Pattern Recognition 80 (2018) 129–142

Contents lists available at ScienceDirect

Pattern Recognition

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

Attribute-Based Synthetic Network (ABS-Net): Learning more from pseudo feature representations

Jiang Lu[a,b], Jin Li[a], Ziang Yan[a], Fenghua Mei[b], Changshui Zhang[a,*]

[a] Department of Automation, Tsinghua University, State Key Laboratory of Intelligent Technologies and Systems, Tsinghua National Laboratory for Information Science and Technology (TNList), Beijing, China
[b] China Marine Development and Research Center (CMDRC), Beijing, China

ARTICLE INFO

Article history:
Received 20 September 2017
Revised 30 November 2017
Accepted 4 March 2018
Available online 6 March 2018

Keywords:
Pseudo feature representation
Zero-shot learning
Supervised learning
Data augmentation
Attribute learning

ABSTRACT

In large-scale visual recognition tasks, researchers are usually faced with some challenging problems, such as the extreme imbalance in the number of training data between classes or the lack of annotated data for some classes. In this paper, we propose a novel neural network architecture that automatically synthesizes pseudo feature representations for the classes in lack of annotated images. With the supply of semantic attributes for classes, the proposed *Attribute-Based Synthetic Network (ABS-Net)* can be applied to zero-shot learning (ZSL) scenario and conventional supervised learning (CSL) scenario as well. For ZSL tasks, the pseudo feature representations can be viewed as annotated feature-level instances for novel concepts, which facilitates the construction of unseen class predictor. For CSL tasks, the pseudo feature representations can be viewed as products of data augmentation on training set, which enriches the interpretation capacity of CSL systems. We demonstrate the effectiveness of the proposed ABS-Net in ZSL and CSL settings on a synthetic colored MNIST dataset (C-MNIST). For several popular ZSL benchmark datasets, our architecture also shows competitive results on zero-shot recognition task, especially leading to tremendous improvement to state-of-the-art mAP on zero-shot retrieval task.

1. Introduction

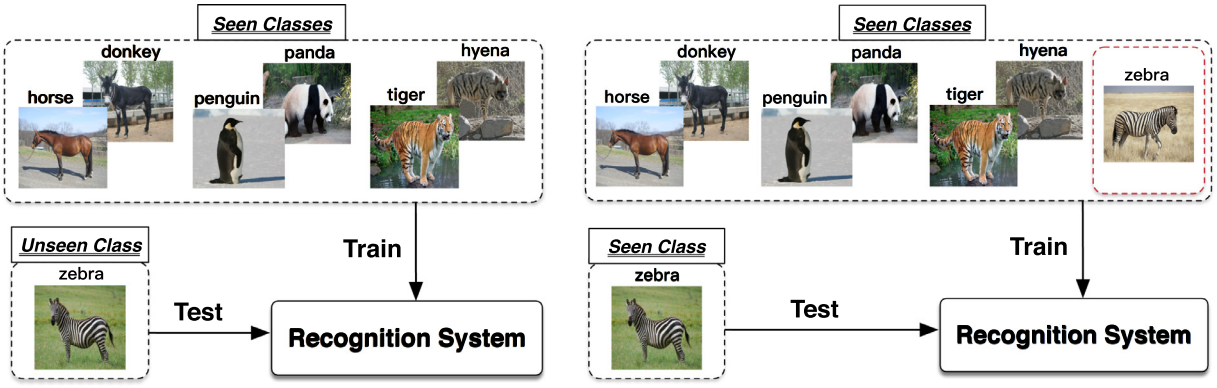
Although large scale classification based on supervised learning has achieved major successes in recent years by deep learning [1–5], the collection and annotation of huge amounts of training data for growing classes are time consuming and expensive, which consequently raises a dilemma of the lack in training data for some classes, forming the 'long-tail' phenomenon in the distribution for the number of training data between classes. The extreme imbalance in training data size between classes or the lack of annotated data for some classes are forcing us to develop more efficient learning paradigms [6–8].

As we all known that collecting semantic attributes or distinctive descriptions could be more easier than collecting massive real images for newly defined classes. As a practicable solution, zero-shot learning (ZSL) has gained growing attention recently [9–12], which aims to recognize previously unseen classes without labeled training data. For ZSL tasks, some intermediate semantic properties, such as attributes [9,13] or category hierarchies [14], are usually revealed and shared for seen and unseen classes, acting as side information by which the unseen classes could be inferred rationally. Explicitly, the dataset of seen classes is well labeled with categories and attributes tags in ZSL settings, whereas the unseen classes are faced with absence of training data but presence of their attribute descriptions. As illustrated in Fig. 1, the purpose of ZSL for visual recognition is to predict for each novel image which of unseen classes it belongs to. Due to the disjoint and unrelated classes for ZSL, the gaps between the distribution of training data (seen classes) and test data (unseen classes) are usually considered to exist [11,15,16]. In contrast, conventional supervised learning (CSL) [17] aims to predict the class labels of novel images for seen classes.

It is generally known that the process to recognize novel concepts for most of people is abstractively from individuality to generality, and then from generality to individuality. More specifically, assuming the fundamental understandings about what features represent the attributes *horselike/stripe/black and white* are obtained from some prepared images, like *horse, donkey, tiger, hyena, penguin, panda* etc, one can surmise roughly what a zebra looks like if told zebra has above attributes. Inspired by humans' behaviors of recognizing a novelty, in this paper we present a novel

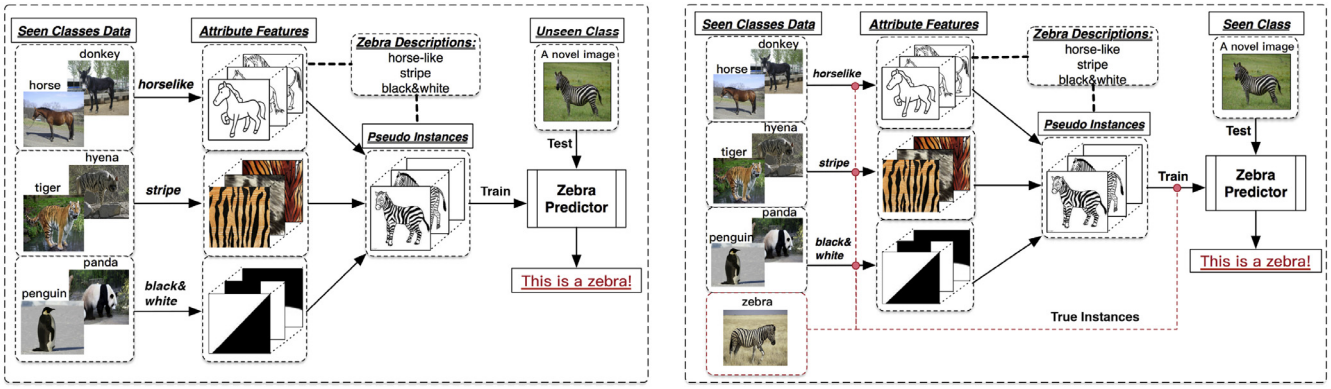
* Corresponding author.
E-mail addresses: lu-j13@mails.tsinghua.edu.cn (J. Lu), lijin14@mails.tsinghua.edu.cn (J. Li), yza15@mails.tsinghua.edu.cn (Z. Yan), mei_fh@21cn.com (F. Mei), zcs@mail.tsinghua.edu.cn (C. Zhang).

https://doi.org/10.1016/j.patcog.2018.03.006
0031-3203/



(a) Scenario of zero-shot learning. (b) Scenario of conventional supervised learning.

Fig. 1. Comparison of zero-shot learning and conventional supervised learning.



(a) Illustration of our proposed method for ZSL task. (b) Illustration of our proposed method for CSL task.

Fig. 2. Illustration of our proposed method. We capture some understandings from prepared dataset about what features represent *horselike/stripe/black and white* respectively. These attribute features associated with key descriptions of zebra will be combined into lots of synthetic pseudo instances about zebras' appearance, which can be used for the training of zebra predictor.

neural network architecture that automatically synthesize pseudo feature representations for the classes in lack of annotated images, called *Attribute-Based Synthetic Network (ABS-Net)*. Given the semantic attribute descriptions, our ABS-Net can be applied to ZSL tasks and CSL tasks as well. As illustrated in Fig. 2, our intuition is to firstly learn some credible feature representations for each attribute by utilizing prepared dataset of seen classes, and then summarize these attribute representations into an combined representation, according to the specified attribute descriptions of each unseen class (ZSL) or seen class (CSL). Finally, the combined representation can be viewed as a so-called pseudo feature-level instance of the unseen class (ZSL) or seen class (CSL), which will offer valuable guidance for the training of overall recognition system.

Leveraging the Convolutional Neural Network (CNN) based image features [3], we firstly train a Joint Attribute Feature Extractor (JAFE) in which each fundamental unit is put in charge of the extraction of one attribute feature. Regardless of labels of seen classes, we extract all possible feature vectors for every attribute tag by JAFE, then store these attribute features into a Fragment Repository based on a confidence factor filter. According to the attribute descriptions of one specified class (unseen class for ZSL and seen class for CSL), a probability based sampling strategy is exploited to select some attribute features from Fragment Repository to synthesize combined vectors. This strategy allows us access to lots of synthetic feature vectors for the specified class, called pseudo feature representations, which fills the gaps between training domain and test domain for ZSL and achieves data augmentation in feature level as well for CSL from another perspective. Taking these pseudo feature representations as inputs, a multi-way predictor for some specified classes can be learned. During test time, a novel image goes through the JAFE to generate its combined feature vector over all attributes, followed by above well-trained predictor to perform the ultimate inference. Experimental results on a synthetic colored MNIST dataset (C-MNIST) demonstrate the effectiveness of our ABS-Net in ZSL and CSL settings. Furthermore, its performance on several ZSL benchmark datasets show improvements to state-of-the-art results, especially for zero-shot retrieval mAP (i.e. mean average precision).

We conclude our contributions as follows. First, we develop an concise architecture ABS-Net to deal with ZSL tasks, which fills the gaps between seen and unseen concepts and achieves competitive results on several challenging ZSL benchmark datasets. Second, our ABS-Net realizes inherently the data augmentation in feature level, which can be easily extended to CSL scenario.

2. Related work

2.1. Zero-shot learning

Some ZSL methods are based on a two-stage recognition: attribute prediction and classification inference. [18–20] regarded each unseen class as a binary vector, i.e. signature, where each entry delegates the presence or absence of one attribute. For a