Better Transferability with Attribute Attention for Generalized Zero-Shot Learning

Ruofan Guo [1] **Li Du** [1] **Yuan Du** [1] **Minzhe Liu** [1] **Xiaoliang Chen** [2]

Abstract

Generalized zero-shot learning (GZSL) models, trained on images from seen classes, are targeted to classify new images from both the seen and unseen classes. Previous works take the original semantic description as the input without further exploring its intrinsic information, which may hinder the knowledge transfer from the seen classes to the unseen classes. To facilitate knowledge transfer, a new approach is proposed in this paper to view the GZSL as a multilabel classification task. The labels, obtained from semantic attributes, are the similarities between seen and unseen classes. The proposed attribute attention module in our network is capable of extracting and enhancing the discriminative parts of image features. Moreover, two loss functions with pseudo data generation are implemented in the network to balance the transferability and the discriminability. Comprehensive evaluations show that the proposed approach outperforms the state-of-the-art methods in three datasets, APY, AWA2 and SUN, in the GZSL setting.

1. Introduction

Convolutional neural network (CNN) have achieved revolutionary successes in image classification. Most of the CNN based methods rely on abundant labeled training data, e.g. ImageNet (He et al., 2016), and can only recognize the objects whose categories are included in the training dataset. However, obtaining massive training data of all the categories is unfeasible. Therefore, zero-shot learning (ZSL) that aims to train a model to recognize the object that has no relevant training data becomes an attractive topic.

The ZSL model is targeted to recognize images belonging to novel categories that have not appeared in the training stage. In ZSL, the ability to recognize unseen category images is generally obtained by transferring knowledge learned from seen classes to unseen classes, which is accomplished with the help of semantic descriptions. The ZSL setting is based on the assumption that the input images only come from unseen classes in the test stage, which may not be feasible in some practical applications. Therefore, the model should have the ability to classify images from seen and unseen classes, which is called generalized zero-shot learning (GZSL). The search space of GZSL is expanded to all the seen and unseen classes. GZSL models typically suffer from the domain shift problem (Fu et al., 2014) due to the distributional differences between seen and unseen classes. As pointed out in (Chao et al., 2016), because of the imbalance between seen and unseen data, unseen class images tend to be recognized to belong to seen classes in GZSL setting. In this paper, we propose a new method for knowledge transfer that can alleviate the bias of classification and solve the domain shift problem.

In GZSL, the utilization of semantic information is one of the key points to achieve better knowledge transfer. The previous work (Xie et al., 2019) focuses on the transformation from image space to semantic space. The classification results are obtained by comparing predicted semantic attributes with ground truth attributes. The internal information in semantic attributes is not fully uncovered, which may hinder the further improvement of knowledge transfer. To obtain better transferability, we view GZSL as a multilabel classification task and construct labels from the internal information of semantic description.

The knowledge transfer in our work is accomplished by utilizing the similarities between different classes. Although two images belong to different classes, they still share some common features, and the common features are transferable between different classes, e.g. the similar tail of cow and buffalo, the similar color of cheetah and tiger. Therefore, we use continuous value ranging from 0 to 1 to describe the similarity of the input image to each class, instead of using 0 and 1 to indicate whether the input belongs to a category or not. In this way, the proposed network will learn the common features in different classes, which are then used

[1] School of Electronic Science and Engineering, Nanjing University, Nanjing, China [2] University of California, Irvine, CA, USA. Correspondence to: Ruofan Guo <rfguo@smail.nju.edu.cn>.

2020 ICML Workshop on Human in the Loop Learning (HILL 2020).

to recognize the images belonging to unseen classes.

Compared with class labels, semantic attributes have abundant information that is sufficient to guide the knowledge transfer. Semantic attributes are labeled by human experts, which are the most accurate and reliable knowledge compared to the inexplicit knowledge learned during training. Therefore, the ground truth similarities between different classes are constructed from semantic attributes.

The attention mechanism is combined in our work to extract the discriminative parts of features, aiming to enhance the discriminability and preserve the independence of recognition between seen and unseen classes. In the majority of previous works, the semantic descriptions are equally treated. However, the information in semantic descriptions can be divided into two categories, the discriminative part and the non-discriminative part, such as the strip and tail of zebra. Few previous research combined attention module in their work (Xie et al., 2019; Liu et al., 2019a). Due to the complicated network structure and lacking in effective supervision, training the attention module becomes troublesome. A new method of designing the attention module in ZSL is proposed in our work.

The combined approach, which involves knowledge transfer and attention mechanism, outperforms the state-of-the-art methods in GZSL setting. To sum up, our contributions are:

- We view ZSL as a multilabel classification task to obtain better knowledge transferability.
- We design an attention module that is easy to train to enhance the discriminability of image features.
- Experiments on four benchmark datasets show the superiority of the designed approach in GZSL settings.

2. Related Work

Zero-shot learning aims at training models to recognize images whose categories have never appeared in the training dataset. To achieve this, the model are trained to learn the relation between the training image data and their semantic descriptions. According to the methods that are used to fuse the visual information and semantic descriptions, current ZSL/GZSL approaches can be grouped into the following two types: 1. Semantic information is directly used to help the classifier to determine the category of the input image. 2. Semantic information is used to generate pseudo features of unseen classes, and the classifier is trained with the pseudo features to recognize unseen classes.

In the first category, images and semantic information are projected onto the same space to fuse their information. The methods proposed in (Fu et al., 2015; Ye & Guo, 2017) both project the visual features onto semantic space, and classification scores are measured by the distance between their semantic embeddings and semantic attributes. However, there is still a gap between high-dimensional visual features and low-dimensional semantic attributes, resulting in a loss of the visual information in the projection. To bridge the gap, a new latent visual embedding that is visually semantic is presented in (Zhu et al., 2019). Due to the lack of training data from unseen classes, this method is prone to classifying unseen class images into seen classes. Instead of only striving to extract semantic information from input images, the method in (Sung et al., 2018) projects both visual and semantic features onto a common intermediate space.

Generative models, e.g. generative adversarial network (GAN) and conditional variational autoencoder (CVAE), are widely used to generate pseudo features of unseen classes in ZSL. The generative model is trained on seen classes to acquire the distribution of visual representations based on semantic information. Provided with semantic information of unseen classes, the pseudo unseen samples can be generated by the generative model. The classifier is trained on real seen samples and pseudo unseen samples. (Xian et al., 2018) used conditional Wasserstein-GAN to synthesize CNN features, and the semantic information is used as the condition. GDAN (Huang et al., 2019), which combines GAN and dual learning, unifies both visual-semantic and semantic-visual generation and metric learning to further improve the performance. To improve the quality of generated features, (Li et al., 2019) proposed to define multiple soul samples for each class based on the investigation in the multi-view nature of different images.

Inspired by the learning process of humans, some research has been done to study the importance of different parts of semantic attributes in classifying an image. Attention mechanism (Xu et al., 2015) has been successfully applied to various fields. In ZSL, the attention module in (Xie et al., 2019) separates image features into different parts by the attention masks generated on image feature. Each part is responsible for predicting one part of the semantic attribute. However, there is no guidance on the attention mask generation, especially with the absence of semantic information. (Liu et al., 2019a) proposed an attribute attention framework to weight the attribute, aiming to tackle the semantic ambiguity problem. The proposed LFGAA network extracts attention masks from the outputs of the different layers of feature extractor. We find that these methods rely on a complicated network and lack in effective supervision with respect to the various inputs, making it intractable to transfer knowledge in the training stage.

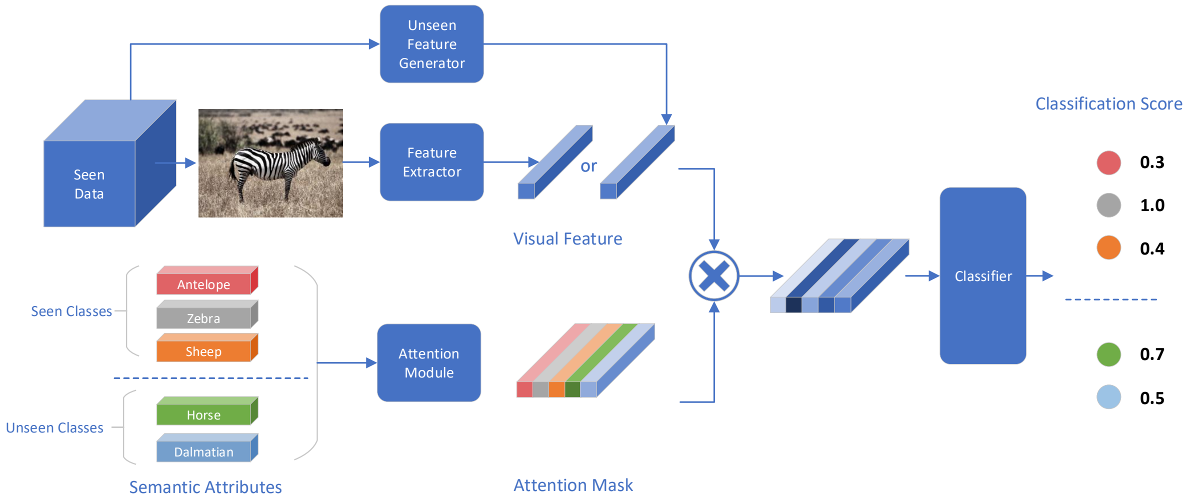

Figure 1. The framework of proposed network. The main body consists of a feature extractor, an attention module and a multilabel classifier. The image feature is weighted by the attention mask generated by the attention module to extract the discriminative parts, which is then taken as the input of the classifier to predict the similarity to each category.

3. Methodology

3.1. Problem Definition

We start by defining the zero-shot learning (ZSL) and generalized zero-shot learning (GZSL) problem settings. Given a training dataset $\mathcal{D}_s = \{(x_i^s,\ y_i^s)\}_{i=1}^{N^s}$, where $x_i^s \in \mathcal{X}^{\mathcal{S}}$ is the ith data sample, of totally N^s samples, with a corresponding seen class label as y_i^s, here $y_i^s \in \mathcal{Y}^{\mathcal{S}}$. $\mathcal{X}^{\mathcal{S}}$ and $\mathcal{Y}^{\mathcal{S}}$ is the image set and label set of seen classes. The test dataset is defined in the same way, $\mathcal{D}_u = \{(x_i^u,\ y_i^u)\}_{i=1}^{N^u}$. There is no overlap between training classes and test classes, $\mathcal{Y}^{\mathcal{S}} \cap \mathcal{Y}^{\mathcal{U}} = \emptyset$. Different from conventional image classification dataset, a semantic description dataset is supplied as complementary information in ZSL dataset. Semantic description for every class is available in that dataset: $\mathcal{A} = \{a_i^s\}_{i=1}^{C^s} \bigcup \{a_i^u\}_{i=1}^{C^u}$, where $a_i^{s/u} \in \mathbb{R}^n$ is the semantic vector corresponding to the ith seen/unseen class. The C^s and C^u are the number of seen and unseen classes, and n is the dimension of the semantic space. The goal of ZSL is to learn a classifier $f_{zsl} : \mathcal{X}^{\mathcal{U}} \to \mathcal{Y}^{\mathcal{U}}$, and the goal of GZSL is to learn $f_{gzsl} : \mathcal{X}^{\mathcal{S}} \cup \mathcal{X}^{\mathcal{U}} \to \mathcal{Y}^{\mathcal{S}} \cup \mathcal{Y}^{\mathcal{U}}$.

3.2. Model Overview

For better knowledge transferability, we view ZSL as a multilabel classification task. As shown in Figure 1, our model takes an image and all the attribute vectors as input and outputs classification scores on every class. The output is interpreted as the similarities of the input image to the target classes. The attention module generates attention masks based on the semantic attributes. There are three components in our model, an image feature extractor, an attribute attention module, and a multilabel classifier.

As suggested in (Xian et al., 2019), we use ResNet101 as our feature extractor, of which the parameters will not be updated during training. The input image feature of the classifier is denoted as $x_i \in \mathbb{R}^d$, where d is the dimension of the image feature. The attribute attention module takes the attributes of all classes as input, generating attention mask corresponding to every class. The image feature is weighted by the attention masks to enhance the discriminative part corresponding to every class. Then the classifier will predict the score for each category.

3.3. Attention Module

Other works (Xie et al., 2019; Liu et al., 2019a) also combine the attention mechanism into their work. They extract attention masks from image features and then weight the image features or attributes. Due to the complex network structure and diversity of the input images, it is hard to train the attention module without effective supervision.

The image feature, which is extracted by a well-trained network, has a fixed spatial structure. Each dimension of the feature vector corresponds to some certain information of the image. The semantic attribute vector is defined in the same way. Therefore, the attention module that generates attention masks based on the fixed attributes of all classes is easier to train. During training, the structure transformation from semantic attributes to visual features is learned by the attention module. Then the semantic attributes are projected onto image feature space as attention masks, which are expected to augment the discriminative part and suppress the non-discriminative part.

We use multilayer perceptrons (MLP) to construct the attention module, denoted as $g(\cdot)$. We experimented with two output functions, sigmoid function and ReLU. The sigmoid function produces weights with values ranging from 0 to 1, indicating the importance of each dimension. However, the properties of the sigmoid function severely hinder the training, resulting in a slower training speed and relatively bad performance. Therefore, we resort to ReLU. We conducted experiments to test the effects of ReLU layer, and the experiments show that the weights produced by the ReLU layer range from 0 to 1, which are suitable for attention mask. The output attention mask of jth class attribute is $g(a_j) \in \mathbb{R}^d$ where d is the dimension of the image feature vector. Let $f(\cdot)$ denote the feature extractor, and the weighted image features are obtained by:

$$z_{i,j} = f\left(x_i\right) \otimes g(a_j) \tag{1}$$

Where $\otimes$ is element-wise product operation, $z_{i,j}$ is the image feature weighted by the attention mask of class j. Then the weighted image features of all classes are concatenated together and fed to the classifier.

3.4. Multilabel Classifier

Different classes in a ZSL dataset share some common characteristics. Like humans recognizing an unseen category, the ZSL model transfer knowledge from the seen classes to the unseen classes to acquire the unseen data distribution. In the ZSL dataset, though two objects belong to different categories, they still resemble each other to some degrees. Therefore, we view GZSL as a multilabel classification task. The output scores represent how similar the input is to a certain category.

Let $h(\cdot)$ denote the classifier, which is implemented as a MLP with a sigmoid output function. The classification scores, regarded as similarities, of image x_i to classes j is formulated as:

$$o_{i,j} = h\left(z_{i,j}\right) \tag{2}$$

Instead of taking one feature as input and predicting its classification score on all classes, the multilabel classifier takes $C^s + C^u$ weighted features as input and predicts, in sequence, the probability that the original feature falls into each category. In this way, the criteria of classifying the input image are independent of which category the input belongs to, but only related to the distribution of weighted feature.

3.5. Knowledge Transfer

We did not focus our efforts on the design of the network structure, because the image feature extracted by a well-trained ResNet101 contains enough information for us to imply our algorithm. Instead, we mainly focus on how to utilize semantic information to achieve better knowledge transfer.

3.5.1. Class Similarity

The similarities between different classes are used as labels to guide the training. Semantic attributes, labeled by the human experts, contain abundant information of every classes. Therefore, the class similarities in our approach are obtained from semantic attributes. Inspired by (Jiang et al., 2019), we use the source classes attributes to reconstruct the target class attribute, and the reconstruction coefficients are interpreted as the similarity of the target class to the source classes. The reconstruction is accomplished by Ridge Regression, of which the object function is:

$$\theta_t = \underset{\theta_t}{\arg\min}\, ||a_t - \sum_{s=0}^{n^s} \theta_{ts} \cdot a_s||_2^2 + \beta||\theta_t||_2 \tag{3}$$

Where a_s and a_t are the source and target class attributes, β is a regularization parameter. The similarity is a normalized version of θ:

$$s_{t,s} = \frac{\max(0, \theta_{ts})}{\sum_{i=1}^{n^s} \max(0, \theta_{ti})} \tag{4}$$

Different from (Jiang et al., 2019), we construct two set of similarities, which are similarities $S_{ss} \in \mathbb{R}^{C^s, C^s}$ among seen classes and similarities $S_{su} \in \mathbb{R}^{C^s, C^u}$ between seen and unseen classes. Note that, when doing Ridge Regression to construct S_{ss}, the a_t is also in the set of $\{a_s\}$. Therefore, the a_t is removed and the similarity to itself is set to 1 after normalization.

3.5.2. Loss Function for Knowledge Transfer

With the supervision of similarities between seen and unseen classes, the network learns the unseen feature distribution by transferring knowledge of seen classes. Given an image feature as the input to the classifier, the classifier is expected to produce 1 for its own class. For other class, the output score should be close to the similarity between them. When taking an image feature belonging to unseen classes as input in the test stage, the classifier will assign the highest value to its own class y_u based on the knowledge learned from the similar seen classes. We treat the similarity as the probability of x_i belongs to class y_j,

$$P\left(y_j|x_i\right) = s_{y_i, y_j} \tag{5}$$

Let $o_{i,j}$ denote the predicted similarity of input image x_i to target class $y_j \in Y^s \cup Y^u$. Then the loss function on seen classes is formulated as:

$$\mathcal{L}_{Ts} = -\sum_{i=1}^{N^s} \sum_{j=1}^{C^s} s_{y_i, y_j} \log o_{i,j} + \left(1 - s_{y_i, y_j}\right) \log\left(1 - o_{i,j}\right)$$

The loss function on unseen class formulated as:

$$\mathcal{L}_{Tu}=-\sum_{i=1}^{N^s}\sum_{j=C^s}^{C^s+C^u} s_{y_i,y_j}\log o_{i,j}+\left(1-s_{y_i,y_j}\right)\log\left(1-o_{i,j}\right)$$

Where $s_{y_i,y_j} \in S_{ss} \cup S_{su}$ is the ground truth similarity. In further experiments, we find that a hyperparameter to balance the transfer loss on seen and unseen classes leads to a better result. The loss for knowledge transfer is the weighted sum of them:

$$\mathcal{L}_T = \mathcal{L}_{Ts} + \alpha \mathcal{L}_{Tu} \tag{6}$$

3.5.3. Loss Function for Discriminative Property

When doing recognition, the predicted category is that with the highest similarity value. Since we view it as a multilabel classification task, the predicted similarity of other classes will weaken the discrimination ability. Therefore, we add an auxiliary loss to improve the discriminative property by maximizing the score of their own class. The auxiliary loss is formulated as a cross-entropy loss:

$$\begin{aligned}\mathcal{L}_{A_s} = -\sum_{i=1}^{N^s}\sum_{j=1}^{C^s} & \mathbb{1}(y_i, j)\cdot \log o_{i,j} \\ & + (1-\mathbb{1}(y_i,j))\log(1-o_{i,j})\end{aligned} \tag{7}$$

Where $\mathbb{1}\left(y_i, j\right)$ is an indicator function that returns 1 if $y_i = j$ or 0 otherwise. Since the unseen data is unavailable during training, we can not enhance the discriminative property on unseen classes by simply duplicating the $\mathcal{L}_{A_s}$ However, we do need a counterpart loss on unseen classes because the loss inconsistency between seen and unseen classes will confuse the network. Therefore, we use an intuitive method to generate pseudo unseen data based on the concept mentioned in the class similarity.

Objects belong to different categories share some common characteristics. So the unseen class image features can be generated by synthesizing the seen class image features. We reconstruct unseen class attributes using seen class attributes by Eq. 3. The coefficients are used as weights to synthesize pseudo features, which is formulated as:

$$x_j = \sum_{i=1}^{C^s} \theta_{j,i}\cdot Random\left(\{x_i\}\right) \tag{8}$$

Where x_j is the pseudo image feature belongs to unseen class $j \in Y^u$, and $Random\left(\{x_i\}\right)$ is the image feature chosen randomly from the seen features belonging to class $y_i \in Y^s$. The generation will be repeated many times until there are proportionable pseudo features. Then the auxiliary loss on unseen data is formulated as:

$$\begin{aligned}\mathcal{L}_{A_u} = -\sum_{i=1}^{N^p}\sum_{j=1}^{C^u+C^s} & \mathbb{1}(y_i, j)\cdot \log o_{i,j} \\ & + (1-\mathbb{1}(y_i,j))\log(1-o_{i,j})\end{aligned} \tag{9}$$

Taking into account the scores of pseudo features on seen classes, $\mathcal{L}_{A_u}$ will not only increase the discriminative property on unseen classes, but also suppresses the scores of input features belonging to seen classes. Previous works (Chao et al., 2016; Atzmon & Chechik, 2019) present that reducing the scores of seen classes can improve the recognition result in GZSL setting. The loss for discriminative property is the weighted sum of Eq.7 and Eq.9:

$$\mathcal{L}_A = \mathcal{L}_{A_s} + \beta\mathcal{L}_{A_u} \tag{10}$$

The magnitudes of the four losses are related to the number of categories involved in the calculation, and the degree of impacts they have on the network varies. Therefore, the final object function is the weighted sum of the four loss function, which is formulated as:

$$\mathcal{L} = \mathcal{L}_{As} + \alpha\mathcal{L}_{Au} + \beta\mathcal{L}_{Ts} + \gamma\mathcal{L}_{Tu} \tag{11}$$

4. Experiments

4.1. Datasets and Settings

Following the instruction in (Xian et al., 2019), we conduct zero-shot and generalized zero-shot recognition experiments on four widely used ZSL datasets: APY(Farhadi et al., 2009), AWA2(Xian et al., 2019), CUB(Wah et al., 2011), SUN(Patterson et al., 2014). Specifically, APY is a small-scale coarse-grained dataset, with 64D attributes and a total of 15,339 images, and the seen/unseen splits are 20/12. AWA2 is an extension of AWA1(Lampert et al., 2009). AWA2 includes 37,322 images of animals from 50 classes, and the seen/unseen splits are 40/10. The semantic attribute provided in AWA2 is an 85D vector associated with each class. CUB is a fine-grained and medium-scale dataset, which contains 200 different types of birds annotated with 312attributes and 11,788 images in total. SUN is a scene image dataset, consisting of 14,340 images from 717 categories, with splits of 645/72 for seen/unseen classes. A 102D continuous semantic vector is supplied for each class.

4.2. Implementation Details and Parameters

The survey (Xian et al., 2019) reproduced a variety of previous methods using the 2048D feature extracted from ResNet101(He et al., 2016). And following (Xian et al., 2019), many methods proposed recently use the 2048D features as their input. For fair comparison with those published approaches, we take ResNet101 as our feature extractor.

Table 1. Generalized zero-shot recognition results on APY, AWA2, CUB and SUN, in %. H is the harmonic mean of the accuracies on seen and unseen classes. '/' denotes that the result is not reported.

Method	APY			AWA2			CUB			SUN		
	SEEN	UNSEEN	H	SEEN	UNSEEN	H	SEEN	UNSEEN	H	SEEN	UNSEEN	H
(Sung et al., 2018)	/	/	/	93.4	30.0	45.3	61.4	38.1	47.0	/	/	/
(Verma et al., 2018)	/	/	/	68.1	58.3	62.8	53.3	41.5	46.7	30.5	40.9	34.9
(Zhu et al., 2018)	78.6	14.2	24.0	84.2	35.4	50.3	61.3	31.7	41.8	39.3	22.1	28.3
(Liu et al., 2018)	75.0	14.2	23.9	/	/	/	60.7	28.4	38.7	37.0	22.5	30.2
(Liu et al., 2019b)	/	/	/	93.4	27.0	41.9	80.9	36.2	50.0	40.0	18.5	25.3
(Cacheux et al., 2019)	/	/	/	83.2	48.5	61.3	55.8	52.3	**53.0**	30.4	47.9	36.8
(Xian et al., 2018)	/	/	/	/	/	/	57.7	43.7	49.7	36.6	42.6	**39.4**
(Jiang et al., 2019)	64.0	24.1	**35.1**	65.8	61.2	**63.4**	52.6	52.0	52.3	37.3	31.2	34.0
ours	57.7	32.5	**41.5**	70.6	63.2	**66.7**	52.1	52.8	**52.5**	34.1	47.5	**39.7**

Table 2. Zero-shot recognition results on APY, AWA2, CUB, and SUN (in %), '/' denotes that the result is not reported.

Method	APY	AWA2	CUB	SUN
(Sung et al., 2018)	/	64.2	55.6	/
(Xian et al., 2018)	/	/	57.3	60.8
(Zhu et al., 2018)	41.1	70.2	55.8	61.3
(Liu et al., 2018)	43.6	/	56.2	61.8
(Liu et al., 2019b)	/	68.1	**67.6**	61.5
(Cacheux et al., 2019)	/	67.9	63.8	**63.5**
(Li et al., 2019)	43.1	70.6	58.8	61.7
(Jiang et al., 2019)	38.9	71.2	59.5	61.8
ours	**44.2**	**73.4**	60.9	61.3

The attention module is implemented as a multilayer perceptron (MLP) with one hidden layer followed by a ReLU output layer. The outputs, ranging from 0 to 1, are interpreted as attention masks. The multilabel classifier is implemented as a two layers fully connected (FC) neural network followed by a sigmoid output function.

As we discussed in Section 4.1, each of the four datasets has its own characteristics, which affects the discriminability of image features and the transferability between seen and unseen classes. Therefore, to better balance the discriminative property and the transferability, three hyperparameters α, β, γ are set differently in four datasets.

4.3. Evaluations in ZSL Settings

In the ZSL setting, the predicted class of one input image is the class with the highest classification score among all unseen classes. We formulated it as:

$$C_{zsl}\left(x_i\right) = \arg\max_{j} \left\{o_{i,j}\right\},\ j \in Y^u \tag{12}$$

We compare the proposed method against current state-of-the-art approaches on four datasets. As suggested in (Xian et al., 2019), average class accuracy (ACA) is adopted as the evaluation metric. Table. 2 presents the result on four datasets. Our approach achieves state-of-the-art results in ZSL setting, but not the best among all datasets.

The seen classes are not involved in the comparison in ZSL setting, making discriminative property decisive in recognition. Since we mainly focus on the knowledge transfer in the GZSL setting, multilabel classification loss plays a major role in guiding the training. In our work, the transferability is gained by sacrificing the discriminative property, which accounts for the relatively low accuracy in the ZSL setting.

4.4. Evaluation in GZSL Settings

The ZSL setting relies on the assumption that the test classes consist of only unseen classes, which can be overly strict and unrealistic. Hence we mainly optimize the performance in the GZSL setting, where the test images come from both seen and unseen classes. In the GZSL setting, the label search space is expanded to all classes, which is formulated as:

$$C_{gzsl}\left(x_i\right) = \arg\max_{j} \left\{o_{i,j}\right\},\ j \in Y^s \cup Y^u \tag{13}$$

Besides, the network is expected to recognize both seen and unseen images. Therefore, a good network should achieve high accuracy on both seen and unseen classes. Suppose that the ACA for the test samples from unseen classes is ACA_u and ACA_s is for the samples from seen classes. Their Harmonic mean H is obtained by $H = \frac{2 \times ACA_s \times ACA_u}{ACA_s + ACA_u}$, which is taken as the evaluation metric of GZSL setting.

Table. 1 shows the ACA_s, ACA_u, H on four datasets. We achieve the best performance in three of four datasets except for the CUB dataset. CUB is a fine-grained dataset containing 200 different types of birds. To separate them apart from each other, the network must have a strong discriminative property, which is not the strength of our approach. We owe the success in the GZSL setting to the excellent transferabil-

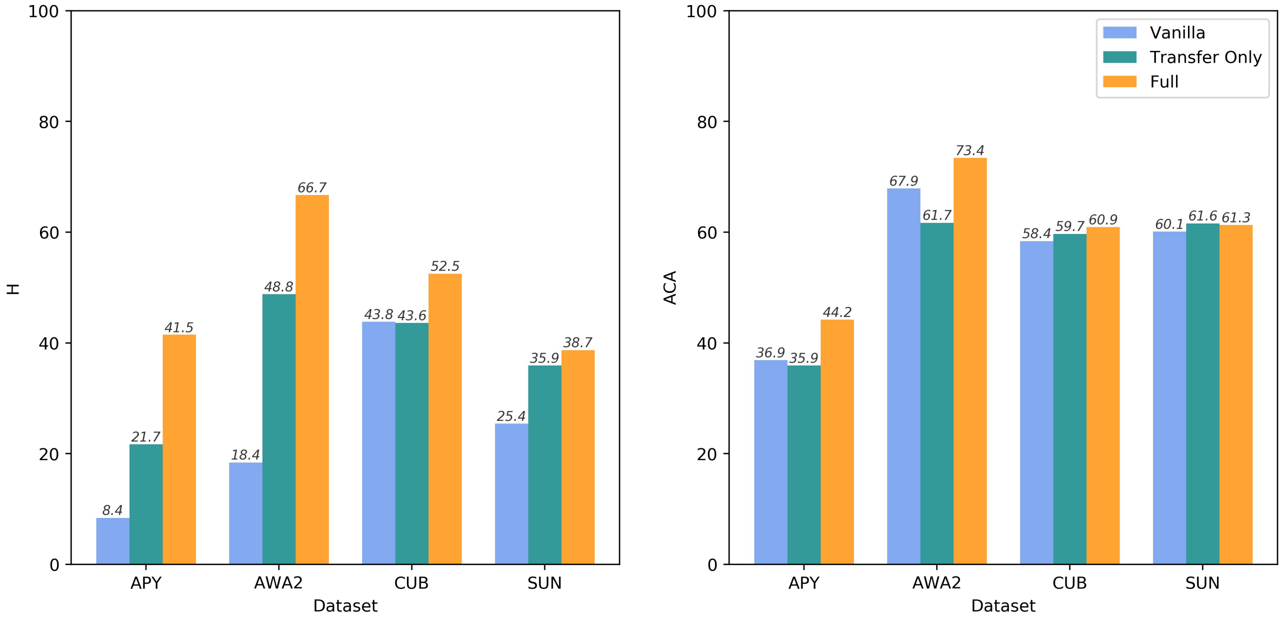

Figure 2. Ablation study results on four datasets, in %. The left is the result in GZSL setting, while the left is in ZSL setting 'Full' denotes the complete approach proposed in this paper. 'Vanilla' represents that loss function is a Cross-Entropy loss on seen classes without knowledge transfer. 'Transfer only' denotes that only transfer loss is available in the experiments.

ity, reflected by gaining high accuracy on unseen data while maintaining the accuracy of seen data.

4.5. Ablation Study

4.5.1. Loss for Knowledge Transfer

Viewing ZSL as a multilabel classification task and using class similarities to facilitate knowledge transfer are the base concepts of our approach. Without the loss for knowledge transfer, the network is totally blind on any information of unseen classes. To measure the importance of knowledge transfer, we compare our results with the vanilla network, where the knowledge transfer loss and pseudo unseen feature are removed. Figure. 2 shows the comparison.

In the GZSL setting, there is a sharp drop on H with the absence of knowledge transfer, indicating that there is little knowledge of unseen data learned from seen data. In the ZSL setting, the ACA on unseen data is low but still comparative. Because the attributes of seen and unseen classes have the same structure and only differ in values, the transformation from attributes to attention masks learned in seen classes can be easily transferred to unseen classes. Also the classifier does not depend on the category of the input. Therefore the vanilla network can achieve a relatively good result in the ZSL setting. In the GZSL setting, due to the different magnitudes of classification scores, many unseen class instances are wrongly classified into seen classes by the vanilla network, resulting in the bad performance.

4.5.2. Loss for Discriminative Property

As we pointed out before, viewing ZSL as a multilabel classification task will weaken the discriminative property. The loss for discriminative property is designed to help the classifier to distinguish different classes. Without the discriminative loss, our model does not perform well in both ZSL and GZSL settings. See more details in Table 2.

5. Conclusion

In this paper, we view GZSL as a multilabel classification task and propose a simple network structure with an attention module. We combine two losses to enhance the transferability and discriminative property of our model on both seen and unseen data. The loss for knowledge transfer makes use of the similarities between seen and unseen classes, which is obtained by semantic attribute, to teach classifier to learn the information of unseen classes. The loss for discriminative property fixes the damage on the discriminative property caused by multilabel classification. Extensive experiments on four datasets verify the effectiveness of the proposed approach, and demonstrate the advantages over the state-of-the-art methods in GZSL setting.

References

Atzmon, Y. and Chechik, G. Adaptive confidence smoothing for generalized zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 11671–11680. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.01194.

Cacheux, Y. L., Borgne, H. L., and Crucianu, M. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 10332–10341. IEEE, 2019. doi: 10.1109/ICCV.2019.01043.

Chao, W., Changpinyo, S., Gong, B., and Sha, F. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In Leibe, B., Matas, J., Sebe, N., and Welling, M. (eds.), *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II*, volume 9906 of *Lecture Notes in Computer Science*, pp. 52–68. Springer, 2016. doi: 10.1007/978-3-319-46475-6_4.

Farhadi, A., Endres, I., Hoiem, D., and Forsyth, D. A. Describing objects by their attributes. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 1778–1785. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206772.

Fu, Y., Hospedales, T. M., Xiang, T., Fu, Z., and Gong, S. Transductive multi-view embedding for zero-shot recognition and annotation. In Fleet, D. J., Pajdla, T., Schiele, B., and Tuytelaars, T. (eds.), *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, volume 8690 of *Lecture Notes in Computer Science*, pp. 584–599. Springer, 2014. doi: 10.1007/978-3-319-10605-2_38.

Fu, Y., Hospedales, T. M., Xiang, T., and Gong, S. Transductive multi-view zero-shot learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(11):2332–2345, 2015. doi: 10.1109/TPAMI.2015.2408354.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Huang, H., Wang, C., Yu, P. S., and Wang, C. Generative dual adversarial network for generalized zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 801–810. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00089.

Jiang, H., Wang, R., Shan, S., and Chen, X. Transferable contrastive network for generalized zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9764–9773. IEEE, 2019. doi: 10.1109/ICCV.2019.00986.

Lampert, C. H., Nickisch, H., and Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 951–958. IEEE Computer Society, 2009. doi: 10.1109/CVPR.2009.5206594.

Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., and Huang, Z. Leveraging the invariant side of generative zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 7402–7411. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00758.

Liu, S., Long, M., Wang, J., and Jordan, M. I. Generalized zero-shot learning with deep calibration network. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 2009–2019, 2018.

Liu, Y., Guo, J., Cai, D., and He, X. Attribute attention for semantic disambiguation in zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 6697–6706. IEEE, 2019a. doi: 10.1109/ICCV.2019.00680.

Liu, Y., Guo, J., Cai, D., and He, X. Attribute attention for semantic disambiguation in zero-shot learning. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 6697–6706. IEEE, 2019b. doi: 10.1109/ICCV.2019.00680.

Patterson, G., Xu, C., Su, H., and Hays, J. The SUN attribute database: Beyond categories for deeper scene understanding. *Int. J. Comput. Vis.*, 108(1-2):59–81, 2014. doi: 10.1007/s11263-013-0695-z.

Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H. S., and Hospedales, T. M. Learning to compare: Relation network for few-shot learning. In *2018 IEEE Conference* *on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1199–1208. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00131.

Verma, V. K., Arora, G., Mishra, A., and Rai, P. Generalized zero-shot learning via synthesized examples. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4281–4289. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00450.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. In *Technical Report CNS-TR-2011-001*, 2011.

Xian, Y., Lorenz, T., Schiele, B., and Akata, Z. Feature generating networks for zero-shot learning. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5542–5551. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00581.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning: A comprehensive evaluation of the good, the bad and the ugly. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768.

Xie, G., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., and Shao, L. Attentive region embedding network for zero-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 9384–9393. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00961.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A. C., Salakhutdinov, R., Zemel, R. S., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Bach, F. R. and Blei, D. M. (eds.), *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, volume 37 of *JMLR Workshop and Conference Proceedings*, pp. 2048–2057. JMLR.org, 2015.

Ye, M. and Guo, Y. Zero-shot classification with discriminative semantic representation learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5103–5111. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.542.

Zhu, P., Wang, H., and Saligrama, V. Generalized zero-shot recognition based on visually semantic embedding. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2995–3003. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00311.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., and Elgammal, A. A generative adversarial approach for zero-shot learning from noisy texts. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1004–1013. IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00111.