Deep Semisupervised Zero-Shot Learning with Maximum Mean Discrepancy

Lingling Zhang
zhanglingling@stu.xjtu.edu.cn
Jun Liu
liukeen@mail.xjtu.edu.cn
Minnan Luo
minnluo@mail.xjtu.edu.cn
MOEKLINNS Lab, Department of Computer Science and Technology,
Xi'an Jiaotong University, 710049, China

Xiaojun Chang
cxj273@gmail.com
Centre for Quantum Computation and Intelligent Systems, University
of Technology Sydney, Ultimo NSW 2007, Australia

Qinghua Zheng
qhzheng@mail.xjtu.edu.cn
MOEKLINNS Lab, Department of Computer Science and Technology,
Xi'an Jiaotong University, 710049, China

Due to the difficulty of collecting labeled images for hundreds of thousands of visual categories, zero-shot learning, where unseen categories do not have any labeled images in training stage, has attracted more attention. In the past, many studies focused on transferring knowledge from seen to unseen categories by projecting all category labels into a semantic space. However, the label embeddings could not adequately express the semantics of categories. Furthermore, the common semantics of seen and unseen instances cannot be captured accurately because the distribution of these instances may be quite different. For these issues, we propose a novel deep semisupervised method by jointly considering the heterogeneity gap between different modalities and the correlation among unimodal instances. This method replaces the original labels with the corresponding textual descriptions to better capture the category semantics. This method also overcomes the problem of distribution difference by minimizing the maximum mean discrepancy between seen and unseen instance distributions. Extensive experimental results on two benchmark data sets, CU200-Birds and Oxford Flowers-102, indicate that our method achieves significant improvements over previous methods.

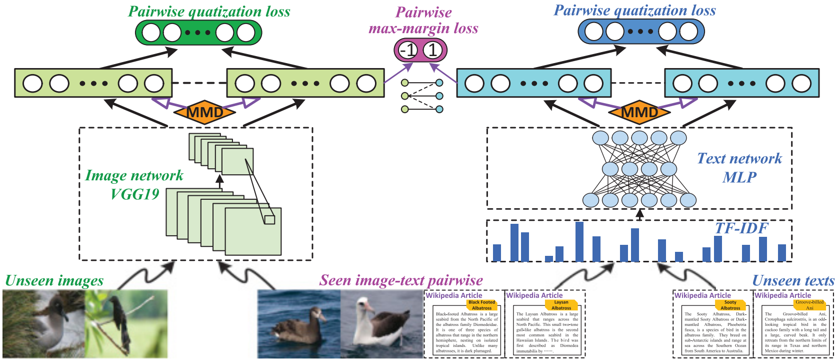
Neural Computation **30, 1426–1447** (2018)
doi:10.1162/neco_a_01071

1 Introduction

Image data are booming along with the development of the Internet, and the categories of images are also dramatically increasing. Collecting adequate labeled images for each category is difficult because of two limitations: only a few images are available for most of the visual categories, and prior work on image data annotation, which is tedious and time-consuming, might have a serious effect on algorithm performance because of poor quality. Inspired by the human ability to recognize a new category without ever seeing a visual instance, zero-shot learning (Palatucci, Pomerleau, Hinton, & Mitchell, 2009) has attracted increasing interest. Zero-shot learning is a special case of classification in which unseen categories do not have any labeled instances in the training stage. It aims to improve the scalability of traditional classification by exploiting shared knowledge between seen and unseen categories (Yu & Aloimonos, 2010). Zero-shot learning has been applied to face verification (Kumar, Berg, Belhumeur, & Nayar, 2011), image annotation (Kovashka, Vijayanarasimhan, & Grauman, 2011; Peng, Wu, & Ernst, 2017), and image retrieval (Kovashka, Parikh, & Grauman, 2012; Scheirer, Kumar, Belhumeur, & Boult, 2012), among other applications (Kansky et al., 2017).

Two main strategies are employed to accomplish zero-shot learning. The first focuses on transferring knowledge from seen to unseen categories by attribute sharing (Lampert, Nickisch, & Harmeling, 2014). Common attributes such as the color, shape, and other properties of visual objects are artificially constructed by several experts. These attribute-based methods are limited because of the challenging prior work on constructing attribute space. For this issue, the second group of strategies mines the correlation between seen and unseen categories by projecting entire category labels into a semantic space and then learns the visual classifier for unseen categories. These methods are more practical than attribute-based methods because they do not require constructing the attribute space manually. Therefore, these methods based on label embedding have been attracting attention. Implementing zero-shot learning by label embedding, however, is still a significant challenge because the category labels are insufficient to represent the semantics of seen and unseen categories. However, the textual descriptions of visual categories could adequately describe their characteristics, and these descriptions are widely available and relatively affordable. For example, Wikipedia currently contains more than 5 million articles.[1] Therefore, utilizing the rich information from textual descriptions of seen and unseen categories is a preferable strategy for accomplishing zero-shot learning (Elhoseiny, Saleh, & Elgammal, 2013; Lei Ba, Swersky,

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



Black-footed Albatross is a large seabird from the North Pacific of the albatross family Diomedeidae. It is one of three species of albatross that range in the northern hemisphere, nesting on isolated tropical islands. Unlike many albatrosses, it is dark plumaged.

The Laysan Albatross is a large seabird that ranges across the North Pacific. This small two-tone gull-like albatross is the second most common seabird in the Hawaiian Islands. The bird was first described as Diomedea immutabilis by ······.

The Sooty Albatross, Dark-mantled Sooty Albatross or Dark-mantled Albatross, Phoebetria fusca, is a species of bird in the albatross family. They breed on sub-Antarctic islands and range at sea across the Southern Ocean from South America to Australia.

The Groove-billed Ani, Crotophaga sulcirostris, is an odd-looking tropical bird in the cuckoo family with a long tail and a large, curved beak. It only retreats from the northern limits of its range in Texas and northern Mexico during winter.

Figure 1: The architecture of deep cross-modal network for semisupervised zero-shot learning with MMD.

Fidler, & Salakhutdinov, 2015). In addition, the previous studies on zero-shot learning have the limitation of grasping the common properties of seen and unseen categories. In fact, the label embedding distributions of seen and unseen categories may be quite different even in the same dimensional semantic space (Long et al., 2013; Zhang, Yu, Chang, & Wang, 2015). In summary, it is necessary to mine the category's semantics better using the corresponding textual descriptions and decrease the distribution difference of seen and unseen instances in a theoretical manner.

Consequently, we propose an end-to-end deep zero-shot classification method by jointly considering the heterogeneity gap between cross-modalities and the correlation among unimodal instances in a semisupervised framework. This method uses the textual descriptions acquired from Wikipedia to replace the image labels for representing the category's semantics better. This method also alleviates the distribution difference of seen and unseen instances by using the distribution-matching strategy maximum mean discrepancy (MMD) because of its efficiency in computing and optimizing. For better understanding, we illustrate the proposed deep cross-modal network (DCMN) in Figure 1. Through deep nonlinear mapping, the original images and texts lying in heterogeneous feature spaces are projected into a shared embedding space. In the shared space, we bridge the semantic gap between image and text modality by minimizing the max-margin loss. Then we design two unimodal quantization losses for images and texts, respectively. The image quantization loss aims to reduce differences in image representations in the same category; the text quantization loss is constructed to capture the semantic correlation between seen and unseen categories. Finally, we use two MMD constraints to guarantee the distribution similarity of seen and unseen instances. The contributions of this letter are summarized as follows:

1. A novel deep cross-modal network is proposed for semisupervised zero-shot learning, where the cross-modal max-margin loss and two unimodal losses are combined to transfer knowledge from seen to unseen instances.
2. We use the MMD constraints to decrease the distribution difference of seen and unseen instances efficiently for capturing their commonality better.
3. An efficient alternative algorithm is proposed to optimize the deep cross-modal network, which is consistent with traditional backpropagation (BP) in theory.
4. We conduct extensive experiments on two benchmark data sets (CU200-Birds and Oxford Flowers-102) to illustrate the effectiveness and superiority of our method. The experimental results indicate that our method consistently outperforms other methods.

The remainder of this letter is organized as follows. Related work on zero-shot learning is introduced in section 2. A novel deep cross-modal network to solve the semisupervised zero-shot learning problem is proposed in section 3. An efficient alternative algorithm is proposed in section 4 to solve the optimization problem. Extensive experiments on two benchmark data sets are conducted in section 5. The conclusions are given in section 6.

2 Related Work

The key point of zero-shot learning is sharing common knowledge between seen and unseen categories (Xian, Schiele, & Akata, 2017), in which attribute learning and label embedding are two main streams for this task. Recently, several studies have tried to use textual descriptions of visual categories to implement zero-shot learning.

2.1 Attribute Learning.

Attribute learning is a main avenue for research on zero-shot learning, which transfers knowledge from seen to unseen categories by attribute sharing (Lampert, Nickisch, & Harmeling, 2009). The seen and unseen visual category labels are all represented as attribute vectors, in which the attributes consist of shape, color, or some other geographic information. In the training stage, the attribute-based methods capture the semantic relationship between attributes and seen images and then obtain the attribute classifier. In the testing stage, the classifier can be used to derive new attribute vectors for unseen images. After that, the scores between unseen categories and unseen images could be calculated and we could obtain labels for unseen images. Specifically, Liu, Zhang, and Chen (2014) proposed a unified framework to learn the attribute-attribute relations and the attribute classifier jointly and then to boost the performance of attribute predictor. Parikh and Grauman (2011) presented a novel method for learning relative attributes and explained how to use relative attributes to enhance the accuracy of zero-shot learning. For these attribute-based methods, the attributes are manually defined by humans, which remains a challenging prior work for zero-shot learning.

2.2 Label Embedding. Another notable body of zero-shot learning belongs to label embedding. In contrast to attribute learning, these methods project independent category labels into a semantic embedding space. After that, the semantic relationships among all labels can be characterized, and thus supervised knowledge can be transferred from seen to unseen categories. The results of these methods are as follows: in the embedding space, the representation of the seen image is close to that of its corresponding label, and similar labels would also have similar representations in the space. The representation of a cat image is close to the "cat" label in the semantic space. The distance of "cat" and "dog" labels is closer compared to the distance of "cat" and "fish" labels because the cat species is more similar to dog than to fish. Norouzi et al. (2013) mapped images into the semantic embedding space via convex combination of the label embeddings. Zhang and Saligrama (2015) developed a novel semantic similarity embedding (SSE) method based on a max-margin framework for zero-shot recognition. Fu, Xiang, Kodirov, and Gong (2015) proposed to model the semantic manifold using a semantic class label graph. Fu, Hospedales, Xiang, and Gong (2015) proposed a transductive multiview embedding space for exploiting the multiple semantic representations of visual data. Fu and Sigal (2016) also proposed using an open set semantic vocabulary to train the classifiers for seen and unseen categories in supervised learning. In recent years, many researchers have focused on utilizing deep learning methods to mine the semantic embedding space effectively (Frome et al., 2013; Socher, Ganjoo, Manning, & Ng, 2013; Dauphin, Tur, Hakkani-Tur, & Heck, 2013). Compared with the previous attribute-based methods, label embedding methods are more practical because they do not require constructing the attribute space artificially. However, label embeddings are insufficient to represent the semantics of the categories. Moreover, the distribution difference of seen and unseen labels brings the limitation of capturing their commonality for zero-shot learning.

2.3 Visual Categories with Text Descriptions. Leveraging some textual descriptions for visual categories has been proved to be powerful at zero-shot learning tasks (Shojaee & Baghshah, 2016). Compared with the original simple image tags or captions, abundant textual descriptions are capable of detecting the comprehensive properties of seen and unseen categories. The representative work in Elhoseiny et al. (2013) captured the information between visual and textual domains by combining a regression function and a knowledge transfer function to implement zero-shot learning. Related work in Lei Ba et al. (2015) conceived of a deep zero-shot classification convolutional neural network in which the weights of the classifier are generated according to the abundant textual descriptions. Compared with previous work, these methods could capture more abundant correlations between seen and unseen categories.

3 Deep Cross-Modal Network

In the framework of semisupervised zero-shot learning, there is a set of n training images $\mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ over c classes. The first n_s ($n_s < n$) instances $\mathbf{x}_1$, $\mathbf{x}_2, \ldots, \mathbf{x}_{n_s}$ are labeled images with class labels in the first c_s seen classes. The remaining n_u images, $\mathbf{x}_{n_s+1}$, $\mathbf{x}_{n_s+2}, \ldots, \mathbf{x}_n$, are unlabeled instances whose labels belong to the remaining c_u unseen. Note that seen classes and the unseen classes are disjoint. We use $X_s = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_s}\}$ and $X_u = \{\mathbf{x}_{n_s+1}, \mathbf{x}_{n_s+2}, \ldots, \mathbf{x}_n\}$ to denote the seen and unseen image sets, respectively. In this letter, we perform image classification over the unseen classes using additional textual descriptions corresponding to all c categories. The textual description set $\mathbb{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_c\}$ matches c visual classes one by one. $\mathbf{t}_j \in \mathbb{R}^m$ ($j = 1, 2, \ldots, c$) refers to the m-dimensional feature vector of the textual description for the jth visual category. The seen and unseen class descriptions are denoted as $T_s = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{c_s}\}$ and $T_u = \{\mathbf{t}_{c_s+1}, \mathbf{t}_{c_s+2}, \ldots, \mathbf{t}_c\}$, respectively.

3.1 Cross-Modal Network Architecture. As shown in Figure 1, we propose a novel end-to-end deep cross-modal network (DCMN) for semisupervised zero-shot learning. The framework comprises two unimodal deep networks, an image network and a text network, which map entire image instances $\mathbb{X}$ and corresponding textual descriptions $\mathbb{T}$ to a K-dimensional shared semantic embedding space.

3.1.1 Image Network Architecture. The image network on the left of Figure 1 is exploited to accept all images $\mathbb{X}$ and then maps them into a cross-modal embedding space. We use a deep convolutional neural network (CNN) as the image network because of its high performance on object recognition tasks (Lawrence, Giles, Tsoi, & Back, 1997; Zhou, 2016). We start with the 19-layer convolutional network (VGG19) (Simonyan & Zisserman, 2014), which is composed of sixteen 3×3 convolutional layers ($conv1$–$conv16$) and three fully connected layers ($fc17$–$fc19$). The rectifier linear units (ReLU; Glorot, Bordes, & Bengio, 2011), $a(\mathbf{x}) = \max(\mathbf{0}, \mathbf{x})$, are taken as the activation function for the hidden layers $conv1$–$fc19$. After that, we replace the layer of the softmax classifier in the original VGG19 with a new feature map, fcm. The fcm layer recodes the image features from the $fc19$ layer to the new K-dimensional semantic representations. We compress the output of fcm in the range $[-1, 1]$ by using the hyperbolic tangent (tanh) activation: $a(\mathbf{x}) = \tanh(\mathbf{x})$. In this case, the semantic representation $\mathbf{p}_i$ for image $\mathbf{x}_i$ can be computed as

$$\mathbf{p}_i = \tanh\left(f(\mathbf{x}_i;\theta_x); W_x^{\ell}\right), \tag{3.1}$$

where W_x^{ℓ} represents the weight and bias parameter of the fcm layer. $\theta_x = \{W_x^{\ell-19}, \ldots, W_x^{\ell-2}, W_x^{\ell-1}\}$ denotes the parameters of $conv1$–$fc19$ sequentially. The $f(\mathbf{x}_i;\theta_x)$ is the output of the $fc19$ layer for image $\mathbf{x}_i$. We denote the semantic representations of all images as $P = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n] \in \mathbb{R}^{K\times n}$, where the seen and unseen images are represented as $P_s = [\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{n_s}] \in \mathbb{R}^{K\times n_s}$ and $P_u = [\mathbf{p}_{n_s+1}, \mathbf{p}_{n_s+2}, \ldots, \mathbf{p}_n] \in \mathbb{R}^{K\times n_u}$, respectively.

3.1.2 Text Network Architecture. Each textual description $\mathbf{t}_j$ illustrated on the right of Figure 1, passes the multilayer perceptrons (MLP) to be represented as a fixed-length vector in the cross-modal embedding space. The MLP network is designed with two fully connected layers, $fc1$ and fcm. Note that the fcm layer maps text features from the $fc1$ layer into K-dimensional semantic space. The tanh activation function is still used for the fcm layer to control the output range in $[-1, 1]$. Then we obtain the semantic representation $\mathbf{q}_j$ for text $\mathbf{t}_j$ as

$$\mathbf{q}_j = \tanh\left(g(\mathbf{t}_j;\theta_t); W_t^{\ell}\right), \tag{3.2}$$

where θ_t and W_t^{ℓ} denote the parameters of the $fc1$ and fcm layers, respectively. The $g(\mathbf{t}_j;\theta_t)$ is the output of the $fc1$ layer for text $\mathbf{t}_j$. After that, the semantic representations of all texts are represented as $Q = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_c] \in \mathbb{R}^{K\times c}$. The $Q_s = [\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_{c_s}] \in \mathbb{R}^{K\times c_s}$ and $Q_u = [\mathbf{q}_{c_s+1}, \mathbf{q}_{c_s+2}, \ldots, \mathbf{q}_c] \in \mathbb{R}^{K\times c_u}$ denote the representations of seen and unseen texts respectively.

With the inputs of images $\mathbb{X}$ and texts $\mathbb{T}$ for DCMN, we accomplish representation learning for seen and unseen instances by minimizing the following loss function,

$$\min_{\theta, S_u^t} L_f + \lambda(L_x + L_t), \tag{3.3}$$

where $\theta = \{\theta_x, \theta_t, W_x^{\ell}, W_t^{\ell}\}$ denotes the parameter set of the cross-modal network. $S_u^t \in \{-1, 1\}^{n_u\times n_u}$ is the similarity matrix for unseen images. The trade-off parameter λ controls the relative importance of unimodal loss $(L_x + L_t)$ regarding cross-modal loss L_f. The loss function, equation 3.3, could be divided into two parts:

- L_f (section 3.2): The cosine max-margin loss L_f uses the information from labeled seen image-text pairs to narrow the semantic gap between image and text modality.
- L_x and L_t (section 3.3): There are two unimodal losses: image modal loss and text modal loss. For image modal loss L_x, we not only guarantee that the visual instances from the same category are close in the embedding space, but also reduce the distribution difference of seen

and unseen images. For text modal loss L_t, we maintain the semantic correlation between distinct categories in the text network and decrease the distribution difference of seen and unseen texts.

In the following sections, we demonstrate the cross-modal max-margin loss L_f and two unimodal losses L_x and L_t in detail.

3.2 Cross-Modal Max-Margin Loss L_f. Through deep nonlinear mapping of image and text networks, the similarity between seen image $\mathbf{x}_i$ and seen text $\mathbf{t}_j$ can be obtained by computing the cosine distance in the shared embedding space,

$$\cos(\mathbf{p}_i, \mathbf{q}_j) = \frac{\mathbf{p}_i^\top \mathbf{q}_j}{\|\mathbf{p}_i\| \|\mathbf{q}_j\|}, \tag{3.4}$$

where $\| \cdot \|$ denotes the Euclidean norm of a vector. To guarantee the learned representations of an image and its corresponding textual description to be similar in the embedding space, we follow the intuitive strategy used in Cao, Long, Wang, Yang, and Yu (2016) to minimize the following cosine max-margin loss,

$$L_f = \sum_{i=1}^{n_s} \sum_{j=1}^{c_s} \max \left(0, \Delta - h_{ij} \frac{\mathbf{p}_i^\top \mathbf{q}_j}{\|\mathbf{p}_i\| \|\mathbf{q}_j\|} \right), \tag{3.5}$$

where $\Delta > 0$ is the margin parameter. For the seen image-text pair $(\mathbf{x}_i, \mathbf{t}_j, h_{ij})$, $h_{ij} = 1$ if the seen image $\mathbf{x}_i$ belongs to the seen class $\mathbf{t}_j$; otherwise, $h_{ij} = -1$. Obviously, equation 3.5 can guarantee that the image $\mathbf{p}_i$ and text $\mathbf{q}_j$ are similar if $h_{ij} = 1$, whereas $\mathbf{p}_i$ and $\mathbf{q}_j$ are dissimilar. Hence, the cosine max-margin loss can narrow the gap between the seen image and its corresponding text in K-dimensional embedding space.

3.3 Single Modal Losses L_x and L_t. The proposed max-margin loss effectively bridges the gap between different modalities by maintaining the similarity between the seen image and its corresponding text. However, it could not be directly applied for zero-shot learning because the semantic heterogeneity between unseen images and unseen texts still exists. The unseen instance representations, including P_u and Q_u, tend to be uncertain without any constraint factors. In other words, the representations P_u and Q_u may be distorted if we capture information only from seen image-text pairs. This case motivates us to design two specific unimodal losses, L_x and L_t, to control the quality of unimodal network directly. The unimodal loss can be presented in the following unified form,

$$L = G + \beta MMD, \tag{3.6}$$

which consists of two parts: the pairwise quantization loss G (see section 3.3.1) and the *MMD* constraint (see section 3.3.2). The parameter β is the shrinkage coefficient of the *MMD* constraint.

3.3.1 Pairwise Quantization Loss G. For the image network, the image representations from identical categories should be closer compared to others from disparate category intuitively. Thus, we put forward the following equation to minimize the following image pairwise quantization loss:

$$G_x = \sum_{i=1}^{n}\sum_{i'=1}^{n}\left(s_{ii'}^{x} - \frac{\mathbf{p}_i^{\top}\mathbf{p}_{i'}}{\|\mathbf{p}_i\|\|\mathbf{p}_{i'}\|}\right)^2. \tag{3.7}$$

The $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ denote the image network outputs of images $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, respectively. $s_{ii'}^{x} \in \{-1, 1\}$ is the indicator that represents whether images $\mathbf{x}_i$ and $\mathbf{x}_{i'}$ belong to the same category. When x_i and x_i' are from identical categories, the indicator $s_{ii'}^{x} = 1$; otherwise, $s_{ii'}^{x} = -1$. There are three conditions about the value of $s_{ii'}^{x}$:

- If $\mathbf{x}_i \in X_s$, $\mathbf{x}_{i'} \in X_s$, the value of $s_{ii'}^{x}$ is directly determined by whether the categories x_i and x_i' are the same.
- If $\mathbf{x}_i \in X_s$, $\mathbf{x}_{i'} \in X_u$, $s_{ii'}^{x} = -1$ because the seen classes and unseen classes are disjoint for zero-shot learning.
- If $\mathbf{x}_i \in X_u$, $\mathbf{x}_{i'} \in X_u$, the value of $s_{ii'}^{x}$ is a variable because the class labels of $\mathbf{x}_i$ and $\mathbf{x}_i'$ are uncertain. The similarity indicator matrix for unseen images is denoted as $S_x^u \in \{-1, 1\}^{n_u \times n_u}$, which requires optimization in the training procedure.

In function 3.7, the cosine distance of $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ ranges from -1 to 1, which corresponds to the similarity indicator $s_{ii'}^{x}$. To be more specific, the representations of $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ tend to be similar when $s_{ii'}^{x} = 1$; otherwise, $\mathbf{p}_i$ and $\mathbf{p}_{i'}$ would tend to be dissimilar. Hence, function 3.7 efficiently maintains the similarity of images from the same category theoretically.

Enlightened by the human cognitive process, the relevance among categories is conducive to identifying unseen instances better. For example, if one person has seen some images about the animal "lion," then knowledge of the relationship between "lion" and "tiger" will help him learn the new animal "tiger" in the future. The affluent textual descriptions $\mathbb{T} = \{\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_c\}$ for visual categories can be used to measure the relevance among all categories effectively. Considering that, we use the cosine distance of texts $\mathbf{t}_j$ and $\mathbf{t}_{j'}$, that is, $\cos(\mathbf{t}_j, \mathbf{t}_j')$, to represent the relevance between the jth category and the j'th category. For the text network, we naturally maintain the cosine similarity relevance among all categories by minimizing the following pairwise quantization loss,

$$G_t = \sum_{j=1}^{c}\sum_{j'=1}^{c}\left(s^t_{jj'} - \frac{\mathbf{q}_j^\top \mathbf{q}_{j'}}{\|\mathbf{q}_j\|\|\mathbf{q}_{j'}\|}\right)^2, \tag{3.8}$$

where $s^t_{jj'} = \cos(\mathbf{t}_j, \mathbf{t}'_j) \in [-1, 1]$. The $\mathbf{q}_j$ and $\mathbf{q}_{j'}$ denote the text network outputs of textual descriptions $\mathbf{t}_j$ and $\mathbf{t}_{j'}$, respectively. Apparently, function 3.8 could guarantee that the similarity relevance among learned representation $Q = [\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_c] \in \mathbb{R}^{K\times c}$ is consistent with the original relevance among all textual descriptions $\mathbb{T} = \{\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_c\}$.

Overall, in the embedding space, we not only reduce the difference of image representations from the same categories by minimizing the image pairwise quantization loss G_x, but also extract the correlation between seen and unseen categories by text pairwise quantization loss G_t.

3.3.2 MMD Constraint. Although seen and unseen instances are projected into the shared embedding space, the distribution difference between them might still be significant (Long et al., 2013; Zhang et al., 2015). This distribution difference brings the limitation of capturing the commonality of seen and unseen instances. Thus, a distribution matching strategy should be applied to reduce the difference of the seen images P_s and unseen images P_u distributions, as well as the seen texts Q_s and unseen texts Q_u distributions. The MMD (Gretton, Borgwardt, Rasch, Schölkopf, & Smola, 2012), which computes the Euclidean distance between different distributions, is often applied for measuring distribution differences because of its efficiency in computation and optimization. In this letter, two MMD constraints over the image and text domain, MMD$_x$ and MMD$_t$, are calculated using equations 3.9 and 3.10, respectively:

$$\left\|\frac{1}{n_s}\sum_{i=1}^{n_s}\mathbf{p}_i - \frac{1}{n_u}\sum_{i'=n_s+1}^{n}\mathbf{p}_{i'}\right\|^2 = Tr(PM^xP^\top). \tag{3.9}$$

$$\left\|\frac{1}{c_s}\sum_{j=1}^{c_s}\mathbf{q}_j - \frac{1}{c_u}\sum_{j'=c_s+1}^{c}\mathbf{q}_{j'}\right\|^2 = Tr(QM^tQ^\top). \tag{3.10}$$

M^x and M^t are two MMD matrices, which can be computed as

$$M^x_{ii'} = \begin{cases} 1/n_s^2 & x_i, x_{i'} \in X_s \\ 1/n_u^2 & x_i, x_{i'} \in X_u \\ -1/(n_s \times n_u) & \text{otherwise.} \end{cases} \quad M^t_{jj'} = \begin{cases} 1/c_s^2 & t_j, t_{j'} \in T_s \\ 1/c_u^2 & t_j, t_{j'} \in T_u \\ -1/(c_s \times c_u) & \text{otherwise.} \end{cases} \tag{3.11}$$

```
Algorithm 1: Deep Semisupervised Zero-Shot Learning with MMD.
Input: Image set $X = \{X_s, X_u\}$, textual description set $T = \{T_s, T_u\}$, cross-modal indicator matrix $H = \{h_{ij} : 1 \leq i \leq n_s, 1 \leq j \leq c_s\}$, trade-off parameter $\lambda$, $\beta_x$, $\beta_t$.
Output: Representations $P_s$, $P_u$, $Q_s$, and $Q_u$ for seen images, unseen images, seen texts, and unseen texts, respectively.
Initialize: Set initial value of $\theta = \{\theta_x, \theta_t, W_x^\ell, W_t^\ell\}$ and similarity matrix $S_x^u$ for unseen images.
while not converge do
    Update the similarity matrix $S_x^u$ for unseen images accordings section 4.1;
    Update the parameters of image network by solving $\arg\min_{\{\theta_x, W_x^\ell\}} L_f + \lambda L_x$ accordings section 4.2;
    Update the parameters of text network by solving $\arg\min_{\{\theta_t, W_t^\ell\}} L_f + \lambda L_t$ accordings section 4.2.
end while
```

Obviously, the MMD constraint will asymptotically approach zero if two distributions tend to be the same. Therefore, the MMD_x term can ensure that the seen image distribution is consistent with the unseen image distribution in the cross-modal embedding space. Similarly, the MMD_t term could also reduce the distribution difference between seen and unseen texts in theory.

4 Optimization Procedure

In this section, we train the proposed DCMN with the mini-batch stochastic gradient descent (SGD) method. We exploit the alternative optimization algorithm presented in algorithm 1 to solve the objective function 3.3 by updating the network parameters $\{\theta_x, W_x^\ell, \theta_t, W_t^\ell\}$ and the similarity matrix S_u^x for unseen images alternately.

4.1 Optimize S_x^u. When the network parameters $\{\theta_x, W_x^\ell, \theta_t, W_t^\ell\}$ are fixed, the category index y_i of unseen image $\mathbf{x}_i \in X_u$ can be predicted by

$$y_i = \arg\max_j \frac{\mathbf{p}_i^\top \mathbf{q}_j}{\|\mathbf{p}_i\| \|\mathbf{q}_j\|}, \tag{4.1}$$

where $(n_s + 1) \le i \le n$. To be more specific, $(c_s + 1) \le j \le c$ because unseen image $\mathbf{x}_i$ necessarily belongs to the unseen categories. We use $y_{i'}$ to denote the predicted category index for another unseen image, $\mathbf{x}_{i'} \in X_u$. After that, the similarity indicator $s^x_{ii'} = 1$ if $y_i = y_{i'}$ and otherwise, $s^x_{ii'} = -1$. Thus, the unseen image similarity matrix S^u_x can be solved according to the above function.

4.2 Optimize Parameters of DCMN. With the fixed text network parameters $\{\theta_t, W^\ell_t\}$ and similarity matrix S^x_u for unseen images, we could obtain the image network parameters $\{\theta_x, W^\ell_x\}$ by optimizing the following function,

$$\min_{\{\theta_x, W^\ell_x\}} L_f + \lambda L_x, \tag{4.2}$$

where L_x denotes the unimodal loss of image network: $L_x = G_x + \beta_x MMD_x$. The relative importance of the MMD constraint is controlled by the trade-off β_x. We denote the value of loss function 4.2 as F_x and derive the gradient of loss F_x with regard to W^ℓ_x as follows,

$$\frac{\partial F_x}{\partial W^\ell_x} = \frac{\partial L_f}{\partial W^\ell_x} + \lambda \frac{\partial L_x}{\partial W^\ell_x} = \left(\frac{\partial L_f}{\partial \hat{Z}^\ell_x} + \lambda \frac{\partial L_x}{\partial \hat{Z}^\ell_x} \right) \frac{\partial \hat{Z}^\ell_x}{\partial W^\ell_x} = \delta^\ell_x (Z^{\ell-1}_x)^\top, \tag{4.3}$$

where $\hat{Z}^\ell_x = W^\ell_x Z^{\ell-1}_x$ is the output of the ℓth layer before activation $a^\ell_x(\cdot)$, and Z^ℓ_x denotes the output of the ℓth layer after activation $a^\ell_x(\cdot)$. δ^ℓ_x is the residual term that measures the responsibility of the units in the ℓth layer for the loss function 4.2. In particular, $Z^{\ell-1}_x = f(X; \theta_x)$ and $Z^\ell_x = P$. Then we can directly define the residual δ^ℓ_x of the last fcm layer as

$$\delta^\ell_x = \left(\frac{\partial L_f}{\partial Z^\ell_x} + \lambda \frac{\partial L_x}{\partial Z^\ell_x} \right) \odot \frac{\partial Z^\ell_x}{\partial \hat{Z}^\ell_x} = \left(\frac{\partial L_f}{\partial P} + \lambda \frac{\partial L_x}{\partial P} \right) \odot \tanh'(f(X; \theta_x)), \tag{4.4}$$

where $\frac{\partial L_f}{\partial P} = [\frac{\partial L_f}{\partial \mathbf{p}_1}, \frac{\partial L_f}{\partial \mathbf{p}_2}, \ldots, \frac{\partial L_f}{\partial \mathbf{p}_{n_s}}, \underbrace{\mathbf{0}, \ldots, \mathbf{0}}_{n_u}]$. The element $\frac{\partial L_f}{\partial \mathbf{p}_i} (i \le n_s)$ can be computed using function 4.5 where $\mathbb{I}(x)$ is a condition indicator function, that is, $\mathbb{I}(x) = 1$ if x is true and otherwise $\mathbb{I}(x) = 0$:

$$\frac{\partial L_f}{\partial \mathbf{p}_i} = \sum_{j=1}^{c} \mathbb{I}\left(\triangle - h_{ij} \cos(\mathbf{p}_i, \mathbf{q}_j) > 0\right) \left(-h_{ij} \frac{\partial \cos(\mathbf{p}_i, \mathbf{q}_j)}{\partial \mathbf{p}_i} \right). \tag{4.5}$$

Similarly, $\frac{\partial L_x}{\partial P} = [\frac{\partial L_x}{\partial \mathbf{p}_1}, \frac{\partial L_x}{\partial \mathbf{p}_2}, \ldots, \frac{\partial L_x}{\partial \mathbf{p}_n}]$ and $\frac{\partial L_x}{\partial \mathbf{p}_i} = \frac{\partial G_x}{\partial \mathbf{p}_i} + \beta_x \frac{\partial MMD_x}{\partial \mathbf{p}_i}$, which can be achieved by equations 4.6 and 4.7:

$$\frac{\partial G_x}{\partial \mathbf{p}_i} = 2 \times \sum_{i'=1}^{n} \big(\cos(\mathbf{p}_i, \mathbf{p}_{i'}) - s^x_{ii'}\big) \frac{\partial \cos(\mathbf{p}_i, \mathbf{p}_{i'})}{\partial \mathbf{p}_i}, \tag{4.6}$$

$$\frac{\partial \mathrm{MMD}_x}{\partial P} = \frac{\partial Tr(PM^x P^\top)}{\partial P} = 2 \times PM^x. \tag{4.7}$$

According to the functions 4.4 to 4.7, the parameter W_x^ℓ of the fcm layer in the image network is obtained by function 4.3. After that, the parameters $W_x^{\ell-1}$ of $(\ell-1)$-layer in image network can be computed using

$$\frac{\partial F_x}{\partial W_x^{\ell-1}} = \delta_x^{\ell-1}(Z_x^{\ell-2})^\top, \tag{4.8}$$

where the residual term $\delta_x^{\ell-1}$ is as follows:

$$\delta_x^{\ell-1} = ((W_x^\ell)^\top \delta_x^\ell) \odot \dot{a}_x^{\ell-1}(Z_x^{\ell-1}). \tag{4.9}$$

The $\dot{a}_x^{\ell-1}(\cdot)$ denotes the derivative of the $(\ell-1)$th layer activation function, that is, the derivative of ReLU in our image network. Obviously, the computing process of residuals in all layers is consistent with the standard BP procedure. As a result, the parameters θ_x can be computed by the traditional BP algorithm.

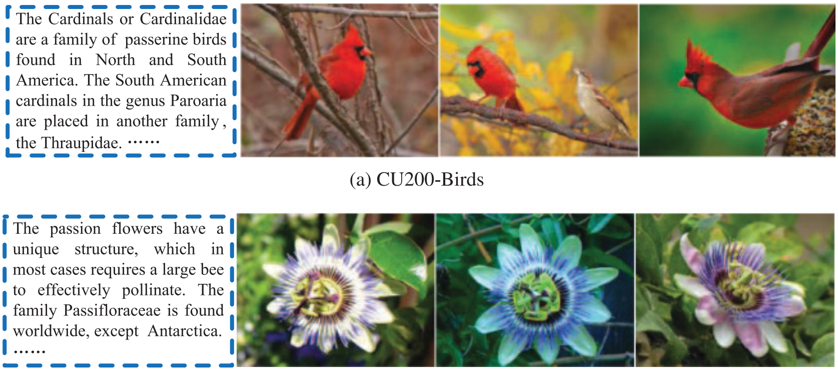
We omit the optimization process of text network parameters $\{\theta_t, W_t^\ell\}$ under the fixed parameters $\{\theta_x, W_x^\ell\}$ because it is analogous to the presented optimization procedure.

5 Experiment

In this section, we conduct extensive experiments on two benchmark data sets to validate the effectiveness of our method.

5.1 Experimental Setup

5.1.1 Data Sets. We perform the experiments on two benchmark data sets, CU200-Birds and Oxford Flowers-102. The CU200-Birds data set (Welinder et al., 2010) consists of 6033 images from 200 bird categories in which each visual class has approximately 30 images. The Oxford Flowers-102 data set (Nilsback & Zisserman, 2008) contains 102 flower species with a total of 8189 images; each class consists of 40 to 258 images. Similar to the work in Elhoseiny et al. (2013), each visual category in these two data sets has a corresponding textual article acquired from Wikipedia or other



The Cardinals or Cardinalidae are a family of passerine birds found in North and South America. The South American cardinals in the genus Paroaria are placed in another family, the Thraupidae. ……

(a) CU200-Birds

The passion flowers have a unique structure, which in most cases requires a large bee to effectively pollinate. The family Passifloraceae is found worldwide, except Antarctica. ……

(b) Oxford Flowers-102

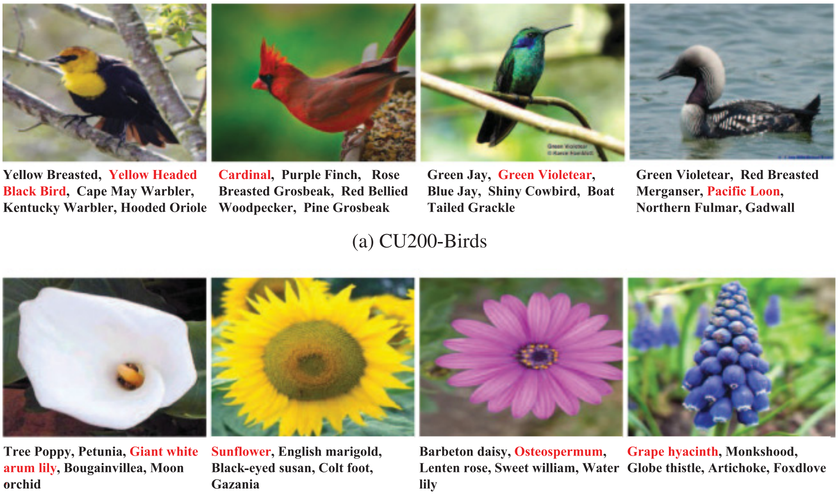
Figure 2: Some examples of images and textual descriptions in two data sets.

authoritative database. Some examples of the images and corresponding textual descriptions for these two data sets are illustrated in Figure 2.

5.1.2 Competitors. We compare the proposed method with the following state-of-the-art baselines to evaluate its effectiveness and superiority:

- Twin gaussian processes (TGP) (Bo & Sminchisescu, 2010), a structured prediction regression method that uses gaussian processes (GP) priors for both covariates and responses
- Domain adaptation using asymmetric kernel transforms (DA) (Kulis, Saenko, & Darrell, 2011), a novel approach to learn a nonlinear transformation for domain adaptation
- Zero-shot learning using purely textual descriptions (DA+GP) (Elhoseiny et al., 2013); which combines a regression function and a knowledge transfer function with additional constraints to predict the classifier parameters for new categories
- Deep zero-shot convolutional neural networks (Deep $fc + conv$) (Lei Ba et al., 2015), which uses text features to predict the weights of the convolutional and the fully connected layers in a deep convolutional neural network. We simply use the fully connected layer model (Deep fc) as the competitor in our experiment. The Deep fc can achieve neck-and-neck performance compared with the Deep $fc + conv$ method on benchmark data sets CU200-Birds and Oxford Flowers-102.

5.1.3 Implementation. We implement the DCMN network based on the open-source Torch framework (Collobert, Kavukcuoglu, & Farabet, 2011). For image network, we resize all images to 224 × 224 pixels. Note that


Yellow Breasted, Yellow Headed Black Bird, Cape May Warbler, Kentucky Warbler, Hooded Oriole
Cardinal, Purple Finch, Rose Breasted Grosbeak, Red Bellied Woodpecker, Pine Grosbeak
Green Jay, Green Violetear, Blue Jay, Shiny Cowbird, Boat Tailed Grackle
Green Violetear, Red Breasted Merganser, Pacific Loon, Northern Fulmar, Gadwall
Tree Poppy, Petunia, Giant white arum lily, Bougainvillea, Moon orchid
Sunflower, English marigold, Black-eyed susan, Colt foot, Gazania
Barbeton daisy, Osteospermum, Lenten rose, Sweet william, Water lily
Grape hyacinth, Monkshood, Globe thistle, Artichoke, Foxdlove

(a) CU200-Birds

(b) Oxford Flowers-102

Figure 3: Examples of predicted classification results on two data sets.

the initial parameters $\theta_x = \{W_x^{\ell-19}, \ldots, W_x^{\ell-2}, W_x^{\ell-1}\}$ of convolutional layers $conv1$ to $conv16$ and fully connected layers $fc17$ to $fc19$ are copied from the pretrained VGG19 network on the ImageNet data set. Additionally, we extract the term frequency-inverse document frequency (TF-IDF) features from raw textual descriptions and then use the clustered latent semantic indexing (CLSI) (Zeimpekis & Gallopoulos, 2005) algorithm to reduce the feature dimension. The TF-IDF features of the CU200-Birds and Oxford Flowers-102 data sets are in $\mathbb{R}^{7086}$ and $\mathbb{R}^{8875}$, respectively. After processing with the CLSI algorithm, the final textual features became the space of in $\mathbb{R}^{200}$ and $\mathbb{R}^{102}$, respectively.

5.2 Performance Comparison. In our experiment, we use fivefold cross-validation to evaluate the performance of our method and those of others. For each data set, 80% categories are considered the seen classes and the remaining 20% categories are considered the unseen classes. Among the seen classes, we randomly separate 80% of the images used for training and 20% of the images used for testing. We use the receiver operating characteristic (ROC) curve and the area under ROC curve (AUC) as the evaluation metrics for all methods, where the images from one certain unseen class are regarded as positive samples and the others are considered negatives. In this case, we show several examples of test images and their top five predicted labels in Figure 3, where the correct predictions are in red. The AUC results on the two data sets are shown in Table 1, in which our method has three versions. +MMD and −MMD represent with and without two MMD

Table 1: Performance Comparison in Terms of AUC over the Benchmark Data sets, CU200-Birds and Oxford Flowers-102.

	Oxford Flowers-102		CU200-Birds	
Methods	Seen	Unseen	Seen	Unseen
TGP	NA	0.58 ± 0.02	NA	0.61 ± 0.02
DA	NA	0.62 ± 0.03	NA	0.59 ± 0.01
DA+GP	NA	0.68 ± 0.01	NA	0.62 ± 0.02
Deep fc	0.96 ± NA	0.63 ± NA	0.93 ± NA	0.69 ± NA
Ours (−MMD TF-IDF)	0.94 ± 0.03	0.73 ± 0.02	0.94 ± 0.02	0.71 ± 0.02
Ours (+MMD TF-IDF)	0.99 ± 0.01	0.78 ± 0.02	0.97 ± 0.01	0.76 ± 0.01
Ours (+MMD FV)	**0.99 ± 0.02**	**0.82 ± 0.01**	**0.98 ± 0.01**	**0.82 ± 0.02**

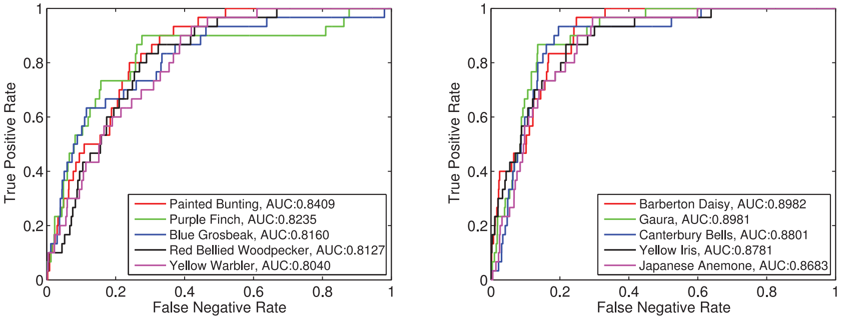
Note: The numbers in bold are the best classification results.

terms, respectively, by setting β_x and β_t equal or not to zero. We also extract the Fisher vectors (FV) for textual descriptions (Klein, Lev, Sadeh, & Wolf, 2014) to further illustrate the availability of our method. We have the following observations according to Table 1:

- DA+GP performs better than DA and GP on two data sets, because it effectively combines a regression function and a knowledge transfer function with additional constraints.
- With the same text features (TF-IDF), our method has better classification performance than the baselines, which indicates that the proposed DCMN successfully captures the common knowledge of seen and unseen instances. In addition, because of the better representation capability of FV features (Yan & Mikolajczyk, 2015; Wang, Li, & Lazebnik, 2016), the classification results are significantly improved when we replace the TF-IDF features with FV text features.
- Comparing the classification results of "Ours (−MMD TF-IDF)" with "Ours (+MMD TF-IDF)," we can conclude that the MMD terms could effectively improve the performance by decreasing the distribution difference of seen and unseen instances. Note that we analyze the influence of MMD terms with the increase of iterations in section 5.3.
- The classification results on Oxford Flowers-102 are generally better than those on CU200-Birds. It is reasonable because more categories in CU200-Birds may increase the uncertainty and difficulty of classification.

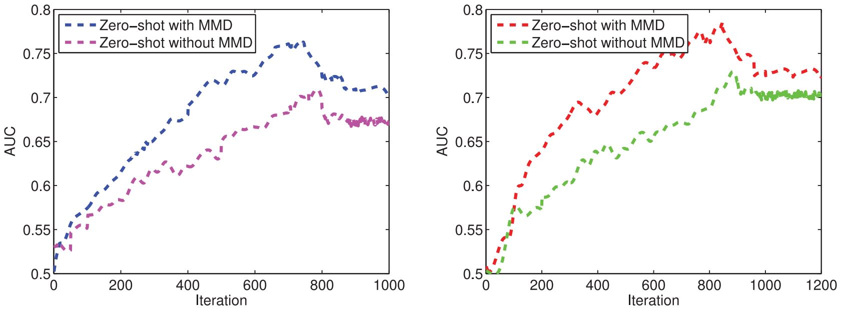
We also show the ROC curves and the corresponding AUC values of the best 10 unseen classes for CU200-Birds and Oxford Flowers-102 in Figure 4.

5.3 Impact of MMD Term. In this section, we conduct an experiment to evaluate whether the MMD terms contribute much to the subsequent performance. By setting $\beta_x = \beta_t = 0$, we obtain a modified framework without



(a) CU200-Birds (b) Oxford Flowers-102

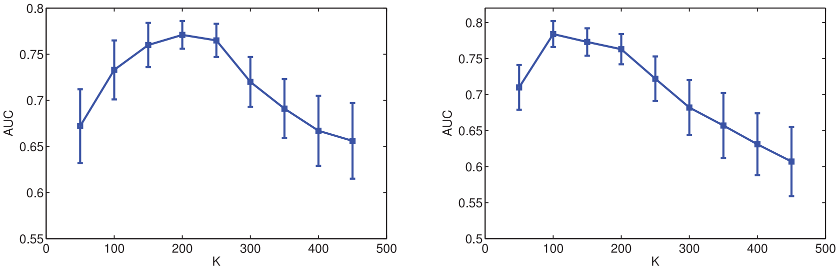
Figure 4: ROC curves of the best five predicted classes for two data sets.



(a) CU200-Birds (b) Oxford Flowers-102

Figure 5: Performance comparison between DCMN with and without MMD terms.

MMD terms. The results shown in Table 1 clearly indicate that the proposed method with MMD constraints outperforms the framework without MMD on the two data sets. In Figure 5, we plot the AUC curves for unseen categories with increased iteration times, where the texts are represented as TF-IDF features. A common phenomenon is that AUC performance improves as the iteration increases at first. After the performance reaches its maximum, it will drop if more iterations are conducted because of overfitting. We also observe that the performance with MMD terms always achieves better classification performance and becomes increasingly better with the increase of iterations than the framework without MMD terms. When the iteration is in the range [450, 750] and [400, 800] for CU200-Birds and Oxford Flowers-102, respectively, the framework with MMD terms shows the more prominent performance compared with the framework without MMD terms. These results indicate that the MMD constraints are



(a) CU200-Birds

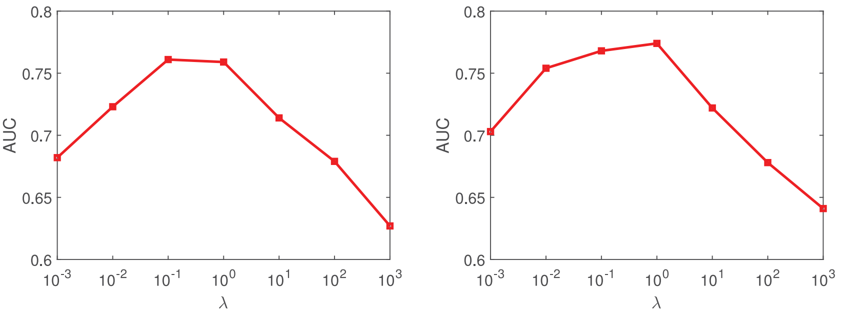
(b) Oxford Flowers-102

Figure 6: The influence of embedding dimension K on classification performance.

effective for semisupervised zero-shot learning by decreasing the distribution discrepancy between seen and unseen instances.

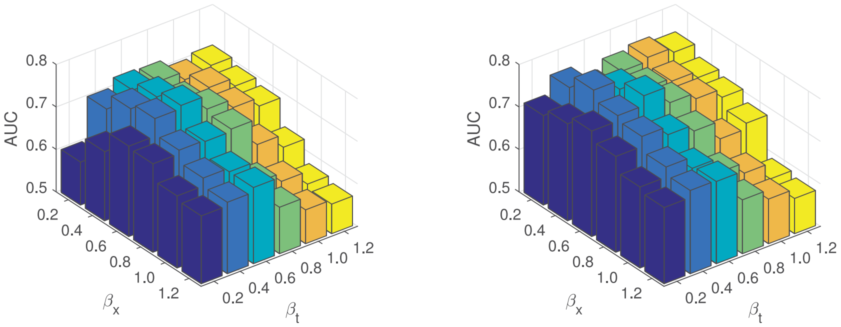
5.4 Impact of Embedding Dimension K**.** In this section, we study the influence of the embedding dimension K on classification performance. We assign the embedding dimension K to be tuned from 50 to 400 with a step size 50. As shown in Figure 6, for the two data sets, the AUC value for the unseen categories improves at first as K increases. After reaching its maximum, performance gradually decreases. Note that the classification result fluctuates widely when the value of K is too small or too large. The results indicate that the appropriate K could achieve satisfactory and relatively stable classification results. The best K is distinguishing on different data sets. When K equals 200 and 100, our method obtains the best classification performance on CU200-Birds and Oxford Flowers-102, respectively.

5.5 Impact of Hyperparameters. We analyze the influences of the tradeoff hyperparameters, λ, β_x, and β_t, on classification performance. Note that parameter λ controls the relative importance of unimodal loss $(L_x + L_t)$ regarding cross-modal max-margin loss L_f. The parameters β_x and β_t are the shrinkage coefficients of the MMD constraints on image modal loss and text modal loss, respectively. In the condition of parameter $\beta_x = \beta_t = 0.4$, we plot the performance curves for unseen categories as the increase of λ in Figure 7, where parameter λ is assigned in $\{10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}\}$. Besides, with $\lambda = 0.1$, we assign β_x and β_t varying from 0.2 to 1.2 and then show the classification results for unseen categories in Figure 8. These results demonstrate that appropriate values of λ, β_x, and β_t could achieve good classification results. The optimal parameters over different data sets are distinguishing. The classification accuracy could be satisfactory and relatively stable when parameter λ is in interval [0.1, 1] and β_x and β_t are in [0.4, 0.6].



(a) CU200-Birds (b) Oxford Flowers-102

Figure 7: The influence of hyperparameter λ on classification performance.



(a) CU200-Birds (b) Oxford Flowers-102

Figure 8: The influences of hyperparameters β_x and β_t on classification performance.

6 Conclusion

In this letter, we propose a novel deep cross-modal network for semisupervised zero-shot learning. We combine the CNN network for image modality with the MLP network for text modality to map all visual objects and textual articles into a common semantic space. In particular, the MMD constraints are proposed in our letter to decrease the distribution discrepancy between seen and unseen instances. The extensive experimental results on the benchmark data sets CU200-Birds and Oxford Flowers-102 show that the proposed method outperforms other state-of-the-art methods on the ROC-AUC metric. In future work, we plan to replace the text TF-IDF feature extraction process and the MLP network part with LSTM recurrent neural network to learn text representations automatically.

Acknowledgments

This work is sponsored by The Fundamental Theory and Applications of Big Data with Knowledge Engineering under the National Key Research and Development Program of China, grant 2016YFB1000903; National Science Foundation of China under grants 61502377, 61532004, 61532015, 61672418, and 61672419; Project of China Knowledge Centre for Engineering Science and Technology; Ministry of Education Innovation Research Team IRT 17R86; and the Innovative Research Group of the National Natural Science Foundation of China (61721002).

References

Bo, L., & Sminchisescu, C. (2010). Twin gaussian processes for structured prediction. *International Journal of Computer Vision*, *87*(1), 28–52.

Cao, Y., Long, M., Wang, J., Yang, Q., & Yu, P. S. (2016). Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (pp. 1445–1454). New York: ACM.

Collobert, R., Kavukcuoglu, K., & Farabet, C. (2011). Torch7: A Matlab-like environment for machine learning. In *BigLearn, NIPS Workshop*, EPFL-CONF-192376.

Dauphin, Y. N., Tur, G., Hakkani-Tur, D., & Heck, L. (2013). *Zero-shot learning for semantic utterance classification*. arXiv:1401.0509.

Elhoseiny, M., Saleh, B., & Elgammal, A. (2013). Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the International Conference on Computer Vision* (pp. 2584–2591). Piscataway, NJ: IEEE.

Frome, A., Corrado, G. S., Shlens, J., Bengio, S., Dean, J., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 2121–2129). Red Hook, NY: Curran.

Fu, Y., Hospedales, T. M., Xiang, T., & Gong, S. (2015). Transductive multi-view zero-shot learning. *Pattern Analysis and Machine Intelligence*, *37*(11), 2332–2345.

Fu, Y., & Sigal, L. (2016). Semi-supervised vocabulary-informed learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 5337–5346). Piscataway, NJ: IEEE.

Fu, Z., Xiang, T., Kodirov, E., & Gong, S. (2015). Zero-shot object recognition by semantic manifold distance. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 2635–2644). Piscataway, NJ: IEEE.

Glorot, X., Bordes, A., & Bengio, Y. (2011). Deep sparse rectifier neural networks. In *Proceedings of the International Conference on Artificial Intelligence and Statistics* (pp. 315–323).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, *13*(1), 723–773.

Kansky, K., Silver, T., Mély, D. A., Eldawy, M., Lázaro-Gredilla, . . . George, D. (2017). *Schema networks: Zero-shot transfer with a generative causal model of intuitive physics*. arXiv:1706.04317.

Klein, B., Lev, G., Sadeh, G., & Wolf, L. (2014). *Fisher vectors derived from hybrid gaussian-Laplacian mixture models for image annotation*. arXiv:1411.7399.

Kovashka, A., Parikh, D., & Grauman, K. (2012). Whittlesearch: Image search with relative attribute feedback. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 2973–2980). Piscataway, NJ: IEEE.

Kovashka, A., Vijayanarasimhan, S., & Grauman, K. (2011). Actively selecting annotations among objects and attributes. In *Proceedings of the International Conference on Computer Vision* (pp. 1403–1410). Piscataway, NJ: IEEE.

Kulis, B., Saenko, K., & Darrell, T. (2011). What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 1785–1792). Piscataway, NJ: IEEE.

Kumar, N., Berg, A., Belhumeur, P. N., & Nayar, S. (2011). Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence*, *33*(10), 1962–1977.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 951–958). Piscataway, NJ: IEEE.

Lampert, C. H., Nickisch, H., & Harmeling, S. (2014). Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence*, *36*(3), 453–465.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*(1), 98–113.

Lei Ba, J., Swersky, K., Fidler, S., & Salakhutdinov, R. (2015). Predicting deep zero-shot convolutional neural networks using textual descriptions. In *Proceedings of the International Conference on Computer Vision* (pp. 4247–4255). Piscataway, NJ: IEEE.

Liu, M., Zhang, D., & Chen, S. (2014). Attribute relation learning for zero-shot classification. *Neurocomputing*, *139*, 34–46.

Long, M., Ding, G., Wang, J., Sun, J., Guo, Y., & Yu, P. S. (2013). Transfer sparse coding for robust image representation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 407–414). Piscataway, NJ: IEEE.

Nilsback, M. E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *Proceedings of the Conference on Computer Vision, Graphics and Image Processing* (pp. 722–729).

Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., & Dean, J. (2013). *Zero-shot learning by convex combination of semantic embeddings*. arXiv:1312.5650.

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009). Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems, 27* (pp. 1410–1418).

Parikh, D., & Grauman, K. (2011). Relative attributes. In *Proceedings of the International Conference on Computer Vision* (pp. 503–510).

Peng, K.-C., Wu, Z., & Ernst, J. (2017). *Zero-shot deep domain adaptation*. arXiv:1707.01922.

Scheirer, W. J., Kumar, N., Belhumeur, P. N., & Boult, T. E. (2012). Multi-attribute spaces: Calibration for attribute fusion and similarity search. In *Proceedings Computer Vision and Pattern Recognition* (pp. 2933–2940).

Shojaee, S. M., & Baghshah, M. S. (2016). *Semi-supervised zero-shot learning by a clustering-based approach*. arXiv:1605.09016.

Simonyan, K., & Zisserman, A. (2014). *Very deep convolutional networks for large-scale image recognition*. arXiv:1409.1556.

Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems, 26* (pp. 935–943). Red Hook, NY: Curran.

Wang, L., Li, Y., & Lazebnik, S. (2016). Learning deep structure-preserving image-text embeddings. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 5005–5013). Piscataway, NJ: IEEE.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., & Perona, P. (2010). *Caltech-UCSD birds 200*. (CNS-TR-2010-001). Pasadena: CalTech.

Xian, Y., Schiele, B., & Akata, Z. (2017). *Zero-shot learning—the good, the bad and the ugly*. arXiv:1703.04394.

Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *Proceedings of the Conference on Computer Vision and Pattern Recognition* (pp. 3441–3450). Piscataway, NJ: IEEE.

Yu, X., & Aloimonos, Y. (2010). Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of the European Conference on Computer Vision* (pp. 127–140). Berlin: Springer.

Zeimpekis, D., & Gallopoulos, E. (2005). CLSI: A flexible approximation scheme from clustered term-document matrices. In *Proceedings of the SIAM International Conference on Data Mining* (pp. 631–635). Philadelphia: SIAM.

Zhang, X., Yu, F. X., Chang, S. F., & Wang, S. (2015). *Deep transfer network: Unsupervised domain adaptation*. arXiv:1503.00591.

Zhang, Z., & Saligrama, V. (2015). Zero-shot learning via semantic similarity embedding. In *Proceedings of the International Conference on Computer Vision* (pp. 4166–4174). Piscataway, NJ: IEEE.

Zhou, T. (2016). An image recognition model based on improved convolutional neural network. *Journal of Computational and Theoretical Nanoscience*, *13*(7), 4223–4229.

Received August 15, 2017; accepted January 4, 2018.