Neurocomputing 238 (2017) 76–83

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Zero-shot classification by transferring knowledge and preserving data structure

Xiao Li, Min Fang*, Jinqiao Wu

School of Computer Science and Technology, Xidian University, 710071, China

ARTICLE INFO

Article history:
Received 29 September 2016
Revised 12 December 2016
Accepted 13 January 2017
Available online 28 January 2017

Communicated by T. Mu

Keywords:
Zero-shot learning
Object recognition
Semantic correlation
Manifold structure

ABSTRACT

In practical object recognition tasks, one often encounters a problem to recognize some unseen objects, which are some new categories without labeled images at training stage. For solving the challenging problem, zero-shot learning has been studied widely which can be seen as a special case of transfer learning. Thus, zero-shot learning deals with the problem of predicting labels of target images based on source images and their common semantic knowledge. Most existing zero-shot learning methods focus on how to project the images into the semantic space. However, the projection function learnt by source images and attributes has a shift for the prediction of target attributes. In this paper, we proposed a zero-shot classification method by transferring knowledge from source domain and preserving target data structure (TKDS). Particularly, we directly learn the target classification model by the semantic correlation with source classification model. Different from existing similarity based zero-shot learning methods, we also utilize the data properties of target data themselves. We simultaneously consider transferring knowledge from source domain to explicit the source images and utilizing the manifold structure of target data to rectify domain shift problem, thus, boosting the prediction performance. Extensive experiments on the widely used datasets show that our model outperforms significantly the state-of-the-art methods.

1. Introduction

Object recognition tasks require large amount of labeled data for all the categories. However, there are many categories without training data in practice. It is expensive to annotate instances for all the categories when the number of categories is large (large-scale visual recognition [1,2]). It is unnecessary to annotate instances for all the categories when some categories are similar (fine-grained classification [3,4]). Therefore, it is urgent and important to develop algorithms that can reduce human labor to label instances for all classes. Zero-shot learning [5–7] and few-shot learning [8,9] have been widely studied in recent years to reduce the labeling cost and recognize an instance of a new class.

Zero-shot learning has shown to be of utility in various applications, such as human action recognition [10,11], activity recognition [12,13], object recognition [14,15], event detection [16,17]. Zero-shot learning aims to deal with the problem by utilizing class prototypes and the seen classes to recognize the unseen classes. Seen classes are with sufficient labeled instances. Unseen classes are new categories of instances without labeled instances. A class prototype means the embedded label representation in a semantic space. The semantic spaces include attribute space where every class is represented by an attribute vector [5,18] and word vector space where every class is represented by the textual description of it which is learnt by large text-corpora [2,19,20]. Zero-shot learning is a special case of transfer learning [21] where seen classes are source domain images and unseen classes are target domain images. However, there are also differences between them. In zero-shot learning setting, there are no instances in target domain and no overlap between source and target labels. The subspace shared between source and target domain is semantically meaningful. While the subspace in transfer learning does not have semantic correlation between source and target labels.

The key problems addressed in zero-shot learning are what are the relationship between unseen classes and seen classes and how to predict the unseen samples accurately even though we have no access to their data when training. Existing zero-shot classification methods mainly focus on the two-step method, mapping image to semantic space and then to label space. However, the final goal of zero-shot learning is to classify the unseen images. We propose a novel zero-shot classification method that can directly learn the classification model rather than using a two-step method. We utilize the attribute representations to learn the semantic correlation, instead of as the middle level representations. The semantic correlation between unseen and seen classes is obtained by capturing

* Corresponding author.
E-mail addresses: fanglabtg@163.com, mfang@mail.xidian.edu.cn (M. Fang).

http://dx.doi.org/10.1016/j.neucom.2017.01.038
0925-2312/

the linear relationship of their class prototypes. The transferred target classification model is learnt by transferring source classification model using the semantic correlation. Considering the disjoint unseen classes and seen classes, the mapping learnt by source images and attributes suffers from domain shift problem [15] when applied in target domain. In order to overcome the problem, when classifying target images, we not only take advantage of the transferred target classification model, but also consider the data properties of target data.

The main contributions of this paper are summarized as follows:

(1) We propose a novel zero-shot classification method by transferring knowledge from source domain and preserving data structure of target domain. With the transferred knowledge from source domain, the target classification model which directly connects images and labels can be easily established instead of predicting the attribute firstly.

(2) Attributes are utilized to learn the semantic correlation rather than used as the middle level representations in standard zero-shot learning methods. We exploit the semantic correlation to transfer the classification model from source to target domain.

(3) To handle the domain shift problem, manifold structure of target data is considered along with the transferred target classification model when classifying target images.

(4) To verify our method, we evaluate it on the state-of-the-art zero-shot learning datasets and demonstrate that it achieves much better performance than existing methods.

This paper is organized as follows. Section 2 provides a brief review of related work. In Section 3, we develop a novel zero-shot classification method by transferring knowledge and preserving data structure (TKDS). Section 4 provides the optimization of our algorithm. In Section 5, extensive experiments are conducted. In Section 6, the conclusions and future work are presented.

2. Related work

Zero-shot learning is actually a two-step process: firstly, mapping between images and attributes is learnt; secondly, predicting the labels of the mapped unseen images by comparing their similarity with the unseen class prototypes.

Attributes are important semantic representations in zero-shot learning problem. There are many methods aiming to predict the attributes for the unseen objects. Then the images are classified in the semantic space based on the similarity of the predicted attributes and the class prototypes. Direct Attribute Prediction (DAP) [5] first predicts attributes and then uses maximum a posteriori to predict the labels of the unseen images. Since attribute classifiers are learned independently of the recognition task. The overall strategy of DAP might be optimal at predicting attributes but not necessarily at predicting classes. Wang et al. [22] capture the relationships between attributes and objects for attribute prediction and object recognition. The work of Liu et al. [6] learns the attribute-attribute relation automatically and explicitly. Jayaraman et al. [23] consider the unreliability of attribute predictions and propose a random forest model.

Akata et al. [24] propose a bilinear function to measure the consistency between an image and a label embedding. ESZSL proposed in [25] is easy and simple to learn the mapping using a bilinear model. These methods all adopt the semantic space as the bridge while no classification model with images and labels is learnt. Since source and target domains have disjoint classes, the mapping function learnt from source images and semantic representations differs greatly with target domain. However, these methods also do not consider the difference between source and target domains. The mapping function learnt by source images and attributes has a shift for the prediction of target attributes. While we not only utilize the knowledge transferred from source domain, but also preserve properties of target data when recognizing target images. Thus, we can overcome the domain shift problem.

There are many methods to deal with the domain shift problem. A TME method is proposed in [14] to learn an embedding space for the images and multiple semantic representations to tackle the projection domain shift problem. In [26], the method maps images into the semantic embeddings by sparse coding and regularizes the amount of adaptation from the source dictionary. The above methods consider the domain shift problem. But they also belong to the two-step methods. However, solving the target problem directly is better than introducing a intermediate problem. While our method directly learns the target classification model based on the source classification model with the semantic correlation.

There are many methods aiming to deal with the zero-shot classification problem directly. A semi-supervised max-margin classification framework [27] that integrates the semi-supervised classification problem over the seen classes and the unsupervised clustering problem over the unseen classes into a unified max-margin multi-class classification formulation, which exploits both labeled and unlabeled data. Li et al. [28] propose a method to tackle the prediction problem directly based on multi-class classification over all seen and unseen classes. A SMS method is proposed in [29] which can directly get the labels of target data. They are all transductive methods where unlabeled target images are available. However, the target images are unavailable in the training process. Our method does not use the data of the unseen classes when training.

There are many methods similar to us, such as [30–32]. The method in [30] estimates a classifier for a new label, as a weighted combination of related classes, using the co-occurrences of visual concepts to define the weight. Norouzi et al. [31] represent the unseen images as a convex combination of seen class prototypes in the semantic space. The idea in [32] is also combining the base classifiers, which is smaller than the seen classifiers, for the unseen classes. They all create classifiers for unseen classes by linear combinations of classifiers for seen classes which resemble our method. However, we consider not only representing unseen classifiers by seen classifiers, but also the manifold structure of target data.

3. The proposed method

In this work, we have a large number of images from seen classes. Our goal is to utilize the seen images to help predicting the unseen images. The overview of the proposed method for zero-shot classification problem is shown in Fig. 1.

Let $D_s = \{\mathbf{X_s}, \mathbf{Z_s}\}$ be the source data with the labels of seen classes $\{1, \ldots, c\}$ and $D_t = \{\mathbf{X_t}, \mathbf{Z_t}\}$ be the target data. The labels of unseen classes are $\{c+1, \ldots, c+u\}$ which are unknown. $\mathbf{X_t}$ are also unknown in the training stage. The semantic representations of the seen classes and unseen classes are $\mathbf{P_s} = \{\mathbf{P_1}, \ldots, \mathbf{P_c}\}$ and $\mathbf{P_t} = \{\mathbf{P_{c+1}}, \ldots, \mathbf{P_{c+u}}\}$, separately. $\mathbf{P_i} \in R^{a \times 1}$ is the semantic representation of the ith class, in which a is the dimension of semantic representation. $\mathbf{X_s} \in R^{n_s \times d}$ and $\mathbf{X_t} \in R^{n_t \times d}$ are source (seen) and target (unseen) images, d is the dimension of image features and n_s is the number of source samples and n_t target samples. $\mathbf{Z_s} \in R^{n_s \times c}$ and $\mathbf{Z_t} \in R^{n_t \times u}$ are source and target labels. $\mathbf{z_i} = [0, 0, \ldots, 1, \ldots, 0] \in R^{1 \times c}$ is the element of $\mathbf{Z_s} \in R^{n_s \times c}$, the column where 1 is represents the label of the sample. There is no overlap between $\mathbf{Z_s}$ and $\mathbf{Z_t}$. Table 1 shows the list of notations and descriptions used in this paper.

The transpose of matrix is represented by the superscript '. $\mathbf{I_k}$ is k identity matrix. For a matrix $\mathbf{B} \in R^{n \times k}$, denote the ith column as $\mathbf{b^i}$, the jth row as $\mathbf{b_j}$. B_{ij} means the elements at the ith row and jth