

Semantic Softmax Loss for Zero-Shot Learning

Zhong Ji, *Member, IEEE*, Yunxin Sun, Yulong Yu*, Jichang Guo, and Yanwei Pang, *Senior Member, IEEE*

Abstract—A typical pipeline for Zero-Shot Learning (ZSL) is to integrate the visual features and the class semantic descriptors into a multimodal framework with a linear or bilinear model. However, the visual features and the class semantic descriptors locate in different structural spaces, a linear or bilinear model can not capture the semantic interactions between different modalities well. In this letter, we propose a nonlinear approach to impose ZSL as a multi-class classification problem via a Semantic Softmax Loss by embedding the class semantic descriptors into the softmax layer of multi-class classification network. To narrow the structural differences between the visual features and semantic descriptors, we further use an L_2 normalization constraint to the differences between the visual features and visual prototypes reconstructed with the semantic descriptors. The results on three benchmark datasets, i.e., AwA, CUB and SUN demonstrate the proposed approach can boost the performances steadily and achieve the state-of-the-art performance for both zero-shot classification and zero-shot retrieval.

Index Terms—Zero-shot learning, semantic embedding, multi-class classification.

I. INTRODUCTION

Zero-Shot Learning (ZSL) [1], [3], [5], [19], [20], [9] aims at building classifiers to predict the unseen classes without any visual instances in the training stage. This task is achieved by transferring the information from seen classes to unseen ones with the knowledge about how each unseen class is semantically related to the seen classes. In order to measure the semantic relations between different classes, both the seen classes and unseen ones are represented as a high dimensional vector embedded in a semantic space. Such a space can be semantic attribute space or semantic word vector space.

Most of the existing ZSL approaches address this task as two different independent subtasks, which can be divided into two categories. The first one associates attribute prediction followed by classification inference [1], [16], [17]. One of the most popular among these approaches is the direct attribute prediction (DAP) approach [1], which predicts attributes independently using SVMs and infers zero-shot predictions by a maximum a posteriori rule that assumes attribute independence. The other one decomposes ZSL into a multimodal learning process and a similarity measurement process. To construct the interactions between the visual instances and the class semantic descriptors, exiting approaches either project the features from one modality to another [8], [10], [18] or project the features from both modalities into a common space [5], [6], [12], [19], [20]. To measure the similarity, most

approaches use nearest neighbour classifier (NN) [1], [5], [6] or label propagation [21].

Although existing approaches for ZSL have achieved impressive performances, they still suffer from issues below. (1) Most existing methods use a linear or bilinear approach to train the multimodal learning model that may not capture the semantic interactions between different modalities well. (2) Existing approaches perform ZSL as two disjoint subtasks, which leads to the information loss.

In this work, we present an end-to-end nonlinear embedding paradigm for ZSL based on the multi-class classification, as illustrated in Fig.1. Specifically, we embed the class semantic descriptors into a multi-class classification framework with the proposed Semantic Softmax Loss (SSL). It divides the classifier parameters into two matrices, a learned generative matrix and an off-the-shelf class semantic matrix. In this way, the visual instances, class semantic descriptors and the class labels are formulated into a unified multi-class classification model, which can be trained in an end-to-end way. We call the proposed method for ZSL as SSL-ZSL for short. Besides, the classification parameters for each class can be seen as a visual prototype reconstructed by the corresponding class semantic descriptor. We impose an L_2 normalization constraint to reconstruction task for semantic embedding so that the reconstructed prototypes preserve most of the information.

In summary, this paper contributes to the following aspects:

- We propose an end-to-end framework for ZSL by embedding the class semantic descriptors into the softmax layer in a multi-class classification pipeline, in which the compatibility between the class semantic descriptors and visual instances are optimized under the supervision of labels. In this way, the classifiers of unseen classes can be obtained with the semantic descriptors.
- To narrow the structural differences between the visual and the class semantic spaces, we add an L_2 normalization constraint on the visual features and reconstructed visual prototypes such that they lie on the same hypersphere.
- The performances of the proposed approach yield a consistent and significant boost on three benchmark ZSL datasets, namely AwA, CUB and SUN.

II. SEMANTIC SOFTMAX LOSS FOR ZERO-SHOT LEARNING

Given a training dataset with M images and their corresponding labels, a traditional classification model is trained to classify a given image to its correct label. In a typical

This work was supported by the National Basic Research Program of China (973 Program) under Grant 2014CB340400, the National Natural Science Foundation of China under Grants 61472273 and 61632018.

Z. Ji, Y. Sun, Y. Yu* (corresponding author), J. Guo and Y. Pang are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mails: {jizhong, sunyuxin, yuyunlong, jeguoguo, pyw}@tju.edu.cn).

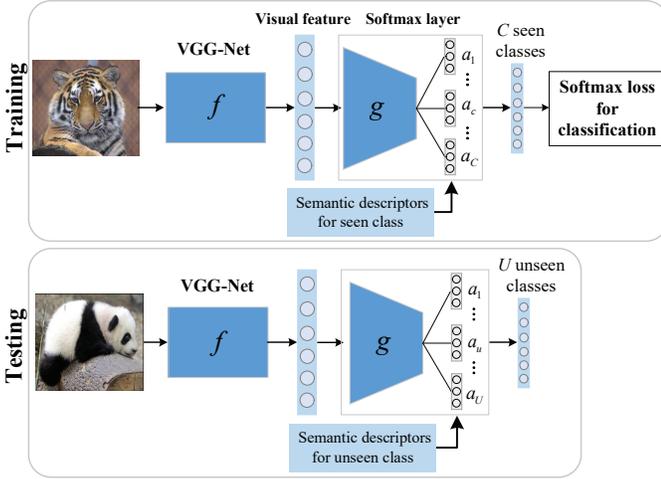


Fig. 1. The proposed pipeline for zero-shot learning. In the training stage, the images and the class semantic descriptors from seen classes are taken as input to predict the class label. The VGG-Net is adopted to extract the visual feature, and the class semantic descriptors are embedded in the softmax layer, f and g are models to be trained. In the testing stage, the test image and the class semantic descriptors of all candidate unseen classes are taken as input, and outputs the predicted label.

convolutional neural network (CNN), a softmax loss function is commonly used for training the network, given by Eq.(1)

$$L_S = -\frac{1}{M} \sum_{i=1}^M \log \frac{e^{W_{y_i}^T f(\mathbf{x}_i) + b_{y_i}}}{\sum_{j=1}^C e^{W_j^T f(\mathbf{x}_i) + b_j}}, \quad (1)$$

where C is the number of training classes, \mathbf{x}_i is the i^{th} instance. In the softmax loss, $f(\mathbf{x}_i)$ is usually the corresponding output of the penultimate layer of a CNN, y_i is the corresponding class label, and W and b are the weights and bias for the last layer of the network which act as a classifier.

For the seen classes, the classifier parameters $\{W, b\}$ can be obtained by training the network with training instances. For the unseen classes, however, no instances are available for training the classifiers. Consequently, this pipeline can not be applied for ZSL directly.

To address the above issue, existing ZSL approaches introduce the class semantic descriptors to transfer the knowledge from seen classes to unseen ones. Considering that the class semantic descriptor characterizes the properties of a class, it is reasonable to assume that the class classifier can be derived from its corresponding class semantic descriptor,

$$W_j = g(\mathbf{a}_j), \quad (2)$$

where \mathbf{a}_j and W_j are the semantic descriptor and classifier for j^{th} class, respectively. Function $g(\cdot)$ denotes the mapping between the class semantic descriptors and classifiers, which can be linear or nonlinear function. In this paper, we consider a simple linear model, i.e., $W_j = V^T \mathbf{a}_j$.

With the learned mapping function V , a class classifier can be deduced by Eq.(2) from its class semantic descriptor. Subsequently, ZSL can be performed as follows:

(1). The visual feature $f(\mathbf{x}_i)$ of a test image \mathbf{x}_i from an unseen class is first extracted using the pre-trained CNN, and normalized to unit length.

(2). The classification of \mathbf{x}_i is achieved by calculating the compatibility scores of the visual feature $f(\mathbf{x}_i)$ and the classifiers of all candidate unseen classes:

$$c(\mathbf{x}_i) = \arg \max_j S(f(\mathbf{x}_i), W_j), \quad (3)$$

where $S(f(\mathbf{x}_i), W_j)$ denotes the compatibility score between the instance \mathbf{x}_i and class j .

Usually, the compatibility score is computed with the inner product between the two vectors or by adapting the cosine similarity as given by Eq.(4):

$$S(\mathbf{x}_i, W_j) = \frac{f(\mathbf{x}_i)^T W_j}{\|f(\mathbf{x}_i)\|_2 \|W_j\|_2}. \quad (4)$$

From another view, W_j in Eq.(2) can be seen as the visual prototype that is reconstructed by the class semantic descriptor. Since each class is represented as a single semantic descriptor which is insufficient to fully represent what that class looks like. Consequently, even if the reconstructed visual prototype is learned by enforcing the class semantic descriptor to be projected to the center among its visual instances in the visual space, the classifiers will still struggle to assign the correct class labels.

To solve this issue, we impose the L_2 -norm of the features to be fixed for every training instance as well as the reconstructed visual prototypes. Specifically, we add an L_2 normalization constraint to the subtraction of the visual features and the reconstructed visual prototypes such that they lie on the same hypersphere. This approach has two advantages. Firstly, on a hypersphere, minimizing the softmax loss is equivalent to maximizing the cosine similarity for the positive pairs and minimizing it for the negative pairs, which strengthens the discriminative ability of the classifier. Secondly, the softmax loss is able to narrow the information loss caused by the different structure from different modalities, since all the features have same L_2 -norm.

The proposed semantic softmax loss is given by Eq.(5).

$$L_S = -\frac{1}{M} \sum_{i=1}^M \log \frac{e^{\mathbf{a}_{y_i}^T V f(\mathbf{x}_i) + b_{y_i}}}{\sum_{j=1}^C e^{\mathbf{a}_j^T V f(\mathbf{x}_i) + b_j}} + \lambda \|V\|_F^2, \quad (5)$$

$$s.t. \quad \|f(\mathbf{x}_i) - V^T \mathbf{a}_{y_i}\|_2 = \alpha, \quad \forall i = 1, 2, \dots, M,$$

where $\{V, b\}$ are the parameters to be trained, $\|V\|_F^2$ is the regularizer, and λ is the constant hyper-parameter.

Here, we provide the details of implementing the semantic embedding in Eq.(5) in the framework of multi-class classification. This module is added just after the penultimate layer of CNN which acts as a feature descriptor. The class semantic descriptors are embedded into the softmax layer based on the inner product. The parameters for softmax layer are derived from the class semantic descriptors and a compatible matrix that shared across all classes. Meanwhile, the difference between the visual feature and the reconstructed visual prototype are scaled to a hypersphere of a fixed radius with an L_2 -normalization layer.

Fixed α as a constant, the L_2 normalization constraint is added as a regularizer to the loss function. In this way, the

module is fully differentiable and can be used in an end-to-end training of the network. During training, we need to back-propagate the gradient of loss L_S through this module as well as the gradient with respect to the parameters $\{V, b\}$ in the proposed module. If the visual feature is abstracted with the pre-trained CNN in advance, $\{V, b\}$ are the only parameters to be trained.

III. EXPERIMENTS

In this section, we conduct zero-shot classification and zero-shot retrieval on three benchmark datasets, respectively, and compare the proposed approach with a number of ZSL approaches. We will show the superior performances of our approach against a number of state-of-the-art methods.

A. Datasets and Settings

Datasets. Three datasets are chosen for our evaluations, Animal with Attributes (AwA) [1], Caltech UCSD Birds (CUB) [2] and SUN attribute dataset [4]. AwA provides 30,475 images from 50 animal classes, and 85 associated class-level attributes. We follow the standard seen/unseen split [1], where 40 classes with 24,295 images are taken as the seen domain and the remaining 10 classes with 6180 images are adopted as the unseen domain. CUB dataset contains 11,788 images from 200 bird species with 312 associated attributes. In this dataset, we use the same zero-shot split as [5] with 150 classes for seen data and 50 disjoint classes for the unseen data. SUN dataset contains 717 scene categories annotated by 102 attributes, and each class has 20 images. In this dataset, we use 707 classes as the seen domain and the remaining 10 classes as the unseen domain, the same as that in [6].

Visual features and class semantic descriptors. To extract the visual feature for each image, we use the pre-trained VGG-Verdeep-16 model [9], where the output of the penultimate layer (before the softmax) is taken as the feature vector. With regard to class semantic descriptors, we use not only the class attributes associated with the datasets but also the word embeddings for each class. We train a word2vector [14] model on the Wikipedia corpus to obtain the 1000-dimensional word vector for each class name. Since few competitors use word embeddings for SUN dataset, we only extract the word vectors for AwA and CUB datasets for the experimental comparison.

1) *Zero-shot Classification:* For zero-shot classification, the model is first trained with the seen data, and then the test images are predicted to the candidate unseen classes with the trained model.

Competitors. We compare our proposed approach with 8 state-of-the-art approaches below:

- 1) **LR** [10] and **RLR** [11]. As a baseline method, Linear Regression (LR) [10] learns a mapping function to project the visual feature to the class semantic space. To alleviate the hubness problem that suffered by nearest neighbor search in a high dimensional space, Reverse

TABLE I

Results on three benchmark datasets in average per-class top-1 accuracy (%). We compare with approaches under different class semantic descriptors including attributes (A) and word vectors (W). ‘†’ denotes the methods are implied by ourselves. ‘-’ indicates that no experiments have been performed under this case in original paper.

Method	AwA		CUB		SUN
	A	W	A	W	A
LR† [10]	63.6	50.6	37.4	28.8	75
RLR† [11]	73.7	58.4	35.2	26.5	76
SSE [13]	76.3	-	30.4	-	82.5
SJE [5]	66.7	51.2	50.1	28.4	-
ESZSL† [12]	76.5	71.5	47.6	30.9	82.0
JLSE [6]	80.5	-	41.8	-	83.8
MLZSC [7]	77.3	-	43.3	-	84.4
MCME [8]	-	67.0	-	32.6	-
SSL-ZSL	82.69	72.02	55.72	33.33	88.00

Linear Regression (RLR) [11] learns a reverse projection to project the class semantic descriptors to the visual space.

- 2) **ESZSL** [12] and **SJE** [5]. Embarrassingly Simple ZSL (ESZSL) [12] is a simple but effective approach that integrates the compatibility scores and class labels into a linear framework, where the compatibility scores are the similarities between the visual feature and class semantic descriptor obtained with a bilinear formulation. Likewise, Structured Joint Embedding (SJE) [5] also uses bilinear compatibility function to associate the visual and class semantic descriptors and adopts a weighted approximate ranking loss inspired from the structured SVM [15].
- 3) **SSE** [13] and **JLSE** [6]. Semantic Similarity Embedding (SSE) [13] and Joint Latent Similarity Embedding (JLSE) [6] express visual images and class semantic descriptors as a mixture of seen class proportions. Specifically, SSE leverages the similar class relationships both in visual and class semantic space and JLSE poses both the visual images and class semantic descriptors into a latent space where the semantic information matches.
- 4) **MLZSC** [7] and **MCME** [8]. Metric Learning for Zero-Shot Classification (MLZSC) [7] formulates zero-shot classification as a metric learning problem via improving semantic embedding consistency. Manifold Regularized Cross-Modal Embedding (MCME) [8] improves the cross modal embedding ability with an effective manifold regularizer.

Evaluation Criteria. We average the correct prediction independently for each class before dividing the number of classes, i.e., the average per-class top-1 accuracy, which is popular for zero-shot classification.

Comparison Results. Table I presents the comparative results of SSL-ZSL on three datasets. It is worth mentioning that SJE [5] extracts GoogleNet features as image representations, the others all use VGG features. From the results, we can observe that our proposed approach achieves the best results on all datasets. Specifically, it has an impressive gains over the other state-of-the-art methods ranging from 0.52% to 5.62% in different datasets with different class semantic descriptors.

TABLE II

The classification performances (%) with and without L_2 normalization constraint on three datasets. SSL-ZSL/with and SSL-ZSL/without denote the methods with and without L_2 normalization constraint, respectively.

Method	AwA	CUB	SUN
SSL-ZSL/without	80.92	50.91	86.5
SSL-ZSL/with	82.69	55.72	88.0
Improvement	1.77	4.81	1.5

Besides, the proposed approach has overwhelming superiority than ESZSL [12], which is a similar approach with linear model. This indicates the superiority of our proposed nonlinear model.

The impact of L_2 normalization constraint. We also conduct experiments to verify the effectiveness of the L_2 normalization constraint. We list the classification results of with and without L_2 normalization constraint in Table II. We can find that the L_2 normalization constraint has large gains of on three datasets. More specifically, it brings 1.77%, 4.81% and 1.5% improvements on AwA, CUB and SUN with attributes, respectively. For displayed directly, we also give a visualization of unseen instances from AwA dataset with and without L_2 normalization constraint, as illustrated in Fig. 2. As we can see, the reconstructed visual prototypes with L_2 normalization constraint are closer to the centers of respective classes than those of without L_2 normalization constraint.

2) *Zero-shot Retrieval*: Given a specified class semantic descriptors of unseen classes, the task of Zero-shot Retrieval (ZSR) is to search some visual images from an image database related to it. In the experiment, the model is first trained with the seen instances and the class semantic descriptors of unseen classes are then taken as queries to rank the images from unseen classes based on the similarity with the specified query.

We select three existing state-of-the-art ZSR approaches which are published in the past two years for comparison. Table III presents the comparative results for mAP on three benchmark datasets. We can find that the proposed approach achieves the best performances on all datasets. Specifically, SSL outperforms the state-of-the-art methods in 5.52%, 15.79% and 6.14% on AwA, CUB and SUN datasets, respectively. Besides, the proposed SSL achieves 68.23% average on three datasets, which has a 9.29% gain than the runner up [6]. We argue that the superior performances benefit from our proposed effective optimization model that narrows the structural differences between the visual space and class semantic space.

Fig.3 shows the precision-recall curves for unseen classes on three datasets. Specifically, we provide the precision-recall curves with attribute and word vectors for AwA dataset. As to CUB dataset, we only show the first 10 classes from 50 unseen classes for the convenience of display. Compared with the precision-recall curves in original paper [7], our approach obviously performs superior for most classes on AwA and SUN datasets and has a larger area under the curves for CUB dataset. We also can find that the areas under the curves on AwA dataset are larger than those on CUB and SUN datasets. This is because CUB and SUN are fine-grained datasets which

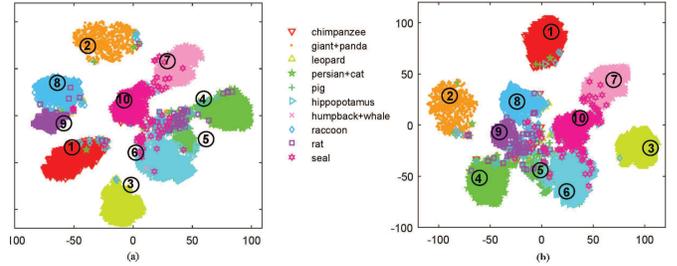


Fig. 2. t-SNE visualization of unseen instances from AwA dataset. The black circles indicate the reconstructed visual prototypes with the corresponding class semantic descriptors. (a) denotes the visualization without normalization constraint while (b) denotes the visualization with normalization constraint. Best view in color.

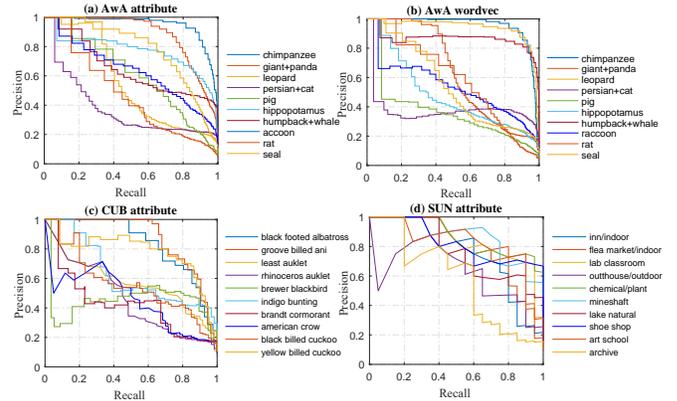


Fig. 3. Precision-Recall curves for unseen classes on three datasets. For AwA, we plot the curves with attribute (a) and word vectors (b), respectively. For CUB dataset, we show the first 10 classes from 50 unseen classes. Best viewed in color.

are more challenging than AwA dataset.

IV. CONCLUSION

We have proposed an end-to-end approach for zero-shot learning in which the semantic descriptors are embedded into the softmax layer in a multi-class classification framework. To narrow the structural differences between different modalities, an L_2 normalization constraint is introduced to imposed on the differences of visual features and the visual prototypes reconstructed with class semantic descriptors. We have shown experimentally that our method outperforms the state-of-the-arts methods both for zero-shot classification and zero-shot retrieval on AwA, CUB and SUN datasets, respectively.

TABLE III

Zero-shot retrieval mAP (%) comparison on three benchmark datasets. The results of the selective comparative methods are cited from the original papers.

Method	AwA	CUB	SUN	Ave.
SSE [13]	46.25	4.69	58.94	36.62
JLSE [6]	67.66	29.15	80.01	58.94
MLZSC [7]	68.1	25.33	52.68	48.69
SSL-ZSL	73.62	44.94	86.15	68.23

REFERENCES

- [1] C. H. Lampert, H. Nickisch, and S. Harmeling, "Learning to detect unseen object classes by between-class attribute transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 951-958.
- [2] C. Wah, and S. Branson, P. Welinder, P. Perona and S. Belongie, "The Caltech-UCSD Birds-200-2011 Dataset," in *California Institute of Technology*, 2010.
- [3] J. Qin, Y. Wang, L. Liu, J. Chen and L. Shao "Beyond Semantic Attributes: Discrete Latent Attributes Learning for Zero-Shot Recognition," in *IEEE Signal Process. Letters*, vol. 23. no. 11, pp. 1667-1671, Nov. 2016.
- [4] G. Patterson, C. Xu, H. Su and J. Hays, "The sun attribute database: Beyond categories for deeper scene understanding," in *Int. Jour. of Comput. Vis.*, vol. 108, no. 4, pp. 59-81, May 2014.
- [5] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, "Evaluation of output embeddings for fine-grained image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2927-2936.
- [6] Z. Zhang, V. Saligrama, "Zero-shot learning via joint latent similarity embedding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 6034-6042.
- [7] M. Bucher, S. Herbin, and F. Jurie "Improving semantic embedding consistency by metric learning for zero-shot classification," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 730-746.
- [8] Z. Ji, Y. Yu, Y. Pang, Z. Guo and Z. Zhang, "Manifold regularized cross-modal embedding for zero-shot learning," in *Inf. Sci.*, vol. 378, no. 2, pp. 48-58, Feb. 2017.
- [9] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Comput. Sci.*, 2015.
- [10] A. Lazaridou, E. Bruni and M. Baroni, "Is this a wampimuk? Cross-modal mapping between distributional semantics and the visual world," in *Association for Comput. Lingui.*, 2014, pp. 1403-1414.
- [11] Y. Shigeto, I. Suzuki, K. Hara, M. Shimbo, and Y. Matsumoto. "Ridge Regression, Hubness, and Zero-Shot Learning," in *Joint Eur. Conf. on Mach. Learn. and Know. Discovery in Data.*, 2015, pp. 135-151.
- [12] B. Romera-Paredes and P. H. Torr. "An embarrassingly simple approach to zero-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2152-2161.
- [13] Z. Zhang and V. Saligrama, "Zero-Shot learning via semantic similarity embedding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4166-4174.
- [14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Int. Conf. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111-3119.
- [15] I. Tsochantaris, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," in *Jour. of Mach. Learn. Research*, vol. 6, no. 2, pp. 1453-1484, Sep. 2005.
- [16] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, "Describing objects by their attributes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1778-1785.
- [17] X. Yu and Y. Aloimonos, "Attribute-based transfer learning for object categorization with zero/one training example," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 127-140.
- [18] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. "Zero-shot learning through cross-modal transfer," in *Int. Conf. Adv. Neural Inf. Process. Syst.*, 2013, pp. 935-943.
- [19] Y. Fu, T. M. Hospedales, T. Xiang, "Transductive multi-view zero-shot learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2332-2345, Nov. 2015.
- [20] Y. Yu, Z. Ji, X. Li, "Transductive Zero-Shot Learning with a Self-training dictionary approach," arXiv preprint arXiv:1703.08893, 2017.
- [21] E. Kodirov, T. Xiang, Z. Fu, "Unsupervised Domain Adaptation for Zero-Shot Learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2452-2460.