DOMAIN-AWARE ZERO-SHOT LEARNING WITH GENERATIVE MODELS

[1]*Yuan-Hao Mike Lee* (李元顥), [1]*Yu-Chiang Frank Wang* (王鈺強), [2]*Chiou-Shann Fuh* (傅楸善)

[1]Graduate Institute of Communication Engineering,
[2]Department of Computer Science and Information Engineering,
National Taiwan University, Taiwan, Republic of China
E-mail: r07942074@ntu.edu.tw

ABSTRACT

Zero-shot learning is the classification task where images belonging to a subset of classes are not present in the training dataset. For these unseen classes, the classification relies entirely on their semantic representations. This seemingly easy task for a human being is, however, extremely difficult for a machine to accomplish. Previous works have suggested learning a relationship between the visual and the semantic space, assuming that the two spaces share a similar distribution. Recently, generative adversarial networks (GAN) have been utilized to bridge the gap between the two spaces, generating a potentially infinite amount of visual samples from unseen classes to improve the classifier. Observing that the softmax results exhibit similar patterns for samples from within seen or unseen class categories, a domain-aware gating module has been further proposed which tries to tell whether a sample comes from seen or unseen classes, coupled with adaptive confidence smoothing which applies soft weighting between softmax outputs from classifiers specialized in different domains. In this paper, we combine the two methods mentioned above, and conduct extensive experiments that demonstrate the performance of the combined model.

Index Terms— Zero-Shot Learning, Domain Classification, Deep Generative Model, Cycle-Consistency, Deep Learning.

1. INTRODUCTION

In recent years, the performance of image classification and recognition has seen significant improvement with the use of deep convolutional neural networks. However, the strong performance relies heavily on the abundance of labeled training images, which may not be present when it comes to real-life applications. This may come as a surprise since the task is easy even for a human child, yet is extremely difficult for a machine to achieve comparable performance.

Few-shot learning describes the generalized problem where an imbalance of categories exists in the training dataset, and has since gained considerable attention. Furthermore, *zero-shot learning* (ZSL) is defined as the classification task where images to be classified from a subset of classes are completely absent in the training dataset, which is an extreme case of few-shot learning.

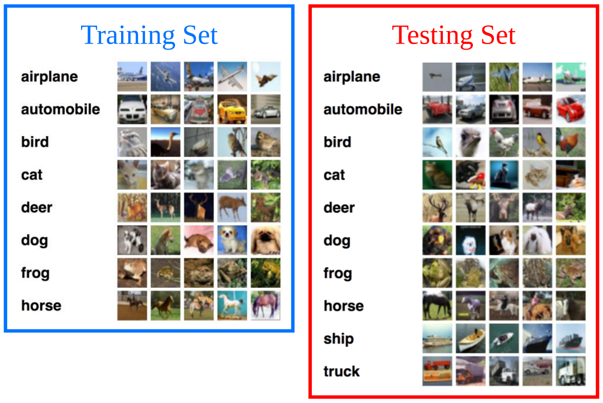

Fig. 1. A typical GZSL setting. Visual representations of unseen classes (*ship*, *truck*) are absent in the training dataset.

In the ZSL setting, the image categories of interest are split into two subsets: the *seen* and *unseen* classes. At training, visual representation is only available for seen classes, while side information (usually semantic representation) is provided for both seen and unseen classes. At testing, the samples to be classified only come from unseen classes [1, 2]. A more difficult setting — *generalized zero-shot learning* (GZSL) — is the case where testing samples may come from either seen or unseen classes. Classifiers trained under the GZSL setting tend to bias toward seen classes, which often leads to poor accuracy for samples from unseen classes.

With the assumption that the distribution of samples in the visual and semantic space are similar, recent works dealing with GZSL have suggested learning a transformation from semantic to visual space [1]. Testing can thus be performed by first transforming images from visual to semantic representations. Generative adversarial networks (GAN) [3] has also been utilized to bridge the gap between the two spaces, where a potentially infinite number of samples from unseen classes can be generated to facilitate the training of classifiers and alleviate the dataset bias [4, 5]. It has also been suggested that a cycle-consistency loss [6] can be added to regularize the synthetic output of such generative models [7].

Based on the observation that samples from seen and unseen classes generally exhibit distinctive patterns in the softmax activation output of the classifier, an additional module has been proposed to distinguish samples of seen classes from those of unseen classes [8]. The function of this domain-aware gating module is similar to that of a "domain classifier", where the two *domains* of interest here are the seen and unseen classes, respectively. With the posterior probability predicted by this model, one can build two final classifiers tasked with classification *within* each domain, then combine the two outputs in a soft manner (*e.g.*, weighting).

In this paper, we seek to establish a combined model that utilizes the two main concepts mentioned above, and conduct extensive experiments showing comparable results with other currently state-of-the-art methods.

2. RELATED WORKS

In this section, we give a brief summary for the literature review of some recent works related to our topic. Notably, Xian et al. has conducted an extensive survey on zero-shot learning papers in [1], which we recommend as a good starting point to research in this field of knowledge.

Generative Adversarial Network. Proposed by Goodfellow et al., GAN has the ability to learn a generative model which captures any data distribution, such as a set of images. The aim for GAN is to produce an output that is indiscriminable from the original data. Following [3], many improvements have been made to the original GAN structure: DCGAN [9] explores the benefits of training a GAN using deep convolutional neural networks; InfoGAN [10] adds a regularization that maximizes the mutual information between the latent variables and the generated output; [11] and [12] introduces class conditioning by feeding an additional class label to both the generator and discriminator. On the other hand, the theory of GAN has been studied extensively, with Wasserstein-GAN (WGAN) [13] utilizing a novel loss function that improves the stability of the training process, which optimizes an approximation of the Wasserstein distance.

Zero-Shot Learning. Abbreviated as ZSL, zero-shot learning is the setting where the images from a subset of classes are not present in the training data [1, 14, 15, 16, 17, 18]. Since the ZSL problem cannot be addressed in a supervised manner, most works suggests tackling the scenario by solving related sub-problems. Earlier methods include learning the unseen classes as a weighted combination of seen classes, with examples like SSE [2], CONSE [19] and SYNC [20]. Some others measure a compatibility score between each image and each class [21, 22, 23, 24, 25, 26, 27] where SJE [28], ALE [29] and DEVISE [30] are the most notable works, while ESZSL [31] and CMT [32] further improves the method by adding regularization and non-linear extensions, respectively. Finally, some works learn a classifier for intermediate representations (*e.g.*, attributes) [15, 17, 33].

Generalized Zero-Shot Learning. Few other works studied a more difficult yet more realistic setting called generalized zero-shot learning (GZSL) which, on the other hand, requires different approaches as the task often creates bias toward seen classes in the learned models. In this setting, images from both seen and unseen classes are present at test time [34]. Existing methods include the use of label embeddings [30], or learning latent features for both images and classes [35]. [8] and [36] further utilized an additional layer which predicts the probability of whether the input belongs to an unseen class, while [32] proposed a novel mechanism for detecting such information. Notably, f-CLSWGAN [4] utilized generative models to hallucinate visual representations from unseen classes. Using existing networks such as GAN or WGAN, a potentially infinite amount of visual samples can be generated to alleviate the problem of data imbalance. This work was later followed by cycle-WGAN [7], which added a cycle-consistency loss [6] to regularize the synthetic output of the generative models.

3. PROPOSED METHOD

3.1. Problem Formation

We begin this section by giving a formal definition of our generalized zero-shot learning problem.

First, the training dataset containing n samples is denoted as $\mathcal{D} = \{(\mathbf{x}, \mathbf{a}, y) | \mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}, y \in \mathcal{Y}_S\}$. Since our work uses pretrained deep networks to extract image features, we use $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^{d_x}$ to represent the d_x-dimensional latent visual representation instead of the image itself for simplicity. $\mathbf{a} \in \mathcal{A} \subset \mathbb{R}^{d_a}$ represents the d_a-dimensional semantic attributes, and $y \in \mathcal{Y}_S = \{y_1, \ldots, y_S\}$ is the class label where $\mathcal{Y}_S$ is the set containing S discrete seen classes. In addition, a disjoint set of U discrete unseen class labels $\mathcal{Y}_U = \{y_{S+1}, \ldots, y_{S+U}\}$ and their corresponding semantic attributes $\mathcal{U} = \{(\mathbf{a}, y) | \mathbf{a} \in \mathcal{A}, y \in \mathcal{Y}_U\}$ are also available during training. Note that $\mathcal{U}$ does not contain any visual information from unseen classes.

Next, the testing dataset containing m samples is denoted as $\mathcal{D}_t$, where $\mathcal{D}$ and $\mathcal{D}_t$ are mutually exclusive. In GZSL settings, $\mathcal{D}_t$ contains samples from both seen and unseen classes, that is, $\mathcal{D}_t = \{(\mathbf{x}, \mathbf{a}, y) | \mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}, y \in \mathcal{Y}_S \cup \mathcal{Y}_U\}$, while in ZSL settings $\mathcal{D}_t = \{(\mathbf{x}, \mathbf{a}, y) | \mathbf{x} \in \mathcal{X}, \mathbf{a} \in \mathcal{A}, y \in \mathcal{Y}_U\}$. We also note that $\mathcal{Y}_S \cap \mathcal{Y}_U = \varnothing$. Given the training set $\mathcal{D}$ and unseen class attribute information $\mathcal{U}$, our task of zero-

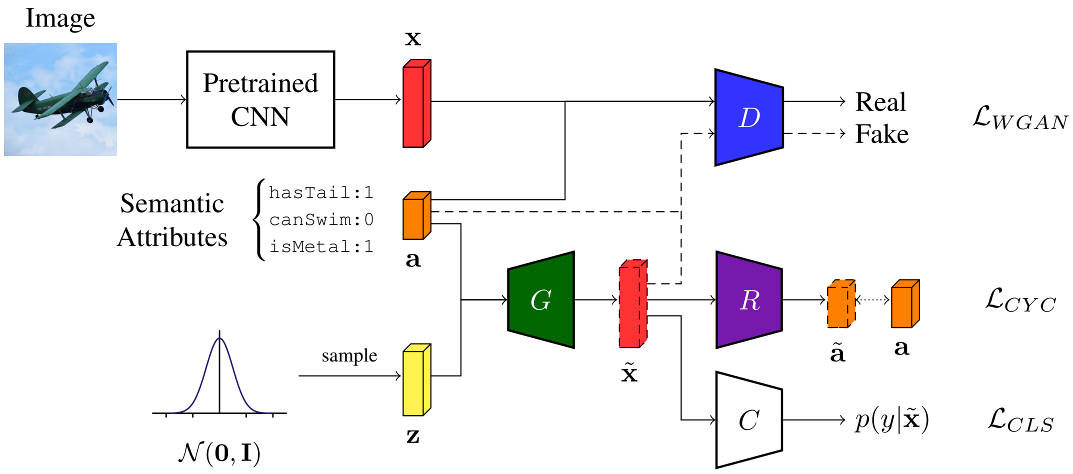

Fig. 2. An illustration of cycle-WGAN [7]. There are four main components in this architecture: 1) a generator G which takes as input semantic attributes and a Gaussian noise then outputs a synthetic visual representation, 2) a discriminator D which takes an input pair of visual representation and semantic attributes then tries to tell whether they come from the real or synthetic distribution, 3) a regressor R which generates back the original semantic attributes from visual representations, and 4) a classifier C which predicts class labels from visual representations.

shot classification is to train a classifier such that it achieves optimal classification accuracy on the testing set $\mathcal{D}_t$.

Given a sample $\mathbf{x}$, we denote the conditional probability that $\mathbf{x}$ belongs to class y as $p(y) = p(Y = y|\mathbf{x})$. Furthermore, the conditional distribution that $\mathbf{x}$ belongs to a seen class is denoted as $p(\mathcal{Y}_S) = p(Y \in \mathcal{Y}_S|\mathbf{x})$, or an unseen class as $p(\mathcal{Y}_U) = p(Y \in \mathcal{Y}_U|\mathbf{x})$. Also, define $p(y|\mathcal{Y}_S) = p(Y = y|Y \in \mathcal{Y}_S, \mathbf{x})$ and $p(y|\mathcal{Y}_U) = p(Y = y|Y \in \mathcal{Y}_U, \mathbf{x})$ as the conditional probabilities that $\mathbf{x}$ is of class y given it belongs to a seen or unseen class, respectively. Note that the symbol $\mathbf{x}$ may be omitted for better readability.

3.2. Generative Model

The method we adopt in this work is a direct extension of cycle-WGAN proposed by Felix et al. [7], which itself was based on a feature generating model (f-CLSWGAN) proposed by Xian et al. [4]. Therefore, we give a detailed introduction to the model of cycle-WGAN in this section.

To begin, the network consists of a conditional generator $G : \mathcal{Z} \times \mathcal{A} \rightarrow \mathcal{X}$. It takes as input a d_z-dimensional Gaussian noise vector $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^{d_z}$ and a semantic attribute vector $\mathbf{a} \in \mathcal{A}$, then generates a CNN feature vector $\tilde{\mathbf{x}} \in \mathcal{X}$ as its output. Since the generator is conditioned, $\tilde{\mathbf{x}}$ should correspond to the input semantic attribute $\mathbf{a}$ and the class y to which $\mathbf{a}$ belongs. A discriminator $D : \mathcal{X} \times \mathcal{A} \rightarrow \{0, 1\}$ then distinguishes whether an input pair of a CNN feature vector $\mathbf{x}$ and a semantic attribute vector $\mathbf{a}$ comes from the real dataset distribution or the fake one generated by G. To achieve this, G and D can be trained in an adversarial manner where G generates fake CNN feature vectors that tries to fool D, and conversely D tries to tell whether its input is real or fake. Following [4] and [7], we adopt Wasserstein-GAN (WGAN) as it has been proven to be one of the most stable GAN training methods [13]. The loss function is defined as

$$\begin{aligned} \mathcal{L}_{WGAN} = \mathbb{E}_{(\mathbf{x},\mathbf{a})}[D(\mathbf{x}, \mathbf{a})] - \mathbb{E}_{(\tilde{\mathbf{x}},\mathbf{a})}[D(\tilde{\mathbf{x}}, \mathbf{a})] \\ - \lambda \mathbb{E}_{(\hat{\mathbf{x}},\mathbf{a})}[(\|\nabla_{\hat{\mathbf{x}}} D(\hat{\mathbf{x}}, \mathbf{a})\|_2 - 1)^2], \end{aligned} \tag{1}$$

where $\tilde{\mathbf{x}} \sim G(\mathbf{a}, \mathbf{z})$ represents the synthetic samples generated by G, and $\hat{\mathbf{x}} \sim \alpha\mathbf{x} + (1-\alpha)\tilde{\mathbf{x}}$ with $\alpha \sim U(0, 1)$ sampled from a uniform distribution and λ as the penalty coefficient.

Moreover, since the loss function (1) does not ensure that the visual representations generated by G are discriminative in terms of different class labels, an additional negative log-likelihood classification loss is proposed as

$$\mathcal{L}_{CLS} = -\mathbb{E}_{(\tilde{\mathbf{x}},y)}[\log p(y|\tilde{\mathbf{x}})], \tag{2}$$

where $p(y|\tilde{\mathbf{x}})$ represents the predicted conditional probability that sample $\tilde{\mathbf{x}}$ belongs to its true class label y.

Lastly, the synthetic samples generated by G requires a further regularization as the aforementioned loss functions does not guarantee any constraint over the generation process from semantic attribute to visual feature space. Inspired by [6], a regressor $R : \mathcal{X} \rightarrow \mathcal{A}$ that encodes the generated visual features back to its original semantic attributes is proposed to enforce cycle-consistency over the generator G. Its loss function is thus defined as

$$\mathcal{L}_{CYC} = \mathbb{E}_{(\mathbf{a},\mathbf{z})}[\|\mathbf{a} - R(G(\mathbf{a}, \mathbf{z}))\|_2^2], \tag{3}$$

where $\mathbf{a} \sim \mathcal{A}$ are semantic attributes sampled from both seen and unseen classes, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are noise vectors sampled

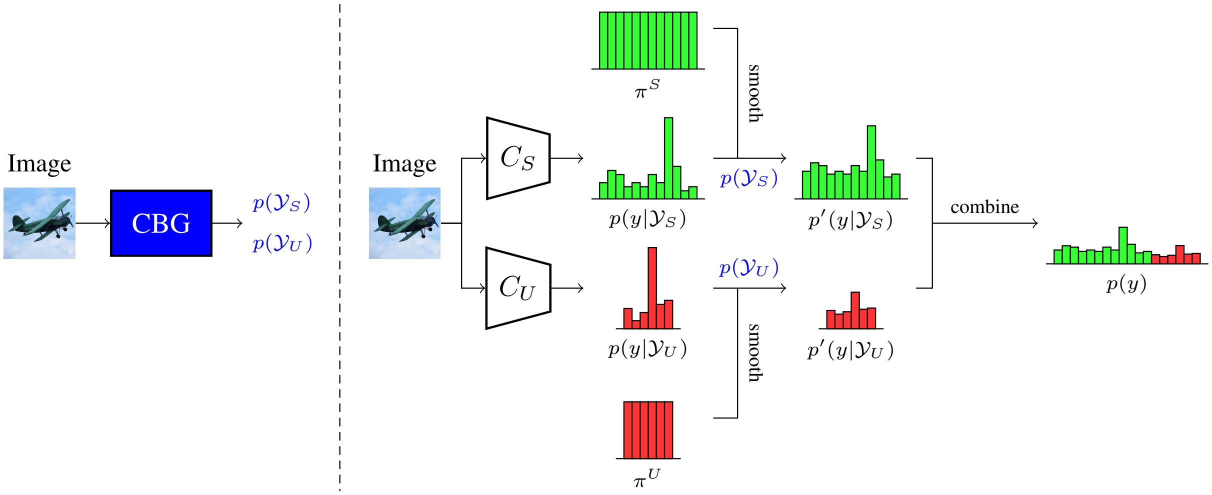

Fig. 3. An illustration of COSMO [8]. In this figure, the seen class predictions are marked in green and the unseen in red. During inference, a test image is first predicted by three classifiers: the domain classifier CBG, the seen class expert C_S, and the unseen class expert C_U. The softmax outputs of C_S and C_U are then smoothed with a uniform distribution based on the values of $p(\mathcal{Y}_S)$ and $p(\mathcal{Y}_U)$ predicted by the CBG, as described in (11). Lastly, the smoothed results are combined by the law of total probability as in (8). Note that for brevity, we omit the pretrained CNN in this figure. Best viewed in color.

from a normal Gaussian distribution, and $R(G(\mathbf{a}, \mathbf{z})) = \tilde{\mathbf{a}}$ is the recovered semantic attribute.

To train this feature generation model, we first pretrain a regressor R with the data from seen classes by minimizing a loss function similar to that defined in (3):

$$\mathcal{L}_{REG} = \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim\mathcal{D}}[\|\mathbf{a} - R(\mathbf{x})\|_2^2], \quad (4)$$

where $R^* = \arg\min_R \mathcal{L}_{REG}$ is henceforth fixed as the optimized regressor.

Next, we also pretrain a softmax classifier C with samples from seen classes by minimizing a loss function similar to that defined in (2):

$$\mathcal{L}_C = -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\log p(y|\mathbf{x})] \quad (5)$$

Note that this loss function is different from $\mathcal{L}_{CLS}$ as it is calculated over samples from seen classes instead of those generated by G. We denote $\tilde{C} = \arg\min_C \mathcal{L}_C$ as this pretrained classifier.

Finally, along with the pretrained R^* and $\tilde{C}$, the generator G and discriminator D in an adversarial manner, with the total loss function defined as

$$\mathcal{L}_t = \mathcal{L}_{WGAN} + \lambda_1\mathcal{L}_{CLS} + \lambda_2\mathcal{L}_{CYC}, \quad (6)$$

where $G^*, D^* = \arg\min_G \max_D \mathcal{L}_t$ and λ_1, λ_2 are weighting coefficients. Note that R^* and $\tilde{C}$ are both fixed when training the WGAN.

After obtaining G^* and D^*, we can train a final zero-shot softmax classifier C^* using a potentially infinite number of synthetic samples $\tilde{\mathbf{x}} \sim G^*(\mathbf{a}, \mathbf{z})$. Formally, this can be achieved by minimizing the loss function defined in (2), such that $C^* = \arg\min_C \mathcal{L}_{CLS}$. Note that G^* is fixed in this stage. Given an input CNN feature vector $\mathbf{x}$, the class label y^* which has the maximum softmax score produced by C^* is selected as the classification result, that is,

$$y^* = \arg\max_{y\in\mathcal{Y}} p(y|\mathbf{x}), \quad (7)$$

where $\mathcal{Y} = \mathcal{Y}_U$ for ZSL, and $\mathcal{Y} = \mathcal{Y}_S \cup \mathcal{Y}_U$ for GZSL settings, respectively.

3.3. Domain-Aware Gating Module

With our zero-shot classifier trained, we now follow COSMO [8], the probabilistic approach proposed by Atzmon et al. to introduce conditional smoothing over the softmax output from classifiers. In this subsection, we give a detailed description of our domain-aware gating module. To begin, we have

$$p(y) = p(y|\mathcal{Y}_S) \cdot p(\mathcal{Y}_S) + p(y|\mathcal{Y}_U) \cdot p(\mathcal{Y}_U) \quad (8)$$

by the law of total probability. This equation can be easily interpreted by separating a zero-shot classification model into different modules. First, $p(y|\mathcal{Y}_S)$ can be obtained by training a naive classifier over seen class samples (denoted as C_S). Then, $p(y|\mathcal{Y}_U)$ is simply the softmax output of the unseen classes from the zero-shot classifier we trained using cycle-WGAN (denoted as C_U), as described in Section 3.2. Lastly, $p(\mathcal{Y}_S)$ and $p(\mathcal{Y}_U)$ weight the two terms accordingly, which can be obtained using an additional "domain gating" classifier that gives the probability of a sample belonging to a seen or unseen class. Typically, $p(\mathcal{Y}_S) = 1 - p(\mathcal{Y}_U)$.

Toward this goal of exploiting the prediction results from two expert classifiers C_S and C_U, works like [32] uses hard gating by assigning $(p(\mathcal{Y}_S), p(\mathcal{Y}_U))$ to be either $(0, 1)$ or $(1, 0)$. In contrast, COSMO [8] combines the two in a soft manner. First, a confidence-based gating network (CBG) was proposed to estimate the values of $p(\mathcal{Y}_S)$ and $p(\mathcal{Y}_U)$. The CBG learns to discriminate the distinct softmax output patterns of samples *within* and *without* the expert's domain by "top-K pooling" [8].

When a probabilistic classifier observes an input that does not belong to any of its candidate classes, its output should be evenly low for all classes as these answers are equally wrong. However, even when this is the case, the classifiers still tend to assign most of the softmax probability to only one or two classes, introducing "false winners". A method similar to Laplace smoothing [37] has thus been utilized to overlay an additional prior distribution π over the original output. For both the seen and unseen distributions,

$$\begin{aligned} p^{\lambda}(y|\mathcal{Y}_S) &= (1 - \lambda^S) \cdot p(y|\mathcal{Y}_S) + \lambda^S \cdot \pi^S \\ p^{\lambda}(y|\mathcal{Y}_U) &= (1 - \lambda^U) \cdot p(y|\mathcal{Y}_U) + \lambda^U \cdot \pi^U, \end{aligned} \tag{9}$$

where λ^S, λ^U are the weighting coefficients. Since π^S and π^U do not depend on the input $\mathbf{x}$, it is intuitive to set them as uniform distributions with maximum entropy, that is,

$$\pi^S = \frac{1}{|\mathcal{Y}_S|}, \pi^U = \frac{1}{|\mathcal{Y}_U|}. \tag{10}$$

To make the additive prior distribution adaptive in accordance with the output of the CBG, we set $\lambda^S = 1 - p(\mathcal{Y}_S)$, such that the more likely $\mathbf{x}$ belongs to a seen class, the larger $p(\mathcal{Y}_S)$ is and the smaller λ^S (*i.e.*, smoothing effect) will be. Similarly, we set $\lambda^U = 1 - p(\mathcal{Y}_U)$. To sum up,

$$\begin{aligned} p'(y|\mathcal{Y}_S) &= p(\mathcal{Y}_S) \cdot p(y|\mathcal{Y}_S) + (1 - p(\mathcal{Y}_S)) \cdot \pi^S \\ &= p(y, \mathcal{Y}_S) + (1 - p(\mathcal{Y}_S)) \cdot \pi^S \\ p'(y|\mathcal{Y}_U) &= p(\mathcal{Y}_U) \cdot p(y|\mathcal{Y}_U) + (1 - p(\mathcal{Y}_U)) \cdot \pi^U \\ &= p(y, \mathcal{Y}_U) + (1 - p(\mathcal{Y}_U)) \cdot \pi^U, \end{aligned} \tag{11}$$

where $p(\mathcal{Y}_S)$ and $p(\mathcal{Y}_U)$ are outputs of the CBG, $p(y, \mathcal{Y}_S)$ and $p(y, \mathcal{Y}_U)$ are the softmax outputs of C_S and C_U, respectively. The final prediction can then be obtained using (8) after replacing $p(y|\mathcal{Y}_S)$ and $p(y|\mathcal{Y}_U)$ with $p'(y|\mathcal{Y}_S)$ and $p'(y|\mathcal{Y}_U)$ in (11).

4. EXPERIMENTS

In this section, we evaluate our combined model of cycle-WGAN [7] and COSMO [8] using four generalized zero-shot learning benchmarks, and compare it to recent state-of-the-art works.

4.1. Datasets

Following many previous works in the field of zero-shot learning, we test our model using four common benchmark datasets, namely AWA, CUB, SUN and FLO. In our experiments, the images are fed into the pretrained CNN as is, without any preprocessing. No data augmentation is performed.

AWA [15] (Animals with Attributes) is a coarse-grained dataset containing a total of 30,475 images from 50 different types of animals, where each animal class corresponds to a semantic vector of 85 attributes (*e.g.*, black, white, stripes, eats fish...). It is split into 27 training, 13 validation, and 10 testing classes.

CUB [43] (Caltech-UCSD Birds) is a fine-grained dataset containing a total of 11,788 images from 200 different types of birds, where each bird class corresponds to a semantic vector of 312 attributes (*e.g.*, bill shape, wing color...). It is split into 100 training, 50 validation, and 50 testing classes.

FLO [44] (Oxford-Flowers) is a fine-grained dataset containing a total of 8,189 images from 102 different types of flowers. While the dataset itself does not contain semantic annotations, we use the sentence embedding collected from [45] as class descriptions. It is split into 62 training, 20 validation, and 20 testing classes.

SUN [46] (Scene Categorization Benchmark) is a fine-grained dataset containing a total of 14,340 images from 717 different types of complex visual scenes, where each scene class corresponds to a semantic vector of 102 attributes (*e.g.*, driving, biking, hiking...). It is split into 580 training, 65 validation, and 72 testing classes.

4.2. Evaluation

We follow the unified evaluation protocol proposed by Xian et al. [1], which has since been adopted by many works and considered as a standard measurement of GZSL model performance. The protocol adopts average per-class top-1 accuracy for evaluation, that is, the class with the highest softmax score is selected as the predicted answer. In the ZSL setting, the accuracy of each unseen class is obtained independently and then averaged over all unseen classes, which we denote as $\mathbf{u}$. Apart from $\mathbf{u}$, we also compute the average per-class accuracy of seen classes in the GZSL setting, which we denote as $\mathbf{s}$. The final result is calculated as the harmonic mean of $\mathbf{s}$ and $\mathbf{u}$, that is, $H = (2 \cdot \mathbf{s} \cdot \mathbf{u})/(\mathbf{s} + \mathbf{u})$.

4.3. Implementation Details

The unified protocol adopts ResNet-101 [47] as the visual feature extractor, which is pretrained on ImageNet [48] with 1,000 classes. The activations of its top-layer pooling units are then selected as the visual representation of the images, which are of 2048 dimensions.

For the cycle-WGAN model, both the generator G and the discriminator D are multi-layer perceptrons (MLP) with

Datasets	AWA			CUB			FLO			SUN		
	$\mathbf{u}$	$\mathbf{s}$	H	$\mathbf{u}$	$\mathbf{s}$	H	$\mathbf{u}$	$\mathbf{s}$	H	$\mathbf{u}$	$\mathbf{s}$	H
Models w/o COSMO												
ESZSL [31]	6.6	75.6	12.1	12.6	63.8	21.0	11.4	56.8	19.0	11.0	27.9	15.8
SJE [28]	11.3	74.6	19.6	23.5	59.2	33.6	13.9	47.6	21.5	14.7	30.5	19.8
DEVISE [30]	13.4	68.7	22.4	23.8	53.0	32.8	9.9	44.2	16.2	16.9	27.4	20.9
SYNC [20]	8.9	87.3	16.2	11.5	70.9	19.8	—	—	—	7.9	43.3	13.4
ALE [29]	16.8	76.1	27.5	23.7	62.8	34.4	34.4	13.3	21.9	21.8	33.1	26.3
LATEM [21]	7.3	71.7	13.3	15.2	57.3	24.0	6.6	47.6	11.5	14.7	28.8	19.5
LAGO [38]	21.8	73.6	33.7	24.6	64.8	35.6	—	—	—	18.8	33.1	23.9
DEM [39]	32.8	84.7	47.3	19.6	57.9	29.2	—	—	—	—	—	—
ICINESS [40]	—	—	—	—	—	41.8	—	—	—	—	—	30.3
TRIPLE [41]	27.0	67.9	38.6	26.5	62.3	37.2	22.2	38.3	28.1	—	—	—
RN [42]	31.4	91.3	46.7	38.1	61.1	47.0	—	—	—	—	—	—
f-CLSWGAN [4]	59.7	61.4	59.6	43.7	57.7	49.7	59.0	73.8	65.6	42.6	36.6	39.4
cycle-WGAN [7]	59.6	63.4	59.8	47.9	59.3	53.0	61.6	69.2	65.2	47.2	33.8	39.4
Models w/ COSMO												
COSMO + f-CLSWGAN [8]	64.8	51.7	57.5	41.0	60.5	48.9	59.6	81.4	**68.8**	35.3	40.2	37.6
COSMO + LAGO [8]	52.8	80.0	63.6	44.4	57.8	50.2	—	—	—	44.9	37.7	**41.0**
COSMO + cycle-WGAN (ours)	61.2	67.7	**64.3**	47.5	60.4	**53.2**	62.4	70.5	66.2	44.1	35.8	39.5

Table 1. Comparison of classification results under the GZSL setting. $\mathbf{u}$ and $\mathbf{s}$ are the average per-class top-1 accuracies measured within the unseen and seen classes, respectively. H is the harmonic mean of $\mathbf{u}$ and $\mathbf{s}$. All results are shown in percentage.

a single hidden layer of 4,096 hidden nodes and a final activation by LeakyReLU [49]. The output of the generator is activated by ReLU [50], and has the same dimension as that of the ResNet-101. The output of the discriminator is of single dimension, and is not activated by ReLU. A linear transformation from the visual to the semantic space is adopted for the regressor R. All components are initialized with a truncated normal initialization with $(\mu, \sigma) = (0, 0.01)$ and zero bias. Following [7], we set $\lambda = 10$ in (1) and $\lambda_1 = \lambda_2 = 0.01$ in (6). After the generative model is trained, we create 300 synthetic visual representations from each class to train the final softmax classifier C_U.

For the COSMO model, we use the GZSL cross-validation splits proposed in [8]. A logistic regression classifier is adopted for the CBG, which is trained on the *Gating-Train* split. It takes as input the top-K pooled softmax outputs from both expert classifiers C_S and C_U, and predicts the values of $p(\mathcal{Y}_S)$ and $p(\mathcal{Y}_U)$ via a fully connected layer.

4.4. Results

Table 1 shows the experiment results on four datasets under the GZSL setting, and compares with several recent state-of-the-art methods. The values of $\mathbf{u}$ and $\mathbf{s}$ represent the average per-class top-1 accuracies measured within the unseen and seen classes, where H is the harmonic mean of the two values. The results of previous works are directly reported from the original papers, except for the models with COSMO which are measured and reported in [8].

Our combined method of COSMO and cycle-WGAN outperforms all listed methods in the AWA and CUB datasets, and achieves comparable performance with previous state-of-the-art works in the other two.

It is worth pointing out that, although the accuracies from our method did not surpass those from other works in some datasets, it still achieves better performances than those of the cycle-WGAN (w/o COSMO) across all four benchmarks, as is the case of f-CLSWGAN reported by [8]. This shows that the addition of the domain-aware gating module proposed by Atzmon et al. [8] is indeed advantageous in terms of generalized zero-shot classification tasks.

5. CONCLUSION

In this work, we adopt the methods of cycle-WGAN [7] and COSMO [8] into a combined model. The former trains an unseen class expert classifier using data hallucination with a generative model, while the latter first trains a naive seen class expert classifier, then adaptively smooths and combines the softmax outputs from two experts by the predicted confidence of whether the input belongs to a seen or unseen class.

As shown in our experiments, the combined model achieves state-of-the-art performance in two of the four common benchmarks, while being comparable in the other two. Notably, our model outperforms the original cycle-WGAN method across all four benchmarks, showing the benefit of an additional domain-aware gating module.

REFERENCES

[1] Yongqin Xian, Bernt Schiele, and Zeynep Akata, "Zero-shot learning-the good, the bad and the ugly," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4582–4591.

[2] Ziming Zhang and Venkatesh Saligrama, "Zero-shot learning via semantic similarity embedding," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4166–4174.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[4] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata, "Feature generating networks for zero-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5542–5551.

[5] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin, "Zero-shot learning via class-conditioned deep generative models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[6] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[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)*, 2018, pp. 21–37.

[8] Yuval Atzmon and Gal Chechik, "Adaptive confidence smoothing for generalized zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 11671–11680.

[9] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[10] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in neural information processing systems*, 2016, pp. 2172–2180.

[11] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[12] Augustus Odena, Christopher Olah, and Jonathon Shlens, "Conditional image synthesis with auxiliary classifier gans," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 2642–2651.

[13] Martin Arjovsky, Soumith Chintala, and Léon Bottou, "Wasserstein generative adversarial networks," in *International Conference on Machine Learning*, 2017, pp. 214–223.

[14] Sheng Huang, Mohamed Elhoseiny, Ahmed Elgammal, and Dan Yang, "Learning hypergraph-regularized attribute predictors," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 409–417.

[15] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 453–465, 2013.

[16] Hugo Larochelle, Dumitru Erhan, and Yoshua Bengio, "Zero-data learning of new tasks.," in *AAAI*, 2008, vol. 1, p. 3.

[17] Marcus Rohrbach, Michael Stark, and Bernt Schiele, "Evaluating knowledge transfer and zero-shot learning in a large-scale setting," in *CVPR 2011*. IEEE, 2011, pp. 1641–1648.

[18] Xiaodong Yu and Yiannis Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *European conference on computer vision*. Springer, 2010, pp. 127–140.

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

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

[22] Yanwei Fu and Leonid Sigal, "Semi-supervised vocabulary-informed learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5337–5346.

[23] Ruizhi Qiao, Lingqiao Liu, Chunhua Shen, and Anton Van Den Hengel, "Less is more: zero-shot learning from online textual documents with noise suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2249–2257.

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

[25] Maxime Bucher, Stéphane Herbin, and Frédéric Jurie, "Improving semantic embedding consistency by metric learning for zero-shot classification," in *European Conference on Computer Vision*. Springer, 2016, pp. 730–746.

[26] Zhenyong Fu, Tao Xiang, Elyor Kodirov, and Shaogang Gong, "Zero-shot object recognition by semantic manifold distance," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2635–2644.

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

[28] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2927–2936.

[29] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid, "Label-embedding for image classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 7, pp. 1425–1438, 2015.

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

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

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

[33] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele, "What helps where–and why? semantic relatedness for knowledge transfer," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 910–917.

[34] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult, "Toward open set recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1757–1772, July 2013.

[35] Hanwang Zhang, Xindi Shang, Wenzhuo Yang, Huan Xu, Huanbo Luan, and Tat-Seng Chua, "Online collaborative learning for open-vocabulary visual classifiers," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2809–2817.

[36] Abhijit Bendale and Terrance E Boult, "Towards open set deep networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1563–1572.

[37] Christopher Manning, Prabhakar Raghavan, and Hinrich Schütze, "Introduction to information retrieval," *Natural Language Engineering*, vol. 16, no. 1, pp. 100–103, 2010.

[38] Yuval Atzmon and Gal Chechik, "Probabilistic and-or attribute grouping for zero-shot learning," *arXiv preprint arXiv:1806.02664*, 2018.

[39] Li Zhang, Tao Xiang, and Shaogang Gong, "Learning a deep embedding model for zero-shot learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2021–2030.

[40] Yuchen Guo, Guiguang Ding, Jungong Han, Sicheng Zhao, and Bin Wang, "Implicit non-linear similarity scoring for recognizing unseen classes.," in *IJCAI*, 2018, pp. 4898–4904.

[41] Haofeng Zhang, Yang Long, Yu Guan, and Ling Shao, "Triple verification network for generalized zero-shot learning," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 506–517, 2018.

[42] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1199–1208.

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

[44] Maria-Elena Nilsback and Andrew Zisserman, “Automated flower classification over a large number of classes,” in *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 2008, pp. 722–729.

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

[46] Genevieve Patterson and James Hays, “Sun attribute database: Discovering, annotating, and recognizing scene attributes,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2751–2758.

[47] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[48] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[49] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng, “Rectifier nonlinearities improve neural network acoustic models,” in *Proc. icml*, 2013, vol. 30, p. 3.

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