sum_{n=1}^{N} p_\theta(x_n|a)\nabla_\theta \log p_\theta(x_n|a)[-\log(1-\mathcal{D}(x_n,a))] \\ &= \mathbb{E}_{x\sim p_\theta(x|a)}[\nabla_\theta \log p_\theta(x|a)(-\log(1-\mathcal{D}(x,a)))] \\ &\simeq \frac{1}{K}\sum_{k=1}^{K} \nabla_\theta \underbrace{\log p_\theta(x_k|a)}_{\text{the action}} \underbrace{(-\log(1-\mathcal{D}(x_k,a)))}_{\text{the reward}},\end{aligned} \tag{5}$$

where x_n denotes the n-th visual feature in the candidate pool, and x_k denotes the the k-th visual feature approximately sampled from the current version of sampler $p_\theta(x|a)$. Inspired by the reinforcement learning terminology, we use the term $\log p_\theta(x|a)$ to denote taking an action x in the environment a, and the term $(-\log(1-\mathcal{D}(x,a)))$ to denote the reward for the policy [21].

Specifically, for the policy gradient based on reinforcement learning, we first calculate the cosine similarity between the N visual features x_n in candidate pool and corresponding index vector $f_\theta(a)$ encoded by the sampler $\mathcal{S}_\theta$. Then, the probability $p_\theta(x_n|a)$ is obtained by softmax operation on the N cosine similarity values $cos(f_\theta(a), x_n)$. After that, we choose the probability value $p_\theta(x_k|a)$ and corresponding visual feature vector x_k of the top K probability value. Then, the log value of k probability value $p_\theta(x_k|a)$ is defined as the action value $log(p_\theta(x_k|a))$, and the value of $-log(1-\mathcal{D}(x_k,a))$ based on the discriminative score $cos(f_\varphi(a), x_k)$ is defined as the reward value. Finally, the average value of the product of the action value and reward value is the loss of the sampler.

Moreover, in order to reduce the expression differences of sampler and discriminator when they encode the attribute into the visual feature space, we introduce the classification loss $\mathcal{L}_{cls}$ based on the score function of discriminator. We apply the SoftMax classifier to both the real visual features and the sampled fake visual features. The classification loss $\mathcal{L}_{cls}$ is defined as follow:

$$\mathcal{L}_{cls} = -\sum_i y_i \ln p(y=i|x) = -\sum_i y_i \ln \frac{\exp(\mathcal{D}(x,a_i))}{\sum_{c_s}\exp(\mathcal{D}(x,a_{c_s}))}, \tag{6}$$

where c_s denotes the number of the seen categories. $p(y=i|x)$ represents the probability that the visual feature x belongs to category i. Specifically, we utilize $\mathcal{L}^{\mathcal{S}}_{cls}$ and $\mathcal{L}^{\mathcal{D}}_{cls}$ to represent the classification loss for real visual features and sampled fake visual features, respectively.

We take the classification losses as the regularizer for enforcing the sampler to select discriminative features, and promoting the disciminator to consider both inter-class and intra-class distance. Over full objective can be derived as follows:

$$\min_\theta \max_\varphi V(\mathcal{S},\mathcal{D}) + \alpha\mathcal{L}^{\mathcal{S}}_{cls} + \beta\mathcal{L}^{\mathcal{D}}_{cls}, \tag{7}$$

where α and β are the hyperparameters weighting the classifiers of sampler and discriminator, respectively.

Through multiple iterations of training, both sampler $\mathcal{S}$ and discriminator $\mathcal{D}$ can be used to classification. Given the attributes of all the unseen classes, i.e., $\mathcal{A}_u$, all the test index vector $f_\theta(\mathcal{A}_u)$ can be obtained by the sampler $\mathcal{S}$. Then, we calculate the cosine similarity between any test visual feature x and the index vector $f_\theta(\mathcal{A}_u)$. For the query image feature x, the label y of $a \in \mathcal{A}_u$ with the highest compatibility score is the classification result. As for the discriminator $\mathcal{D}$, we can directly get the compatibility scores between any query image x and all the test attributes $\mathcal{A}_u$ by Eq. 1. By finding the attribute with the highest cosine value, we can get the label of the query image.

Table 1. The details of five benchmark datasets.

Dataset	SUN	CUB	AWA1	AWA2	aPY
$\boldsymbol{\mathcal{A}}$	102	312	85	85	64
$\boldsymbol{\mathcal{N}}$	14340	11788	30475	37322	15339
$\boldsymbol{\mathcal{S}+\mathcal{U}}$	645+72	150+50	40+10	40+10	20+12
$\boldsymbol{\mathcal{N}_s}$	10320	7057	19832	23527	5932
$\boldsymbol{\mathcal{N}_u}$	1440	2967	5685	7913	7924
$\boldsymbol{\mathcal{N}_{s\to ts}}$	2580	1764	4958	5882	1483

3 Experiments

3.1 Experimental Setup

Dataset. We employ the most widely-used ZSL datasets for performance evaluation, that is, SUN Attribute Database (SUN) [16], Caltech-UCSDBirds 200-2011 (CUB) [15], Animals with Attributes 1 (AWA1) [10], Animals with Attributes 2 (AWA2) [26], Attribute Pascal and Yahoo (aPY) [4]. The GBU train/test split setting proposed in [26] is adopted to evaluate both the conventional ZSL setting and GZSL setting. Table 1 shows the details of five benchmark datasets. $\mathcal{A}$ denotes the dimension of attributes. $\mathcal{S}$ and $\mathcal{U}$ are the categories numbers of seen and unseen classes. N presents the number of images. N_s and N_u are the number of images of seen and unseen classes. Note that $N_{s\to ts}$ denotes the images' number of seen classes during test in the GZSL setting.

Evaluation metrics. To compare the performance with the existing method, we use the unified evaluation protocol, i.e., Mean Class Accuracy (MCA), proposed in [26]. MCA averages the correct predictions independently for each class before dividing the number of classes. In the GZSL setting, we adopt MCA_S on seen test classes, MCA_U on unseen test classes, and their harmonic mean $H = 2 * MCA_S * MCA_U/(MCA_S + MCA_U)$ as the evaluation metrics.

Implementation details. We use the 2048-dimensional top-layer pooling units with ReLU activation of the 101-layered ResNet as the visual features following the pre-trained method in [26]. The sampler and discriminator are Multi-Layer Perceptron (MLP) with ReLU activation. The dimension of the input layer of the MLP is the attribute's dimension of the corresponding dataset. For all datasets, the dimensions of the hidden layer and output layer are 1600 and 2048, respectively. We use Adam optimizer with learning rate 0.00005 to train the SAN framework. The pool size N of candidate pool is set as 100, and we sample $K = 3$ visual features from the candidate pool each time across all the datasets. We apply $\alpha = \beta = 1$ and develop our method based on PyTorch[3].

[3] The source code is provided at: https://github.com/TCvivi/Zero-Shot-Learning.

Table 2. Comparisons in conventional settings. The best results are in **bold**. SAN-D and SAN-S present the discriminator and sampler of SAN model, respectively.

Method	SUN(%)	CUB(%)	AWA1(%)	AWA2(%)	aPY(%)
ESZSL [18]	54.5	53.9	58.2	58.6	38.3
LATEM [25]	55.3	49.3	55.1	55.8	35.2
SYNC [1]	59.1	55.6	54.0	46.6	23.9
SAE [9]	40.3	33.3	53.0	54.1	8.3
RethinkZSL [12]	62.6	54.4	70.9	**71.1**	38.0
GAZSL [30]	61.3	55.8	68.2	70.2	41.1
f-CLSWAGN [28]	60.8	57.3	68.2	-	-
LisGAN [11]	61.7	**58.8**	70.6	-	43.1
SAN-D	**62.9**	57.0	**71.4**	69.7	**43.4**
SAN-S	62.7	55.7	70.3	68.1	40.3

3.2 Comparisons in Conventional Setting

We compare our method with the cross-modal embedding models and generative models in the conventional ZSL setting. Table 2 shows the experimental results. In the legacy challenge of zero-shot learning, both discriminator (SAN-D) and sampler (SAN-A) provide competitive performance, i.e., **62.9%** on SUN, **71.4%** on AWA1, **43.4%** on aPY. We analyze this striking improvements owing to the visual-semantic embedding space built by the adversarial training way, which decreases the intra-class distance and increases the inter-class separation.

Compare to the cross-modal embedding models, e.g., ESZSL [18], SYNC [1], which map the visual or semantic features to the fixed anchor points in the embedding subspace, the sampler of the proposed SAN framework directly samples the unseen features. Moreover, the discriminator distinguishes whether the sampled visual features match the input attribute, rather than whether the sampled visual features are true or fake. The sampler supplies the training unseen classes, and the visual-semantic embedding built by the adversarial training way decreases the intra-class distance and increases the inter-class separation. Then, the hubness problem is mitigated by reducing the universal neighbours surrounding the embedding vectors of unseen classes.

Compare to the generative models based on GAN, e.g., GAZSL [30], and f-CLSWGAN [28], we propose to utilize the encoded semantic features to sample true visual features of seen classes, instead of directly generating visual features. Thus, the sampled visual features space is not affected by the semantic gap between attribute and visual feature space. Experimental results show that the proposed method is better than the the generative models based on GAN. Moreover, both the sampler and discriminator of the proposed SAN are able to achieve good classification results.

3.3 Comparisons in Generalized Setting

Although the generative methods have much better generalization ability than the cross-modal methods on the conventional setting, the performances of these

Table 3. Comparisons in generalized settings. The best results are in **bold**. $U = MCA_U$, $S = MCA_S$, and H is the harmonic mean.

Method	SUN(%)			CUB(%)			AWA1(%)			AWA2(%)			aPY(%)		
	U	S	H	U	S	H	U	S	H	U	S	H	U	S	H
ESZSL [18]	11.0	27.9	15.8	12.6	63.8	21.0	6.0	75.6	12.1	5.9	77.8	11.0	2.4	70.1	4.6
LATEM [25]	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	11.5	77.3	20.0	0.1	73.0	0.2
SYNC [1]	7.9	**43.3**	13.4	11.5	**70.9**	19.8	8.9	**87.3**	16.2	10.0	**90.5**	18.0	7.4	66.3	13.3
SAE [9]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2	0.4	**80.9**	0.9
RethinkZSL [12]	36.3	42.8	39.3	47.4	47.6	47.5	**62.7**	77.0	**69.1**	56.4	81.4	66.7	26.5	74.0	39.0
GAZSL [30]	22.1	39.3	28.3	31.7	61.3	41.8	29.6	84.2	43.8	35.4	86.9	50.3	14.2	78.6	24.0
f-CLSWAGN [28]	42.6	36.6	39.4	43.7	57.7	49.7	57.9	61.4	59.6	-	-	-	-	-	-
LisGAN [11]	42.9	37.8	40.2	46.5	57.9	**51.6**	52.6	76.3	62.3	-	-	-	**34.3**	68.2	**45.7**
SAN-D	**45.6**	37.2	**41.0**	**48.6**	49.4	49.0	61.5	76.5	68.2	**57.6**	80.4	**67.1**	32.8	68.9	44.5
SAN-S	41.3	38.5	39.8	**48.6**	46.1	48.5	58.3	76.7	66.3	55.6	79.3	65.4	31.0	68.8	42.8

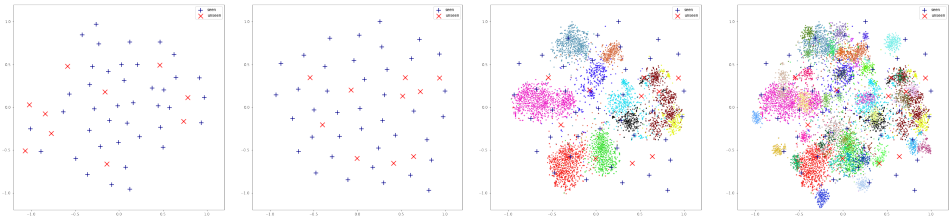
two methods both degrade dramatically on the generalized ZSL. Table 3 shows the experimental results in the generalized ZSL setting. An interesting observation is that the cross-model methods, e.g., ESZSL [18] and SYNC [1], perform well on seen test class (S), but work poorly on the unseen test classes (U). More interesting observation can be found that the generative methods, e.g., GAZSL [30] and f-CLSWGAN [28], perform well on the unseen classes in the GZSL setting, but their classification accuracy for the seen classes is worse than the cross-model methods. From the Table 3, we can see that the proposed SAN framework performs competitive on seen classes, unseen classes, as well as harmonic mean, i.e., **H = 41.0%** on SUN, **H = 67.1%** on AWA2.

The cross-modal methods transfer the knowledge learned from seen classes directly to unseen classes. Thus, the visual-semantic embedding subspace constructed by the seen classes can maintain a high supervised classification accuracy on the seen classes. Obviously, when the search space includes both seen and unseen classes, the images of unseen classes are easily divided into the seen training categories. The generative methods based on GAN build a more complete visual-semantic embedding subspace by generating the pseudo data of unseen class, so as to solve the problem of low recognition accuracy on unseen classes in the GZSL setting. However, due to the semantic gap between attribute and visual feature space, with the improvement of the subspace's ability to recognize unseen classes, it is inevitably sacrifice the classification ability on seen classes.

3.4 Model Analysis

Visualisation of the Learned Representation. To visually investigate the effectiveness of the proposed SAN framework, we adopt the t-SNE [13] approach to embed the representation of the visual features and attributes into a two-dimensional visualisation plane for the AWA1 dataset in Fig. 2. Compare to the distribution of original attributes in Fig. 2(a), the semantic representation of embedded attributes by the discriminator of our method of both seen (bule '+') and unseen (red '×') classes in Fig. 2(b) is more spatially dispersed, which proves that the proposed SAN framework considers the inter-class separation

for all classes. Fig. 2(c) shows the attribute features encoded by discriminator of all classes and visual features embedding of unseen classes. We can see that the discriminator is able to model the discrimination between samples from different semantic categories of unseen classes, and effectively separates the representations into several semantically clusters. It demonstrates that intra-class consistency and inter-class diversity are considered. Fig. 2(d) shows the attributes and visual features embedding of all test classes. Although the proposed method is able to separates the representations of all test classes into several clusters, the distributions of seen classes and unseen classes are too close or even overlapped. It's explains that all methods is fail to achieve perfect results in GZSL setting.



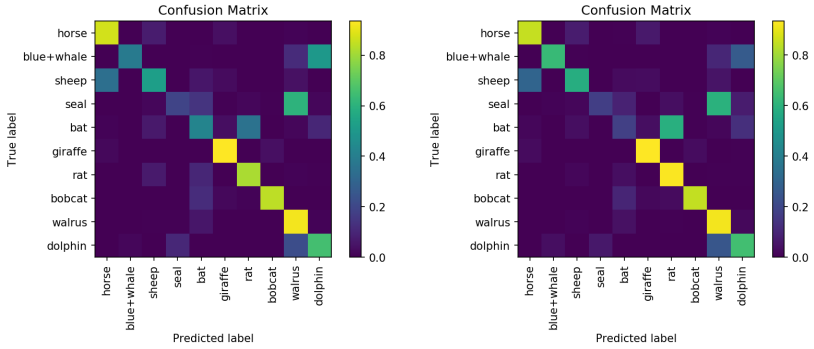
(a) (b) (c) (d)

Fig. 2. The visualization of the learned representation of AWA1 dataset.

Class-wise Accuracy. We use the confusion matrix to show the experimental result of ZSL in a more fine-grained scale. Fig. 3 shows the confusion matrix of both sampler $\mathcal{S}_\theta$ (Fig. 3(a)) and discriminator $\mathcal{D}_\varphi$ (Fig. 3(b)) in the proposed SAN framework on the evaluation of AWA2 dataset. As shown in Fig. 3, sampler $\mathcal{S}_\theta$ and discriminator $\mathcal{D}_\varphi$ of the proposed method generally have better accuracy on most of the test categories. For classes such as 'seal' and 'bat', the low recognition accuracy of the both sampler $\mathcal{S}_\theta$ and discriminator $\mathcal{D}_\varphi$ mainly due to the fact that no similar categories have been seen during training. Therefore, when the visual-semantic embedding constructed in the seen classes is transferred to the novel classes, the model tends to perform poorly on these categories.

4 Conclusion

In this paper, we propose the SAN framework to improve ZSL by mitigating the hubness and semantic gap problem. For the hubness problem, we construct a visual-semantic embedding by the adversarial training way. Moreover, we utilize the discriminator to measure the matching degree of the visual-semantic pairs, rather than distinguish whether the visual features are real or fake. The proposed method considers both the intra-class consistency and inter-class diversity, which alleviates the hubness problem. For the semantic gap problem, we propose to



(a) (b)

Fig. 3. The confusion matrixes on the evaluation of AWA2 dataset.

utilize the encoded attributes to sample visual features of seen classes, instead of directly generating visual features. Then, the sampled visual features space is not affected by the semantic gap between attribute and visual feature space. Extensive experiments on five most widely-used datasets demonstrate that both the sampler and discriminator of the proposed SAN framework outperform state-of-the-art methods both in the conventional and generalized ZSL setting. In the future, we intend to perform SAN for the transductive setting as well. By adding the visual features of the unseen classes to candidate pool of the sampler, the model can learn a more comprehensive visual-semantic embedding space.

References

1. Changpinyo, S., Chao, W., Gong, B., Sha, F.: Synthesized classifiers for zero-shot learning. In: CVPR. pp. 5327–5336 (2016)
2. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild pp. 52–68 (2016)
3. Dinu, G., Baroni, M.: Improving zero-shot learning by mitigating the hubness problem. In: ICLR. pp. 135–151 (2014)
4. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.A.: Describing objects by their attributes. In: CVPR. pp. 1778–1785 (2009)
5. Fu, Y., Hospedales, T.M., Tao, X., Fu, Z., Gong, S.: Transductive multi-view embedding for zero-shot recognition and annotation. In: ECCV. pp. 584–599 (2014)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M.e.a.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
7. Jun, W., Lantao, Y., Weinan, Z., Gong, Y., Yinghui, X., Benyou, W., Peng, Z., Dell, Z.: IRGAN: A minimax game for unifying generative and discriminative information retrieval models. In: SIGIR. pp. 515–524 (2017)
8. Kaiming, H., Xiangyu, Z., Shaoqing, R., Jian, S.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
9. Kodirov, E., Xiang, T., Gong, S.: Semantic autoencoder for zero-shot learning. In: CVPR. pp. 4447–4456 (2017)

10. Lampert, C.H., Nickisch, H., Harmeling, S.: Attribute-based classification for zero-shot visual object categorization. IEEE Trans. Pattern Anal. Mach. Intell. **36**(3), 453–465 (2014)
11. Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., Huang, Z.: Leveraging the invariant side of generative zero-shot learning. In: CVPR. pp. 7402–7411 (2019)
12. Li, K., Min, M.R., Fu, Y.: Rethinking zero-shot learning: A conditional visual classification perspective. In: ICCV. pp. 3582–3591 (2019)
13. van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11), 2579–2605 (2008)
14. Niu, L., Cai, J., Veeraraghavan, A.: Zero-shot learning via category-specific visual-semantic mapping. IEEE Trans. Image Processing **28**(2), 965–979 (2019)
15. P. Welinder, S. Branson, T.M.C.W.F.S.S.B., Perona, P.: Caltech-ucsd birds 200. In: Caltech, Tech. Rep. CNS-TR-2010-001 (2010)
16. Patterson, G., Hays, J.: SUN attribute database: Discovering, annotating, and recognizing scene attributes. In: CVPR. pp. 2751–2758 (2012)
17. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. JMLR **11**, 2487–2531 (2010)
18. Romera-Paredes, B., Torr, P.H.S.: An embarrassingly simple approach to zero-shot learning. In: ICML. pp. 2152–2161 (2015)
19. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015)
20. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.Y.: Zero-shot learning through cross-modal transfer. In: NIPS. pp. 935–943 (2013)
21. Sutton, R.S., McAllester, D.A., Singh, S.P., et al.: Policy gradient methods for reinforcement learning with function approximation. In: NIPS. pp. 1057–1063 (2000)
22. Tang, C., Lv, J., Chen, Y., Guo, J.: An angle-based method for measuring the semantic similarity between visual and textual features. Soft Computing **23**(12), 4041–4050 (2019)
23. Tang, C., Yang, X., Lv, J., He, Z.: Zero-shot learning by mutual information estimation and maximization. Knowledge-Based Systems (2020)
24. Williams, R.J.: Leveraging the invariant side of generative zero-shot learning. Machine Learning **8**, 229–256 (1992)
25. Xian, Y., Akata, Z., Sharma, G., Nguyen, Q.N., Hein, M., Schiele, B.: Latent embeddings for zero-shot classification. In: CVPR. pp. 69–77 (2016)
26. Xian, Y., H., L.C., Bernt, S., Zeynep, A.: Zero-shot learning - a comprehensive evaluation of the good, the bad and the ugly. IEEE Transactions on Pattern Analysis & Machine Intelligence **41**(9), 2251–2265 (2019)
27. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR. pp. 5542–5551 (2018)
28. Yongqin, X., Tobias, L., Bernt, S., Zeynep, A.: Feature generating networks for zero-shot learning. In: CVPR. pp. 5542–5551 (2018)
29. Zhang, L., Xiang, T., Gong, S.: Learning a deep embedding model for zero-shot learning. In: CVPR. pp. 3010–3019 (2017)
30. Zhu, Y., Elhoseiny, M., Liu, B., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: CVPR. pp. 1004–1013 (2018)