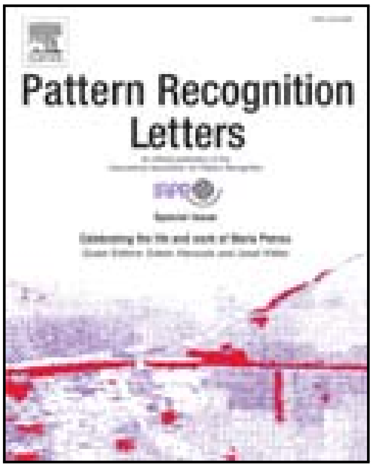
Journal Pre-proof



Pseudo Distribution on Unseen Classes for Generalized Zero Shot Learning

Haofeng Zhang, Jingren Liu, Yazhou Yao, Yang Long

PII: S0167-8655(20)30200-2
DOI: https://doi.org/10.1016/j.patrec.2020.05.021
Reference: PATREC 7907

To appear in: *Pattern Recognition Letters*

Received date: 12 November 2019
Revised date: 5 May 2020
Accepted date: 18 May 2020

Please cite this article as: Haofeng Zhang, Jingren Liu, Yazhou Yao, Yang Long, Pseudo Distribution on Unseen Classes for Generalized Zero Shot Learning, *Pattern Recognition Letters* (2020), doi: https://doi.org/10.1016/j.patrec.2020.05.021

Highlights

1. Attribute similarity is exploited as the Pseudo distribution to solve the over-fitting on GZSL;

2. Attribute similarity is further compressed as one-hot vector to encourage the certainty of the training;

3. Visual space is employed as the embedding space to alleviate the hubness problem;

4. The proposed PSD can significantly outperform the SOTA methods by large margins on GZSL

Pattern Recognition Letters
journal homepage: www.elsevier.com

Pseudo Distribution on Unseen Classes for Generalized Zero Shot Learning

Haofeng Zhang[a,**], Jingren Liu[a], Yazhou Yao[a], Yang Long[b]

[a]School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
[b]School of Computer Science, Durham University, Durham, UK

ABSTRACT

Although Zero Shot Learning (ZSL) has attracted more and more attention due to its powerful ability of recognizing new objects without retraining, it has a serious drawback that it only focuses on unseen classes during prediction. To solve this issue, Generalized ZSL (GZSL) extends the search range to both seen and unseen classes, which makes it a more realistic and challenging task. Conventional methods on GZSL often suffer from the domain shift problem on seen classes because they have only seen data for training. Deep Calibration Network (DCN) tries to minimize the entropy of assigning seen data to unseen classes to balance the training on both seen and unseen classes. However, there are still two problems for DCN, one is the hubness problem and another is the lack of training guidance. In this paper, to solve the two problems, we propose a novel method called PSeudo Distribution (PSD), which exploits the attribute similarity between seen classes and unseen classes as the training guidance to assign the seen data to unseen classes. In addition, the attribute similarity is also compressed to one-hot vector to further encourage the certainty of the model. Besides, the visual space is utilized as the embedding space, which can well settle the hubness problem. Extensive experiments are conducted on four popular datasets, and the results show the superiority of the proposed method.

Keywords: Generalized Zero Shot Learning; Pseudo Distribution; Attribute Similarity

1. Introduction

In the recent decade, deep learning has gained great success in many areas, especially on image classification that the top-5 accuracy on ImageNet Deng et al. (2009) is over 95%, which is considered to have exceeded the recognition ability of human beings. However, the traditional classification is conducted on close-set that the test classes should be same as the training data. In the current era of data explosion, an increasing number of new categories have been emerging everyday, and the close-set classification cannot fulfill the requirement of recognizing new objects. Therefore, Zero Shot Learning (ZSL) was proposed to solve such problem Lampert et al. (2009); Zhang et al. (2019d); Guo et al. (2017), it has attracted more and more attention due to its ability of recognizing unseen categories, and has been applied in many areas such as image classification Long et al. (2018b); Yu et al. (2018b); Chiaro et al. (2019), multimedia retrieval Ji et al. (2020) and object detection Bansal et al. (2018).

ZSL is inspired by the way of human understanding new things, *e.g.*, if a child has never seen a zebra before, but he is told that the zebra has the same shape as a horse and black-and-white stripes, then when he sees a zebra, he must be able to recognize it. Therefore, ZSL usually employs medium information, such as semantic attributes, to bridge the seen classes and the unseen classes. Most ZSL works project visual features into semantic space and find the nearest neighbor from predefined attributes Lampert et al. (2009); Akata et al. (2013); Zhang et al. (2019a,c). This type of methods often suffer from serious domain shift problem due to their negligence of unseen classes during training, *i.e.*, they can obtain good performance on seen classes but perform bad on unseen classes, especially on the

[**]Corresponding author.
e-mail: zhanghf@njust.edu.cn (Haofeng Zhang), chengxuyuangg@njust.edu.cn (Jingren Liu), yazhou.yao@njust.edu.cn (Yazhou Yao), yang.long@ieee.org (Yang Long)

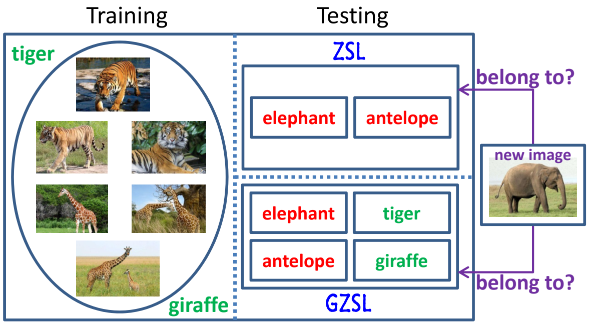


Fig. 1. Illustration of the difference between ZSL and GZSL. The green font stands for seen classes and the red font means unseen classes.

more reasonable and realistic Generalized ZSL (GZSL) setting, which is first defined by Chao *et al.* Chao et al. (2016), because GZSL extends the search scope from only the seen classes to both seen and unseen classes during testing. The difference between ZSL and GZSL is illustrated in Fig. 1.

To cope with such problem, Liu *et al.* proposed a Deep Calibration Network (DCN) Liu et al. (2018a) to balance the training between seen and unseen classes by computing the similarity between the seen data and the unseen classes. DCN assumes that the entropy of assigning a seen data to an unseen class should be minimized to constrain the certainty of the model, and achieves a significant success. However, there are still two problems. The first one is that DCN often suffers from the hubness problem because it use the low-dimensional attribute space or latent space as its embedding space, *i.e.*, a few unseen class prototypes will become the nearest neighbors of many data points. Using the semantic space as the embedding space means that the visual feature vectors need to be projected into the semantic space which will shrink the variance of the projected data points and thus aggravate the hubness problem Zhang et al. (2017). To solve this problem, some recent works Huang et al. (2019); Xian et al. (2018b); Zhu et al. (2018); Li et al. (2019) propose to use generative methods that can generate unseen visual features conditioned on their corresponding attributes, and the generated features are combined with the seen samples to train a fully supervised model. However, these generative methods often include two steps, which make them not end-to-end training. In addition, some researchers Shigeto et al. (2015); Zhang et al. (2017) proposed to project semantic attributes into high dimensional space, which is easy to implement and can achieve significant improvement, so we will also adopt this strategy in our method. The second problem is that DCN does not know which unseen class should be guided to for a seen data point although it can encourage the certainty of classifying it to an unknown unseen class, which is called training guidance problem here.

To solve the above two mentioned problems, in this paper, we propose a novel method called PSeudo Distribution (PSD) of seen data on unseen classes, which can well alleviate the problem of training guidance. Concretely, we employ the attribute similarity between the seen classes and unseen classes as the training guidance, which can directly guide the training direction for the unseen classes. Furthermore, to be more certainty, we also compress the attribute similarities to one-hot vectors. In addition, we take the visual space as the embedding space, and project the semantic attributes to visual space as visual prototypes to mitigate the hubness problem. The contributions of our work are listed as follows,

- We propose a novel method called PSD for the assignment of seen data on unseen classes to alleviate the domain shift problem on GZSL;
- The attribute similarity between the seen classes and unseen classes are exploited as the PSD to guide the direction for training of the unseen classes; besides, the visual space is employed as the embedding space to solve the hubness problem;
- Extensive experiments are conducted on four popular datasets, and the results shows the proposed method can significantly outperform the state-of-the-art methods by large margins.

The main content of this paper is organized as follows: In section 2 we briefly introduce the existing methods for ZSL. Section 3 describes the proposed method in detail. Section 4 gives the experimental results of comparison with existing methods on GZSL. Finally in section 5, we conclude this paper.

2. Related Works

Zero Shot Learning (ZSL) tries to classify unseen data points by transferring the knowledge learned from seen classes to unseen classes with semantic attributes. Recently, a large number of researchers have been endeavoring in this field. The earliest works such as Direct Attribute Projection (DAP) Lampert et al. (2009) directly project the visual features into semantic space, and then learn a SVM classifier to estimate the labels. In Attribute Label Embedding (ALE) Akata et al. (2016) and SJE Akata et al. (2015a), Akata *et al.* projected visual features into semantic space via a bilinear compatibility constraint and maximize the similarity between different attributes with a max margin loss. CONvex combination of Semantic Embeddings (CONSE) Norouzi et al. (2014) and Semantic Similarity Embedding (SSE) Zhang and Saligrama (2015) try to build unseen attributes automatically from the instances of seen categories to reduce the requirement of manual attributes. Furthermore, some researchers such as Kodirov *et al.* Kodirov et al. (2017) introduced the concept of Auto-Encoder and directly use the Euclidean distance to constrain the similarity of projected vectors in both visual and attribute spaces.

To alleviate the hubness problem, Zhang *et al.* proposed a method called Deep Embedding Model (DEM) Zhang et al. (2017) to use the visual space as the embedding space, in which the subsequent nearest neighbour search becomes more effective. Although DEM can achieve better performance than those methods using semantic space as the embedding space, it still suffers from the domain shift problem due to its negligence of unseen classes during training. For fine-grained zero shot image classification problem, Ji *et al.* proposed to exploit an attention

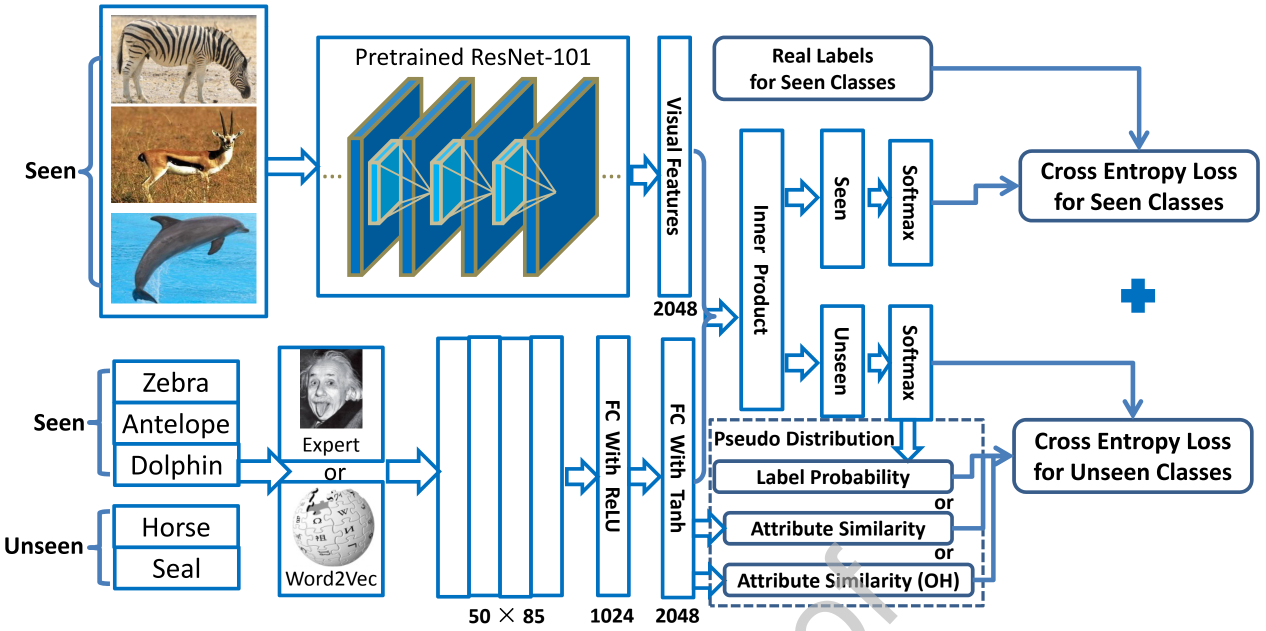


Fig. 2. The flowchart of the proposed method. The dashed box contains three types of pseudo distributions.

network Ji et al. (2019); Yu et al. (2018a) to obtain semantic relevant features by using individual class semantic features, and generate an attention map for weighting the importance of different local regions, which has obtained great success. Liu *et al.* Liu et al. (2020) exploited the one-hot label space as the common embedded space for both visual features and semantic attributes during training, and considered the labels of unseen classes as the linear combination of labels of seen classes in test phase.

In addition, Long *et al.* in Long et al. (2017) proposed to use the attributes of unseen classes to synthesize unseen visual features, and then train a supervised model with seen and synthesized unseen visual features. Thereafter, due to the powerful ability of sample synthesis, an increasing number of Generative Adversarial Net (GAN) Goodfellow et al. (2014) based ZSL methods have been proposed Zhang et al. (2019b); Huang et al. (2019); Xian et al. (2018b); Zhu et al. (2018); Li et al. (2019); Yu et al. (2020), and these methods leverage the unseen attributes to generate synthesized unseen visual features, which are subsequently combined with seen visual features to learn a fully supervised model. Those methods can achieve state-of-the-art performance, but they all suffer from the same problem as the close-set classification that when there is a totally new category it should be retrained by adding the synthesized samples of the new class.

Different from conventional ZSL, which assumes that all the test samples are only from unseen categories, Generalized ZSL (GZSL), which is firstly proposed by Chao *et al.* in Chao et al. (2016), enlarges the search scope to all classes, because we cannot obtain the information that whether the test data only belongs to the unseen classes beforehand in most scenarios, therefore GZSL is a more realistic and challenging task. Besides, due to the fact that there was no agreed upon ZSL benchmark, Xian *et al.* Xian et al. (2018a) defined a new benchmark by unifying both the evaluation protocols and data splits of several publicly available datasets. They also analyzed a significant number of the state-of-the-art methods in depth, both in the classic ZSL setting and the more realistic GZSL setting, which has made a great contribution to this research field.

The most relevant to our method is the Deep Calibration Network (DCN) Liu et al. (2018a), which projects the visual features and attributes into a latent space, where the visual features are assigned to both seen classes and unseen classes to balance the training in order to solve the domain shift problem. However, there are still two problems to be solved, one is the hubness problem caused by using the attribute space as the embedding space, another is the lack of training guidance that DCN only minimizes the entropy of assigning seen data to unseen classes but ignores which class should be assigned.

3. Methodology

3.1. Problem Definition

Let $C_s = \{c_1, c_2, \cdots, c_m\}$ denote a set of seen classes and $C_u = \{c_{m+1}, c_{m+2}, \cdots, c_{m+n}\}$ is the set of unseen classes, where m and n are the numbers of the seen classes and unseen classes. The two sets are disjoint, *i.e.*, $C_s \cap C_u = \emptyset$. In addition, $\boldsymbol{A}^s = \{\boldsymbol{a}_1^s, \boldsymbol{a}_2^s, \cdots, \boldsymbol{a}_m^s\} \in \mathbb{R}^{d_a \times m}$ and $\boldsymbol{A}^u = \{\boldsymbol{a}_{m+1}^u, \boldsymbol{a}_{m+2}^u, \cdots, \boldsymbol{a}_{m+n}^u\} \in \mathbb{R}^{d_a \times n}$ are the two corresponding attribute sets for the seen classes and unseen classes respectively, where d_a is the dimension of the attribute vector.

Suppose it is given a set of labeled features $\boldsymbol{X}^s = \{\boldsymbol{x}_1^s, \cdots, \boldsymbol{x}_i^s, \cdots, \boldsymbol{x}_{N_s}^s\} \in \mathbb{R}^{d_x \times N_s}$ from the seen classes, where d_x is the dimension of the feature and N_s is the number of samples. Let $\boldsymbol{X}^u = \{\boldsymbol{x}_1^u, \cdots, \boldsymbol{x}_i^u, \cdots, \boldsymbol{x}_{N_u}^u\} \in \mathbb{R}^{d_x \times N_s}$ denote the test dataset from the unseen classes, where N_u is the number of the samples, and there is no label for them. The objective of GZSL is to assign labels to the test data $\boldsymbol{X}^u$ by learning a classifier $\mathcal{F}(\boldsymbol{x}_i^u) \rightarrow \mathcal{S}_s \cup \mathcal{S}_u$ with the training data $\boldsymbol{X}^s$ and the whole attribute set $\boldsymbol{A}^s \cup \boldsymbol{A}^u$.

3.2. Architecture

The flowchart of the proposed method is shown in Fig. 2, where the upper branch is designed for images feature extraction, and the bottom branch is utilized for attribute projection. In the upper branch, we use the pre-trained ResNet101 He et al. (2016) as the feature extraction module, and the parameters of it are fixed during training and test. In the bottom branch, the attributes are extracted from class names with Word2Vec or annotated by experts, and we use the expert-annotated attributes in our method. The attributes pass through a two-layer fully connection network, which has a Rectified Linear Unit (ReLU) attached to the first layer and a tanh operation appended to the second layer.

The inner product of visual features and all the attributes are executed, and the results are divided into two parts, the seen one and the unseen one. The seen part is a traditional fully supervised problem, thus we employ the classical cross entropy as its loss function $\mathcal{J}$. For the unseen part, since there is no ground-truth matching for seen data and unseen classes, we introduce the pseudo distribution as its label. Therefore, the unseen part can also be calculated with the cross entropy loss $\mathcal{H}$, and the total loss function can be defined as,

$$\mathcal{L} = \mathcal{J} + \lambda\mathcal{H}, \tag{1}$$

where λ is the balancing coefficient to control the importance of the two parts in Eq. 1. In the following subsections, we will describe the details of the seen part and the unseen part.

3.3. Loss for Seen Classes

Because the visual features are extracted with fixed pre-trained ResNet, they can be considered as the direct input of the upper branch. If we denote the attribute projection function in the bottom part as $f(\boldsymbol{a}_j)$, the probability of the input visual feature $\boldsymbol{x}_i^s$ on the seen category $\boldsymbol{a}_j^s$ can be defined as,

$$p^s(\boldsymbol{x}_i^s, \boldsymbol{a}_j^s) = \frac{\exp^{<\boldsymbol{x}_i^s, f(\boldsymbol{a}_j^s)>}}{\sum_{j=1}^{m} \exp^{<\boldsymbol{x}_i^s, f(\boldsymbol{a}_j^s)>}}, \tag{2}$$

where, $< \cdot, \cdot >$ is the inner product. Then, the entropy loss can be calculated as,

$$\begin{split}\mathcal{J} = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=1}^{m} y_{ij} \log p^s(\boldsymbol{x}_i^s, \boldsymbol{a}_j^s) \\ + (1 - y_{ij}) \log(1 - p^s(\boldsymbol{x}_i^s, \boldsymbol{a}_j^s)),\end{split} \tag{3}$$

where, y_{ij} denotes the one-hot label value of $\boldsymbol{x}_i^s$ on the seen category c_j. If $\boldsymbol{x}_i$ belongs to the class c_j, $y_{ij} = 1$, otherwise $y_{ij} = 0$.

3.4. Loss for Unseen Classes

Similarly, the probability of a seen visual feature $\boldsymbol{x}_i^s$ on an unseen class $\boldsymbol{a}_j^u$ is defined as,

$$p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^u) = \frac{\exp^{<\boldsymbol{x}_i^s, f(\boldsymbol{a}_j^u)>}}{\sum_{j=m+1}^{m+n} \exp^{<\boldsymbol{x}_i^s, f(\boldsymbol{a}_j^u)>}}. \tag{4}$$

Due to the fact that there is no label for a seen feature on an unseen class, thus we have to define a pseudo distribution for it. In the following, we will give three types of pseudo distributions.

Entropy:

It is known that the entropy means when the data source produces a low-probability value (*i.e.*, when a low-probability event occurs), the event carries more "information" than when the source data produces a high-probability value [1]. Generally, entropy refers to disorder or uncertainty, the hight the entropy is the more uncertainty the model is. Therefore, the entropy in our method should be the lower the better. Similar as that in Liu et al. (2018a), we define the entropy loss for the unseen part,

$$\mathcal{H}_E = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=m+1}^{m+n} p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^u) \log p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^u). \tag{5}$$

In Eq. 5, we can consider $p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^u)$ as the pseudo distribution of $\boldsymbol{x}_i^s$ on the unseen classes, and the loss function as the cross entropy for the probability and the pseudo distribution.

Attribute Similarity:

Although the entropy loss defined above can encourage the model to be certainty, there is no clear guidance for a feature to be classified to a fixed class. Therefore, to solve such problem, we employ the attribute similarity as the pseudo distribution of the seen data on unseen classes. The similarity of attributes can be defined as,

$$s_{ij} = Simi(\boldsymbol{a}_i^s, \boldsymbol{a}_j^u) = \frac{< f(\boldsymbol{a}_i^s), f(\boldsymbol{a}_j^u) >}{\|f(\boldsymbol{a}_i^s)\|_2 \|f(\boldsymbol{a}_j^u)\|_2}. \tag{6}$$

In Eq. 6, we use $f(\boldsymbol{a}_i^s)$ and $f(\boldsymbol{a}_j^u)$ instead of $\boldsymbol{a}_i^s$ and $\boldsymbol{a}_j^u$ because some original attributes are very similar and not discriminative, *e.g.*, "blue wale", "humpback wale" and "killer wale", which will make them hard to be classified. Similar as the seen classes, the loss function of the unseen part can be defined as,

$$\mathcal{H}_S = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=m+1}^{m+n} s_{ij} \log p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^u). \tag{7}$$

One-hot Attribute Similarity:

To be more certainty, we further compress the attribute similarity to be one-hot vector as follows,

$$h_{ij} = \begin{cases} 1 & j = \underset{j \in \{m+1, \cdots, m+n\}}{\arg\max} \; s_{ij} \\ 0 & otherwise \end{cases}, \tag{8}$$

[1] https://en.wikipedia.org/wiki/Entropy

Table 1. Summary of the employed four datasets.

Datasets	Dimension		Class Number		Samples Number		
	Feat.	*Att.*	*Seen*	*Unseen*	*SS*	*TS*	*TR*
SUN	2048	102	645	72	10320	1440	2580
CUB	2048	312	150	50	7057	2967	10320
AWA	2048	85	40	10	19832	4958	5685
aPY	2048	64	20	12	5932	7924	1483

where s_{ij} is computed with Eq. 6. The final loss function of the unseen part can be calculated as,

$$\mathcal{H}_O = -\frac{1}{N_s}\sum_{i=1}^{N_s}\sum_{j=m+1}^{m+n} h_{ij}\log p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^s) + (1-h_{ij})\log(1-p^u(\boldsymbol{x}_i^s, \boldsymbol{a}_j^s)). \tag{9}$$

After the definition of three types of the loss functions for the unseen part, $\mathcal{H}$ in Eq. 1 can be replaced with each of $\mathcal{H}_E$, $\mathcal{H}_S$ and $\mathcal{H}_O$. Since the network in Fig. 2 is an end-to-end architecture, the projection parameters for the bottom branch can be easily optimized by applying the mini-batch Stochastic Gradient Descent (SGD).

3.5. Zero Shot Classification

When the network training is finished, the label of the unseen data can be predicted with the following equation,

$$c = \underset{j\in\{1,\cdots,m+n\}}{\arg\max} < \boldsymbol{x}_i^u, f(\boldsymbol{a}_j) >, \tag{10}$$

where $\boldsymbol{a}_j$ denotes $\boldsymbol{a}_j^s$ when $j \in \{1,\cdots,m\}$ and $\boldsymbol{a}_j^u$ for $j \in \{m+1,\cdots,m+n\}$.

4. Experiments

4.1. Datasets and Settings

Datasets: In our experiments, we employ four popular benchmark datasets, *i.e.*, SUN (SUN attribute) Patterson et al. (2014), CUB (Caltech-UCSD-Birds 200-2011) Wah et al. (2011), AWA(Animals with Attributes) Lampert et al. (2009) and aPY(Attribute Pascal and Yahoo) Farhadi et al. (2009a). SUN is a type of fine-grained dataset, which contains many different visual scenes and CUB is also a fine-grained dataset, and it is consisted of 200 bird-species. AWA is a coarse-grained dataset of 50 classes of animals. APY has 20 classes from Pascal VOC Everingham et al. (2010) for training and 12 classes from Yahoo Farhadi et al. (2009b) for test. The other details of these datasets can be found in Tab. 1, where 'SS' refers to number of Seen Samples in training, 'TS' is the number of samples from unseen classes for test, and 'TR' is for the seen ones. In addition, we adopt the split strategy which is proposed by Xian et al. (2018a).

Settings: There are three hyper-parameters in our method, *i.e.*, the learning rate of the deep network, the batch size for SGD and the balancing coefficient λ for Eq. 1. The learning rate and the batch size are set to 3×10^{-4} and 128 respectively for all four datasets. λ is decided by applying cross-validation, which is different from traditional fully supervised strategy. Here we split the part of the seen classes as the validation unseen classes, and the searching range of λ is restricted to $[0.01, 1]$.

Evaluational Metrics: Conventional ZSL metric assumes that the test data in advance are known belonging to unseen classes, and will be tested only on unseen classes, which is unreasonable in realistic scenarios. We usually do not know the ascription of the test data in advance, thus it is necessary to find the best assignment on both seen and unseen classes. Furthermore, the model should be not only suitable for unseen classes but also should maintain the performance on seen classes. The metrics are described as follows,

- Seen test accuracy tr: Average per-class classification accuracy for seen test samples;
- Unseen test accuracy ts: Average per-class classification accuracy for unseen test samples;
- Harmonic accuracy H: traditional arithmetic mean $H = (tr + ts)/2$, which computes the average value of tr and ts, can still generate good results when one of tr and ts is high and the other is very low. However, very low accuracy on single metric often means the trained model fails, thus here we use harmonic accuracy $H = (2 \times tr \times ts)/(tr + ts)$ Xian et al. (2018a) to replace the arithmetic mean.

4.2. Results on GZSL

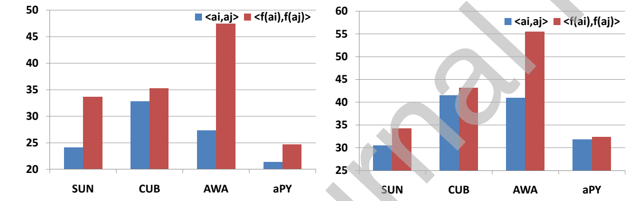
We conduct the the experiments on the above mentioned four datasets, and the results are reported in Tab. 2. Since our method is specially designed for GZSL, which is more reasonable in realistic scenarios, we do not report the results on conventional ZSL. Besides, in order to better show the superiority of the proposed method, we compare it with 22 methods, which can be found in Tab. 2. The results of the first 12 methods are directly cited from Xian et al. (2018a), the results of PRESERVE Annadani and Biswas (2018), CDL Jiang et al. (2018), LAGO Atzmon and Chechik (2018), PSEUDO Long et al. (2018a), KERNEL Zhang and Koniusz (2018), TVN Zhang et al. (2019a), DEM Zhang et al. (2017), LESAE Liu et al. (2018b), ICINESS Guo et al. (2018) and DCN Liu et al. (2018a) are excerpted from their original papers, and the remaining methods such as VZSL Wang et al. (2018) and LESD Ding et al. (2017) are implemented by us according to the original description in those papers.

Compared with these baselines, it can be clearly found that our method can outperform all of them, and exceed them with large margins. Concretely, our method can achieve best performances on both ts and H on all four datasets. Especially, compared to the most similar method DCN, our method can obtain the improvements by 8.8%, 10.2%, 11.5% and 13.8% respectively on ts and 6.6%, 6.0%, 17.1% and 9.8% respectively on H.

In addition, the entropy loss for unseen part in our method is same as that in DCN Liu et al. (2018a) but with different network architecture, so we report the results, which is recorded as **PSD-EN**, to show the effect of the proposed network. It can be seen that although the two methods have same loss function, the different network can improve the performance by 3.4%, 3.9%, 13.1% and 7.4% respectively on H. Furthermore, **PSD-AS** for

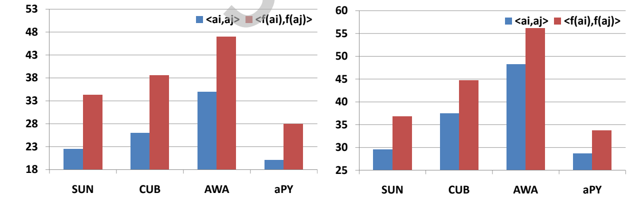
Table 2. Comparison with state-of-the-art baselines on GZSL setting.'-' means not reported.

	SUN			CUB			AWA			aPY		
Method	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**	**ts**	**tr**	**H**
DAP Lampert et al. (2009)	4.2	25.1	7.5	1.7	67.9	3.3	0.0	**88.7**	0.0	4.8	78.3	9.0
CONSE Norouzi et al. (2014)	6.8	39.9	11.6	1.6	**72.2**	3.1	0.4	88.6	0.8	0.0	**91.2**	0.0
SSE Zhang and Saligrama (2015)	2.1	36.4	4.0	8.5	46.9	14.4	7.0	80.5	12.9	0.2	78.9	0.4
LATEM Xian et al. (2016)	14.7	28.8	19.5	15.2	57.3	24.0	7.3	71.7	13.3	0.1	73.0	0.2
ALE Akata et al. (2013)	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	4.6	73.7	8.7
DEVISE Frome et al. (2013)	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	4.9	76.9	9.2
SJE Akata et al. (2015b)	14.7	30.5	19.8	23.5	59.2	33.6	11.3	74.6	19.6	3.7	55.7	6.9
ESZSL Romera-Paredes and Torr (2015)	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	2.4	70.1	4.6
SAE Kodirov et al. (2017)	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	0.4	80.9	0.9
SYNC Changpinyo et al. (2016)	7.0	**43.3**	13.4	11.5	70.9	19.8	8.9	87.3	16.2	7.4	66.3	13.3
GFZSL Verma and Rai (2017)	0.0	39.6	0.0	0.0	45.7	0.0	1.8	80.3	3.5	0.0	83.3	0.0
PRESERVE Annadani and Biswas (2018)	20.8	37.2	26.7	24.6	54.3	33.9	-	-	-	13.5	51.4	21.4
CDL Jiang et al. (2018)	21.5	34.7	26.5	23.5	55.2	32.9	28.1	73.5	40.6	19.8	48.6	28.1
LAGO Atzmon and Chechik (2018)	18.8	33.1	23.9	21.8	73.6	33.7	23.8	67.0	35.1	-	-	-
PSEUDO Long et al. (2018a)	19.0	32.7	24.0	23.0	51.6	31.8	22.4	80.6	35.1	15.4	71.3	25.4
KERNEL Zhang and Koniusz (2018)	21.0	31.0	25.1	24.2	63.9	35.1	18.3	79.3	29.8	11.9	76.3	20.5
TVN Zhang et al. (2019a)	22.2	38.3	28.1	26.5	62.3	37.2	27.0	67.9	38.6	16.1	66.9	25.9
VZSL Wang et al. (2018)	15.2	23.8	18.6	17.1	37.1	23.8	22.3	77.5	34.6	8.4	75.5	15.1
DEM Zhang et al. (2017)	20.5	34.3	25.6	19.6	57.9	29.2	32.8	84.7	47.3	11.1	75.1	19.4
LESAE Liu et al. (2018b)	21.9	34.7	26.9	24.3	53.0	33.3	19.1	70.2	30.0	12.7	56.1	20.1
LESD Ding et al. (2017)	15.2	19.8	17.2	14.6	38.5	21.2	12.6	71.0	21.4	11.8	49.3	19.0
ICINESS Guo et al. (2018)	-	-	32.1	-	-	41.8	-	-	41.0	-	-	25.4
DCN Liu et al. (2018a)	25.5	37.0	30.2	28.4	60.7	38.7	25.5	84.2	39.1	14.2	75.0	23.9
PSD-EN	31	36.7	33.6	34.0	57.0	42.6	40.9	72.2	52.2	26.5	38.4	31.3
PSD-AS	33.7	35.0	34.3	35.3	55.6	43.2	**47.4**	67.0	55.5	24.7	47.0	32.4
PSD-OS	**34.3**	39.7	**36.8**	**38.6**	53.1	**44.7**	47.0	70.0	**56.2**	**28.0**	42.3	**33.7**



(a) *ts* (b) *H*

Fig. 3. The performance on different computations of attribute similarity.



(a) *ts* (b) *H*

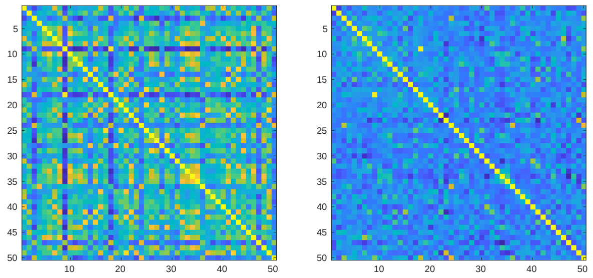
Fig. 4. The performance on different computations of one-hot attribute similarity.

attribute similarity and **PSD-OS** for one-hot attribute similarity can further improve the performances due to their classification guidance and certainty encouragement.

4.3. Ablation Study

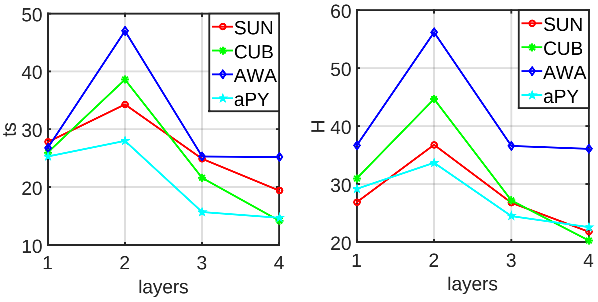
Different attribute similarity: In Eq. 6, we argue that $< f(\boldsymbol{a}_i^s), f(\boldsymbol{a}_j^u) >$ is better than $< \boldsymbol{a}_i^s, \boldsymbol{a}_j^u >$ for computing the similarity of attributes because the processed attributes are more discriminative than their original form. To proof this argument, we conduct experiments on four datasets by replacing the computation of attribute similarity in Eq. 6 from $< f(\boldsymbol{a}_i^s), f(\boldsymbol{a}_j^u) >$ to $< \boldsymbol{a}_i^s, \boldsymbol{a}_j^u >$, and the results are recorded in Fig. 3 and Fig. 4. Form the two figures, it can be clearly discovered that the performances on both *ts* and *H* are significantly improved by employing the processed attributes, especially on AWA. To further show the discriminative attributes after the procession of deep network, we compute the attribute similarity matrices before and after the procession and show them in Fig. 5. It can be clearly seen that the attributes after the deep network process are more discriminative, which is reason for the better performance of the proposed network.

Different network architectures: Our method exploits the two-layer network architecture to project the attribute to visual space and achieve state-of-the-art performance. However, it is necessary to show its performance with other architectures to investigate the importance of the network. In this experiment, we replace the network with 1 layer (semantic→visual), three layers (semantic→512→1024→visual), and four layers (semantic→256→512→1024→visual), and record their performance in Fig. 6. From this figure, it can be clear found that



(a) Before (b) After

Fig. 5. The attribute similarity before and after the procession of the deep network.



(a) *ts* (b) *H*

Fig. 6. The performance on different network architectures for PSD-OS.

the two-layer architecture can achieve the best performance and outperform the others by large margins. This phenomenon reveals the fact that the two-layer architecture is most suitable for this model, and too few parameters can lead to under-fitting while too many will cause over-fitting.

5. Conclusion

In this paper, we have proposed a deep GZSL network, which utilizes the pseudo distribution of seen data on unseen classes to solve the domain shift problem on seen classes. In the network, three types of pseudo distributions including pseudo probability, attribute similarity and one-hot attribute similarity, were employed. In addition, to solve the hubness problem, the network was designed to make inner product in visual space to compute the pseudo probabilities. Extensive experiments on four popular dataset were conducted. The results show the proposed network can significantly improve the performance and the one-hot attribute similarity can achieve best performance due to its strong training guidance and certainty encouragement.

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (61872187), in part by the Medical Research Council (MRC) Innovation Fellowship (MR/S003916/1), and in part by the “111” Program (No.B13022).

References

Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C., 2013. Label-embedding for attribute-based classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

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

Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B., 2015a. Evaluation of output embeddings for fine-grained image classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2927–2936.

Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B., 2015b. Evaluation of output embeddings for fine-grained image classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Annadani, Y., Biswas, S., 2018. Preserving semantic relations for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Atzmon, Y., Chechik, G., 2018. Probabilitic and-or attribute grouping for zero-shot learning, in: The Conference on Uncertainty in Artificial Intelligence.

Bansal, A., Sikka, K., Sharma, G., Chellappa, R., Divakaran, A., 2018. Zero-shot object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 384–400.

Changpinyo, S., Chao, W.L., Gong, B., Sha, F., 2016. Synthesized classifiers for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 5327–5336.

Chao, W.L., Soravit, C., Gong, B., Sha, F., 2016. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild, in: European Conference on Computer Vision, pp. 52–68.

Chiaro, R.D., Bagdanov, A.D., Bimbo, A.D., 2019. Webly-supervised zero-shot learning for artwork instance recognition. Pattern Recognition Letters 128, 420 – 426.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: A large-scale hierarchical image database, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255.

Ding, Z., Shao, M., Fu, Y., 2017. Low-rank embedded ensemble semantic dictionary for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2050–2058.

Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A., 2010. The pascal visual object classes challenge 2007 (voc2007) results. International Journal of Computer Vision 88, 303–338.

Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009a. Describing objects by their attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785.

Farhadi, A., Endres, I., Hoiem, D., Forsyth, D., 2009b. Describing objects by their attributes, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1778–1785.

Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al., 2013. Devise: A deep visual-semantic embedding model, in: Advances in Neural Information Processing Systems.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets, in: Advances in Neural Information Processing Systems, Springer. pp. 2672–2680.

Guo, Y., Ding, G., Han, J., Gao, Y., 2017. Zero-shot learning with transferred samples. IEEE Transactions on Image Processing 26, 3277–3290.

Guo, Y., Ding, G., Han, J., Zhao, S., Wang, B., 2018. Implicit non-linear similarity scoring for recognizing unseen classes, in: Proceedings of the International Joint Conference on Artificial Intelligence, pp. 4898–4904.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778.

Huang, H., Wang, C., Yu, P.S., Wang, C., 2019. Generative dual adversarial network for generalized zero-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 801–810.

Ji, Z., Sun, Y., Yu, Y., Pang, Y., Han, J., 2020. Attribute-guided network for cross-modal zero-shot hashing. IEEE transactions on neural networks and learning systems 31, 321–330.

Ji, Z., Xiong, K., Pang, Y., Li, X., 2019. Video summarization with attention-based encoder-decoder networks. IEEE Transactions on Circuits and Systems for Video Technology doi:`10.1109/TCSVT.2019.2904996`.

Jiang, H., Wang, R., Shan, S., Chen, X., 2018. Learning class prototypes via structure alignment for zero-shot recognition, in: European Conference on Computer Vision, pp. 118–134.

Kodirov, E., Xiang, T., Gong, S., 2017. Semantic autoencoder for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 3174–3183.

Lampert, C.H., Nickisch, H., Harmeling, S., 2009. Learning to detect unseen object classes by between-class attribute transfer, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Li, J., Jing, M., Lu, K., Ding, Z., Zhu, L., Huang, Z., 2019. Leveraging the invariant side of generative zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7402–7411.

Liu, S., Long, M., Wang, J., Jordan, M.I., 2018a. Generalized zero-shot learning with deep calibration network, in: Advances in Neural Information Processing Systems, pp. 2005–2015.

Liu, Y., Gao, Q., Li, J., Han, J., Shao, L., 2018b. Zero shot learning via low-rank embedded semantic autoencoder, in: International Joint Conference on Artificial Intelligence, pp. 2490–2496.

Liu, Y., Gao, X., Gao, Q., Han, J., Shao, L., 2020. Label-activating framework for zero-shot learning. Neural Networks 121, 1–9.

Long, T., Xu, X., Li, Y., Shen, F., Song, J., Shen, H., 2018a. Pseudo transfer with marginalized corrupted attribute for zero-shot learning, in: ACM conference on Multimedia, pp. 1802–1810.

Long, T., Xu, X., Shen, F., Liu, L., Xie, N., Yang, Y., 2018b. Zero-shot learning via discriminative representation extraction. Pattern Recognition Letters 109, 27 – 34.

Long, Y., Liu, L., Shao, L., Shen, F., Ding, G., Han, J., 2017. From zero-shot learning to conventional supervised classification: Unseen visual data synthesis, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 1627–1636.

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G.S., Dean, J., 2014. Zero-shot learning by convex combination of semantic embeddings, in: International Conference on Learning Representation.

Patterson, G., Xu, C., Su, H., Hays, J., 2014. The sun attribute database: Beyond categories for deeper scene understanding. International Journal of Computer Vision 108, 59–81.

Romera-Paredes, B., Torr, P., 2015. An embarrassingly simple approach to zero-shot learning, in: International Conference on Machine Learning.

Shigeto, Y., Suzuki, I., Hara, K., Shimbo, M., Matsumoto, Y., 2015. Ridge regression, hubness, and zero-shot learning, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer. pp. 135–151.

Verma, V.K., Rai, P., 2017. A simple exponential family framework for zero-shot learning, in: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Springer. pp. 792–808.

Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S., 2011. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001. California Institute of Technology.

Wang, W., Pu, Y., Verma, V.K., Fan, K., Zhang, Y., Chen, C., Rai, P., Carin, L., 2018. Zero-shot learning via class-conditioned deep generative models, in: AAAI Conference on Artificial Intelligence.

Xian, Y., Akata, Z., Sharma, G., Nguyen, Q., Hein, M., Schiele, B., 2016. Latent embeddings for zero-shot classification, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition.

Xian, Y., Lampert, C.H., Schiele, B., Akata, Z., 2018a. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. IEEE transactions on pattern analysis and machine intelligence 41, 2251–2265.

Xian, Y., Lorenz, T., Schiele, B., Akata, Z., 2018b. Feature generating networks for zero-shot learning, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5542–5551.

Yu, Y., Ji, Z., Fu, Y., Guo, J., Pang, Y., Zhang, Z.M., 2018a. Stacked semantics-guided attention model for fine-grained zero-shot learning, in: Advances in Neural Information Processing Systems, pp. 5995–6004.

Yu, Y., Ji, Z., Guo, J., Zhang, Z., 2018b. Zero-shot learning via latent space encoding. IEEE transactions on cybernetics 49, 3755–3766.

Yu, Y., Ji, Z., Zhang, Z., Han, J., 2020. Episode-based prototype generating network for zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Zhang, H., Koniusz, P., 2018. Zero-shot kernel learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 7670–7679.

Zhang, H., Long, Y., Guan, Y., Shao, L., 2019a. Triple verification network for generalized zero-shot learning. IEEE Transactions on Image Processing 28, 506–517.

Zhang, H., Long, Y., Liu, L., Shao, L., 2019b. Adversarial unseen visual feature synthesis for zero-shot learning. Neurocomputing 329, 12–20.

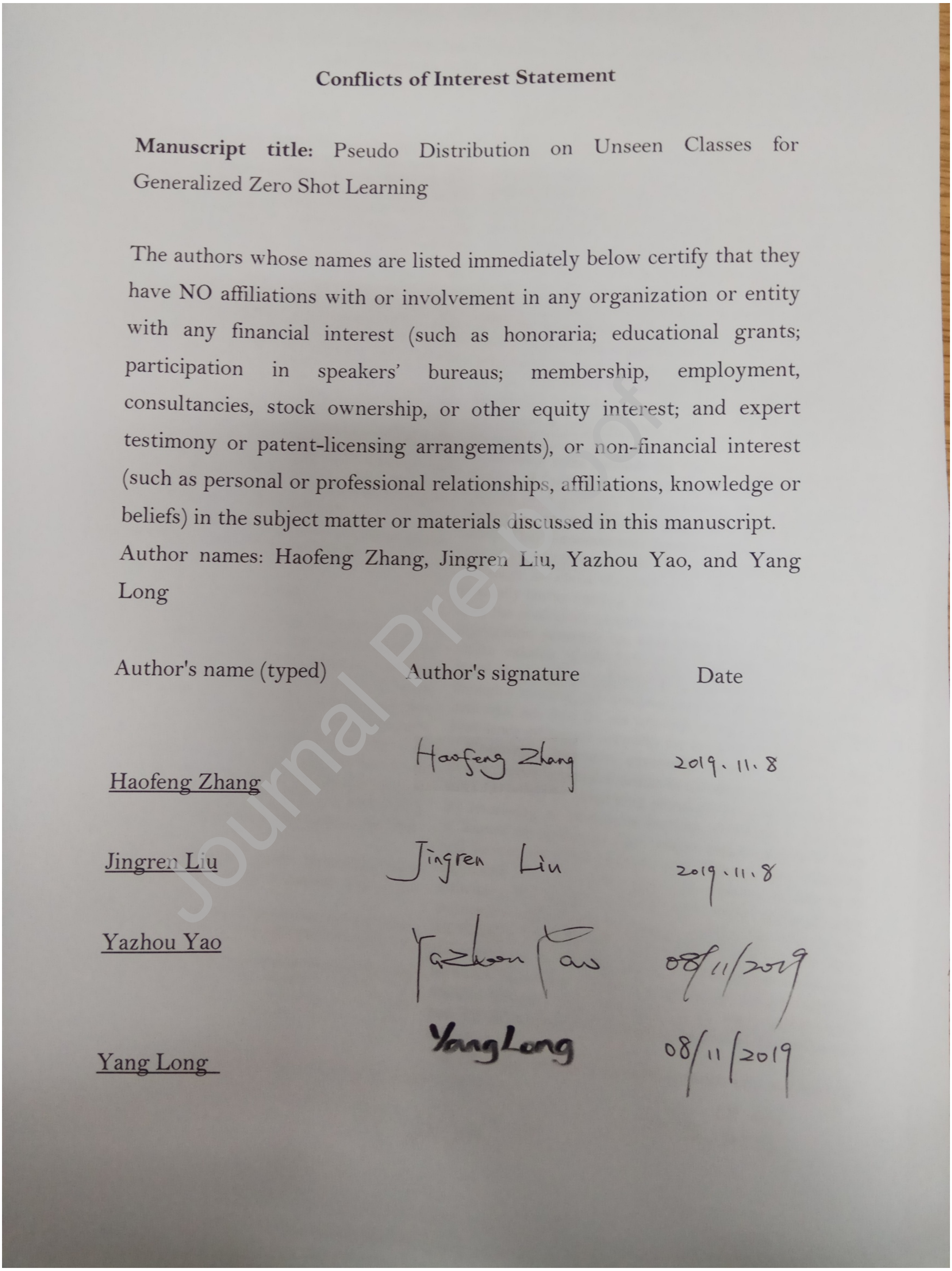
Zhang, H., Long, Y., Shao, L., 2019c. Zero-shot hashing with orthogonal projection for image retrieval. Pattern Recognition Letters 117, 201–209.

Zhang, H., Long, Y., Yang, W., Shao, L., 2019d. Dual-verification network for zero-shot learning. Information Sciences 470, 43–57.

Zhang, L., Xiang, T., Gong, S., 2017. Learning a deep embedding model for zero-shot learning, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 2021–2030.

Zhang, Z., Saligrama, V., 2015. Zero-shot learning via semantic similarity embedding, in: International Conference on Computer Vision, pp. 4166–4174.

Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A., 2018. A generative adversarial approach for zero-shot learning from noisy texts, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1004–1013.



Conflicts of Interest Statement

Manuscript title: Pseudo Distribution on Unseen Classes for Generalized Zero Shot Learning

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names: Haofeng Zhang, Jingren Liu, Yazhou Yao, and Yang Long

Author's name (typed)	Author's signature	Date
Haofeng Zhang	Haofeng Zhang	2019.11.8
Jingren Liu	Jingren Liu	2019.11.8
Yazhou Yao	Yazhou Yao	08/11/2019
Yang Long	YangLong	08/11/2019