Neurocomputing 316 (2018) 369–375

Contents lists available at ScienceDirect

Neurocomputing

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

Brief papers

Semantic softmax loss for zero-shot learning

Zhong Ji[a], Yuxin Sun[a], Yunlong Yu[a,*], Jichang Guo[a], Yanwei Pang[a]

School of Electrical and Information Engineering, Tianjin University, China

ARTICLE INFO

Article history:
Received 12 June 2017
Revised 27 April 2018
Accepted 8 August 2018
Available online 17 August 2018

Communicated by Dr Xiaofeng Zhu

Keywords:
Zero-shot learning
Semantic embedding
Multi-class classification

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 four benchmark datasets, i.e., AwA, CUB, SUN and ImageNet 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.

1. Introduction

Zero-Shot Learning (ZSL) [1–9] 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 [6,10,11]. One of the most popular among these approaches is the direct attribute prediction (DAP) approach [6], 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 [12–14] or project the features from both modalities into a common space [4,7,8,15,16]. To measure the similarity, most approaches use nearest neighbour classifier (NN) [4,6,15,17] or label propagation [18].

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

[*] Corresponding author.
E-mail address: yuyunlong@tju.edu.cn (Y. Yu).

https://doi.org/10.1016/j.neucom.2018.08.014

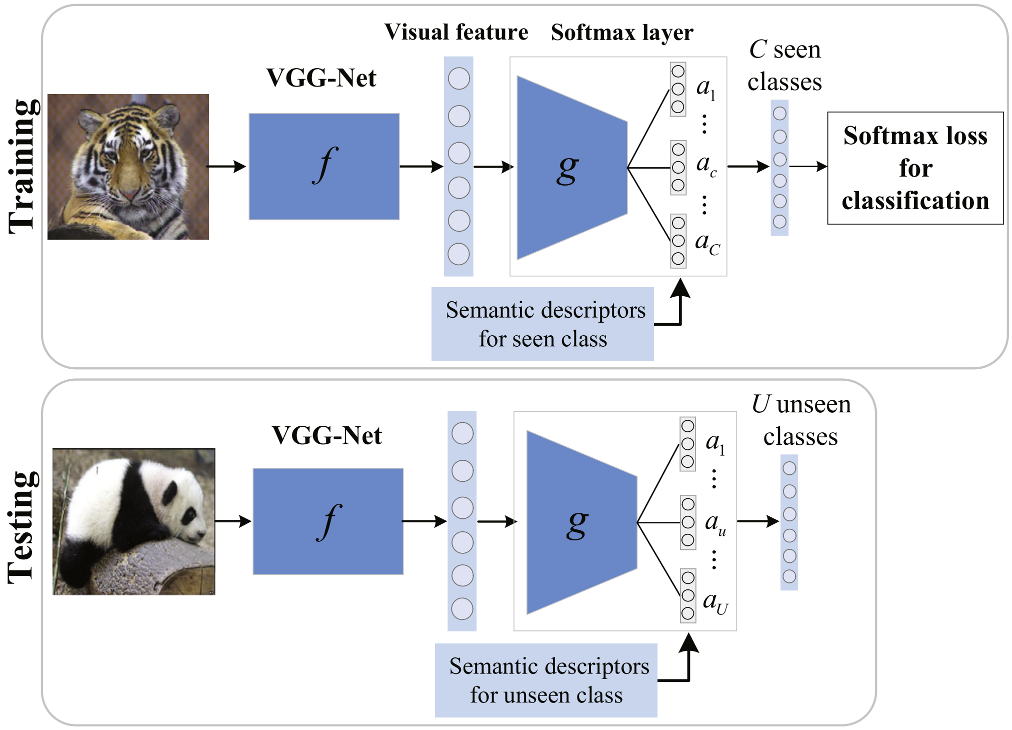

Fig. 1. The proposed pipeline for zero-shot learning. In the training stage, the images and the class semantic descriptors from seen classes are taken as input to predict the class label. The VGG-Net is adopted to extract the visual feature, and the class semantic descriptors are embedded in the softmax 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.

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 four benchmark ZSL datasets, namely AwA, CUB, SUN and ImageNet.

2. Related work

Inspired by the human-being's ability that can intelligently apply learned experiential knowledge to help novel recognition tasks, ZSL performs to classify unseen classes with the class semantic descriptions. As a pioneer work, [19] first proposes to embed seen images to its corresponding class descriptions to learn an embedding protocol. Motivated by this embedding formulation, extensive methods are proposed to establish the embedding relationship between visual space and semantic space to achieve the knowledge transformation from seen classes to unseen ones.

Generally, existing ZSL methods can be divided into three groups according to the embedding mechanism. (1) *Semantic embedding models*. Among these methods, the image features are projected from visual space to semantic space in different approaches. Direct attribute prediction (DAP) [6] is one of the most popular methods which utilizes Bayesian formulation to learn embedding model. Specifically, it first trains SVMs supervised by attribute information in seen classes and then infers the unseen classes with the combination of the posterior probability for each attribute. ConSE [20] constructs the classifier in unseen classes with convex combination of label embedding vectors. In addition, Linear Regression [21] or Neural Network [22] are also utilized to establish the semantic projection. However, such directional embedding methods easily suffer from the hubness issue [23], which is caused by the presence of "universal" unseen instances, or hubs in high-dimensional space. (2) *Visual embedding models*. To alleviate the hubness problem, visual embedding is introduced to ZSL, where the semantic space is mapped to visual space, that is contrary to semantic embedding approaches. Yutaro et al. [24] propose using ridge regression to construct visual embedding. (3) *Common embedding models*. Instead of semantic embedding and visual embedding methods, the embedding relationship can be established by joint exploiting a common latent or physical space. For example, Akata et al. [4] embed supervised attribute information and unsupervised text, hierarchical relationships into joint embedding semantic space. Using deep convolutional neural network, Jimmy et al. [25] transform the visual and semantic features into joint embedding space.

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

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

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