Signal Processing 138 (2017) 265–272

Contents lists available at ScienceDirect

Signal Processing

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

Zero-shot learning with Multi-Battery Factor Analysis

Zhong Ji[a], Yunlong Yu[a], Yanwei Pang[a,*], Lei Chen[a], Zhongfei Zhang[b]

[a] School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072, China
[b] Department of Computer Science, State University of New York, Binghamton, NY 13902, USA

ARTICLE INFO

Article history:
Received 26 October 2016
Revised 30 January 2017
Accepted 22 March 2017
Available online 28 March 2017

Keywords:
Zero-shot learning
Multi-battery factor analysis
Image classification
Attribute
Word vector

ABSTRACT

Zero-shot learning (ZSL) extends the conventional image classification technique to a more challenging situation where the testing image categories are not seen in the training samples. Most studies on ZSL utilize side information such as attributes or word vectors to bridge the relations between the seen classes and the unseen classes. However, existing approaches on ZSL typically exploit a shared space for each type of side information independently, which cannot make full use of the complementary knowledge of different types of side information. To this end, this paper presents an MBFA-ZSL approach to embed different types of side information as well as the visual feature into one shared space. Specifically, we first develop an algorithm named Multi-Battery Factor Analysis (MBFA) to build a unified semantic space, and then employ multiple types of side information in it to achieve the ZSL. The close-form solution makes MBFA-ZSL simple to implement and efficient to run on large datasets. Extensive experiments on the popular AwA, CUB, and SUN datasets show its superiority over the state-of-the-art approaches.

1. Introduction and related work

Zero-shot learning (ZSL) aims at solving the problem when the new testing image categories are not seen in the training samples [1]. Different from the open set recognition and novelty detection which only distinguish abnormalities in the testing data, ZSL seeks to classify the unseen testing classes [2,3].

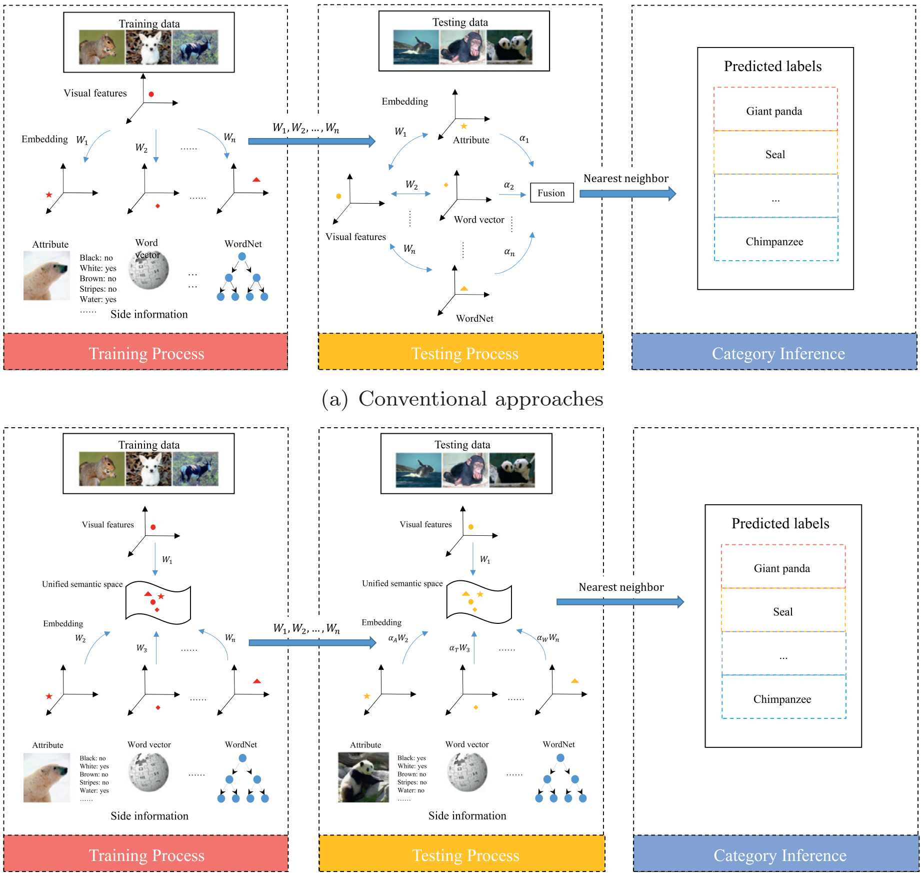
The idea of ZSL is to emulate the ability of humans to recognize new categories without having to look at actual visual samples. Humans have this ability because they are able to relate unseen categories to previously seen categories through semantic information. For example, if a child is able to recognize a horse and has never seen a unicorn, but he was told that unicorn and horse are similar, apart from the unicorn has a long horn in the head. Then, the child is very likely to accurately identify a unicorn the first time it is seen. Similarly, a ZSL system builds a projection between the visual feature space and semantic space through the multimodal projection learned from the labeled training data, i.e., seen categories, and then based on this projection relationship, the testing data (i.e., unseen category) is associated with the semantic feature of the class to be predicted, so that the class of the testing data can be predicted as the category corresponding to the nearest semantic feature.

ZSL is a practical problem setting in image classification [4] and image labeling [5], as there are thousands of categories of objects we intend to recognize, but only a few of them may have been appropriately annotated. Consequently, it is more challenging than the conventional image classification problems. The key ideas of ZSL are to choose better side information (also known as modalities) and to develop an effective common semantic space. The side information provides a bridge to transfer knowledge from the seen classes for which we have training data to the unseen classes for which we do not, and the common space offers a fusion feasibility for the visual features and the side information.

Two types of commonly used side information in ZSL are attributes [6–9] and word vectors [10,11]. Particularly, attributes act as intermediate representations shared across multiple classes, indicating the presence or absence of several predefined properties. Direct attribute prediction (DAP) [6] is one of the first efforts to exploit the attributes to ZSL. It learns attribute-specific classifiers with the seen data and infers the unseen class with the learned estimators. However, attribute-based approaches suffer from a poor scalability as the attributes ontology for each class is generally manually defined. Word-vector-based approaches [12–15] avoid this limitation since word vectors are extracted from a linguistic corpus with neural language models such as GolVe [10] and Word2Vec [11]. Therefore, word vectors have become another popular side information in ZSL. For instance, Socher et al.

* Corresponding author
E-mail addresses: jizhong@tju.edu.cn (Z. Ji), xieyuzhong@tju.edu.cn (Y. Yu), pyw@tju.edu.cn (Y. Pang), article.com.cn@126.com (L. Chen), zhongfei@cs.binghamton.edu (Z. Zhang).

http://dx.doi.org/10.1016/j.sigpro.2017.03.023
0165-1684/© 2017 Elsevier B.V. All rights reserved.



(a) Conventional approaches

(b) MBFA-ZSL approach

Fig. 1. The comparative illustration of the proposed MBFA-ZSL and the conventional approaches. (a) Conventional approaches embed the visual features to each type of side information in its own semantic space independently, (b) MBFA-ZSL employs multiple types of side information in a unified space.

[13] constructed a two layers neural network to project images into the word vector space. In [15], Frome et al. presented a deep visual-semantic embedding model with a hinge loss function, which trains a linear mapping to link the image visual space to the word vector space.

Besides attributes and word vectors, some other side information, such as WordNet [16], visual prototypical concepts [17], class co-occurrence statistics [18], is also applied in ZSL. Further, since different types of side information captures different aspects of the structure of the semantic space, several studies have been made to combine them to achieve higher classification performance [16,19–21]. For example, in [16], Akata et al. first learned the joint embedding weight matrices corresponding to different types of side information, then performed a grid search over the coefficients on a validation set to get the joint compatibility model. In [19], semantic projections are trained for attributes and word vectors independently, followed by a transductive multi-view semantic embedding space to alleviate the projection domain shift problem. In [21], Zhang et al. viewed each source or target data as a mixture of seen class proportions and assumed that instances belonging to the same unseen class have similar mixture patterns. Thus a source/target embedding functions are learned to map an arbitrary source/target domain data into a unified space where similarity can be readily measured. Further, Zhang and Saligrama [32] also developed a joint discriminative learning framework by jointly learning the parameters for both domains with the idea of dictionary learning. These efforts demonstrate that different types of side information complement each other and construct a better embedding space for knowledge transfer. However, although multiple types of side information are utilized, they still exploit each type of side information in its own semantic space independently, and then just combine the predicted scores together [16]. This cannot make full use of the complementary knowledge of different types of side information. A more efficient and robust solution is to investigate multiple types of side information in a unified space. Unfortunately, to the best of our knowledge, there has been little previous work exploiting this idea. To this end, we present a novel approach called MBFA-ZSL to employ multiple types of side information in a unified space, as shown in Fig. 1.

It is worth highlighting several aspects of the proposed MBFA-ZSL approach. (1) It develops an advanced multi-view embedding algorithm named Multi-Battery Factor Analysis (MBFA), which extends Tucker's Inter-Battery Factor Analysis (IBFA) [22]. (2) As far as we know, it represents one of the first attempts that embeds both the image visual features and multiple types of side information into one unified semantic space, which fully utilizes the interrelations among different types of information. (3) The close-form solution makes it simple to implement and efficient to run on large datasets. (4) Extensive experiments on popular datasets demonstrate its superiority over the state-of-the-art approaches.

The reminder of this paper is structured as follows. Section 2 introduces the proposed Multi-Battery Factor Analysis (MBFA) algorithm, and Section 3 describes the proposed MBFA-ZSL approach in detail. Experimental results are presented inSection 4, and conclusions are drawn in the final section.