Transductive Zero-Shot Learning via Visual Center Adaptation

Ziyu Wan,[*1,2] **Yan Li,**[*1] **Min Yang,**[3] **Junge Zhang**[†1]
[1]CRISE, Institute of Automation, Chinese Academy of Sciences
[2]Department of Computer Science, City University of Hong Kong
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
ziyuwan2-c@my.cityu.edu.hk, yan.li@cripac.ia.ac.cn
min.yang1129@gmail.com, jgzhang@nlpr.ia.ac.cn

Abstract

In this paper, we propose a Visual Center Adaptation Method (VCAM) to address the *domain shift* problem in zero-shot learning. For the seen classes in the training data, VCAM builds an embedding space by learning the mapping from semantic space to some visual centers. While for unseen classes in the test data, the construction of embedding space is constrained by a symmetric Chamfer-distance term, aiming to adapt the distribution of the synthetic visual centers to that of the *real* cluster centers. Therefore the learned embedding space can generalize the unseen classes well. Experiments on two widely used datasets demonstrate that our model significantly outperforms state-of-the-art methods.

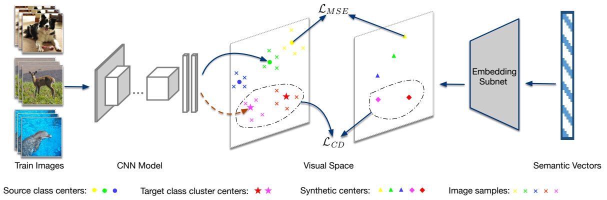

Figure 1: The illustration of the proposed visual center adaptation method.

Introduction

Remarkable success has been achieved by deep neural networks for visual object recognition on domains where a large number of labeled training data is available. Nevertheless, annotating sufficient data is labor-intensive and time-consuming, establishing significant barriers for adapting the learned systems to new domains. To tackle this problem, zero-shot learning (ZSL) has been proposed, which aims to learn recognition models for novel classes without labeled data. Generally, the ZSL approaches can be categorized into two types based on the usage of unlabeled data: inductive ZSL and transductive ZSL. In this paper, we focus on the ZSL with transductive setting in which the unlabeled (target) images from the target classes are available.

Despite the effectiveness of previous studies, ZSL is still challenged by the *domain shift* problem in practice. The source and target classes in ZSL are usually disjoint and even completely unrelated. In this case, applying naive projection function learned from source classes to target classes without any adaptation may lead to a large knowledge gap.

We propose a novel Visual Center Adaptation Method (VCAM) for ZSL. Inspired by (Zhang, Xiang, and Gong 2017), VCAM tries to project semantic information to the visual space to tackle the hubness problem (more details in *supplementary file*). To address the domain shift problem, we add a novel symmetric Chamfer-distance constraint to the learning of projection function in VCAM, which aims to perform structure alignment on target classes, more specially, to adapt the synthetic visual centers obtained using the learned projection function to the *real* cluster visual centers. By maintaining the structure of target classes during the learning of projection on source samples, our model is endowed with a much better generalization ability, which finally leads to the improvement of ZSL.

Our Methodology

Problem Definition We have N_s labeled source samples $\mathcal{D}_s \equiv \{(x_i^s, y_i^s)\}_{i=1}^{N_s}$, where x_i^s is an image and $y_i^s \in \mathcal{Y}_s = \{1, \ldots, S\}$ is the corresponding label. We are also given N_u unlabeled target samples $\mathcal{D}_u \equiv \{(x_i^u\}_{i=1}^{N_u}$ that are from target classes $\mathcal{Y}_u = \{S+1, \ldots, U\}$. The goal of ZSL is to build a recognition model that can predict the label $y_i^u \in \mathcal{Y}_u$ given x_i^u with no labeled training data for target classes. Here, each class $z \in \mathcal{Y}_s \cup \mathcal{Y}_u$ is associated with the pre-defined auxiliary attributes $a_z \in \mathcal{A}$ forming a semantic space. Note that we have $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$ according to the definition of ZSL.

Visual Center Adaptation Method Our VCAM is illustrated in Figure 1. Given the input image x, we use a CNN feature extractor $\phi(\cdot)$ to convert each image into a d-dimensional image representation $\phi(x) \in \mathcal{R}^{d\times 1}$. Motivated by the fact that the image features $\phi(x)$ of samples could form tight and disjoint clusters (Zhang and Saligrama 2016), we argue that each class should have a *real* visual center which is defined as the mean of all feature vectors in the corresponding class. Based on it, an embedding subnet is adopted to transfer the semantic attributes to these centers

[*] The first two authors contributed equally to this work.
[†] Junge Zhang is the corresponding author.

of the corresponding class:

$$c^{syn} = \sigma_2(w_2^T \sigma_1(w_1^T a)) \tag{1}$$

where a denotes the auxiliary attributes of each class. $\sigma_1(\cdot)$ and $\sigma_2(\cdot)$ denote non-linear operation (i.e., Leaky ReLU). w_1 and w_2 are the weights to be learned. c^{syn} is the predicted center for each category. To obtain the projection relation, we adopt the mean square error as the loss function, which minimizes the discrepancy between predicted centers c^{syn} and real centers c in the visual feature space for seen class:

$$\mathcal{L}_{MSE} = \frac{1}{S} \sum_{i=1}^{S} \|c_i^{syn} - c_i\|_2^2 + \lambda \Psi(w_1, w_2) \tag{2}$$

where $\Psi(\cdot)$ is the $l2$-norm parameter regularizer decreasing the model complexity, λ controls the tightness of the constraint and we empirically set $\lambda = 0.0005$.

During the testing phase, we first use Equation 1 to get synthetic center C_u^{syn} for target classes from their semantic attributes. Then for each image x_i, its classification result can be achieved by selecting the nearest synthetic center for it. Formally,

$$su^* = \underset{c_u^{syn}}{argmin} \ \|\phi(x_i) - c_u^{syn}\|_2 \tag{3}$$

However, in fact there is still discrepancy between c^{syn} and real centers for target classes while testing, i.e., domain shift problem, which will result in bad ZSL accuracy. To alleviate this problem, it is necessary to align the structure of the synthetic centers with that of the real centers for target class. Here, we use the class centers calculated by K-means to approximate the *real* centers. A symmetric Chamfer-distance constraint is proposed to measure the similarity between the two unordered high-dimensional point sets:

$$\mathcal{L}_{CD} = \sum_{x \in C_u^{syn}} \min_{y \in C_u^{clu}} \|x - y\|_2^2 + \sum_{y \in C_u^{clu}} \min_{x \in C_u^{syn}} \|x - y\|_2^2 \tag{4}$$

where C_u^{clu} indicates the cluster centers of target class obtained by K-means algorithm. C_u^{syn} represents the synthetic target centers obtained with the learned projection. Combining the above constraint, the final loss function to train VCAM is defined as:

$$\mathcal{L}_{VCAM} = \mathcal{L}_{MSE} + \beta \times \mathcal{L}_{CD} \tag{5}$$

where β controls the effect of these two objectives and we set to $\beta = 0.0005$ empirically.

Experiments

Datasets

We evaluate the effectiveness of the proposed VCAM on two representative ZSL benchmarks: Animals with Attributes2 (AwA2) and Caltech-UCSD Birds 200-2011 (CUB). AwA2 contains 37,322 images from 50 animals categories, where 40 of 50 classes are used for training and the rest are used for testing. For fair comparison with baseline methods, we also report the results on AwA1 that is an old version of animal datasets of ZSL without raw images. CUB is a fine-grained dataset with 200 different bird species and 11,788 images. We use 150 classes as training data and the rest 50 classes are used for testing. We also adopt the same ZSL data splits as used in (Xian, Schiele, and Akata 2017), called PS. We report the results on both the standard splits (SS) and the PS for fair comparison.

		AwA1		AwA2		CUB	
	Method	**SS**	**PS**	**SS**	**PS**	**SS**	**PS**
δ	SJE(2015)	76.7	65.6	69.5	61.9	55.3	53.9
	SYNC(2016)	72.2	54.0	71.2	46.6	54.1	55.6
	SCoRe(2017)	82.8	-	-	69.5	59.5	61.0
	LDF(2018)	83.4	65.8	-	-	70.3	69.2
	SE-ZSL(2018)	83.8	69.5	80.8	69.2	60.3	59.6
μ	UDA(2015)	73.2	-	-	-	39.5	-
	TMV(2015)	80.5	-	-	-	51.2	-
	SMS(2016)	78.4	-	-	-	59.2	-
	TSTD(2017)	90.3	-	-	-	58.2	-
	VCAM(ours)	**94.3**	**77.6**	**93.9**	**78.2**	**74.2**	**71.7**

Table 1: The experimental results in terms of MCA (%). Here, δ denotes inductive ZSL algorithm, μ denotes transductive ZSL algorithm, and "-" means no repeated result available yet.

Experimental Results

Following previous work (Li et al. 2018), the multi-way classification accuracy (MCA) is adopted as our evaluation metric. We summarize the experimental results in Table 1. Compared to the previous approaches, our method achieves significant improvements on all the experimental datasets, verifying the effectiveness of VCAM. For example, VCAM outperforms the best results of baseline methods by a margin of $2\% \sim 13\%$.

We also conduct challenging experiments in the generalized ZSL settings (gZSL) where the larger searching space is provided for testing to verify the effectiveness of our model in dealing with the domain shift problem in ZSL. In addition, we also visualize parts of the zero-shot classification results. Both the gZSL and visualization results are included in *supplementary file*.

Conclusion

In this paper, we have proposed a novel visual center adaptation method for zero-shot learning, which is based on the adaptation of visual centers to solve domain shift problem. Experiments show that VCAM outperforms other state-of-the-art methods on two representative datasets.

References

Li, Y.; Zhang, J.; Zhang, J.; and Huang, K. 2018. Discriminative learning of latent features for zero-shot recognition. In *CVPR*, 7463–7471.

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

Zhang, Z., and Saligrama, V. 2016. Zero-shot recognition via structured prediction. In *ECCV*, volume 9911, 533–548.

Zhang, L.; Xiang, T.; and Gong, S. 2017. Learning a deep embedding model for zero-shot learning. In *CVPR*, 2021–2030.