Transductive Multi-label Zero-shot Learning

Yanwei Fu, Yongxin Yang
{y.fu,yonxin.yang}@qmul.ac.uk
Timothy Hospedales, Tao Xiang
{t.hospedales,t.xiang}@qmul.ac.uk
Shaogang Gong
{s.gong}@qmul.ac.uk

School of EECS
Queen Mary University of London
London, E1 4NS, UK

Zero-shot learning has received increasing interest as a means to alleviate the often prohibitive expense of annotating training data for large scale recognition problems. These methods have achieved great success via learning intermediate semantic representations in the form of attributes and more recently, semantic word vectors. However, many real-world data are intrinsically multi-label. For example, an image on Flickr often contains multiple objects with cluttered background, thus requiring more than one label to describe its content. And different labels are often correlated (e.g. cows often appear on grass). In order to better predict these labels given an image, the label correlation must be modelled: for n labels, there are 2^n possible multi-label combinations and to collect sufficient training samples for each combination to learn the correlations of labels is infeasible.

It is thus surprising to note that there is little if any existing work for general multi-label zero-shot learning. Is it because there is a trivial extension of existing single label ZSL approaches to this new problem? By assuming each label is independent from one another, it is indeed possible to decompose a multi-label ZSL problem into multiple single label ZSL problems and solve them using existing single label ZSL methods. However this does not exploit label correlation, and we demonstrate in this work that this naive extension leads to very poor label prediction for unseen classes. Any attempt to model this correlation, in particular for the unseen classes with zero-shot, is extremely challenging.

In this paper, a novel framework for multi-label zero-shot learning is proposed. Our framework is based on transfer learning – given a training/auxiliary dataset containing labelled images, and a test/target dataset with a set of unseen labels/classes (i.e. none of the labels appear in the training set), we aim to learn a multi-label classification model from the training set and generalise/transfer it to the test set with unseen labels. This knowledge transfer is achieved using an intermediate semantic representation in the form of the skip-gram word vectors [3] which allows vector-oriented reasoning. Such a reasoning is critical for our zero-shot multi-label prediction to synthesise label combination prototypes in the semantic word space. For example, $Vec(\text{'}Moscow')$ should be much closer to $Vec(\text{'}Russia') + Vec(\text{'}capital')$ than $Vec(\text{'}Russia')/Vec(\text{'}capital')$ only. For this purpose, we employ the skip-gram language model to learn the word space, which has shown to be able to capture such syntactic regularities. This representation is shared between the training and test classes, thus making the transfer possible.

More specifically, our framework has two main components: multi-output deep regression (Mul-DR) and zero-shot multi-label prediction (ZS-MLP). Mul-DR is a 9 layer neural network that exploits the widely used convolutional neural network (CNN) layers, and includes two multi-output regression layers as the final layers. It learns from auxiliary data the explicit and direct mapping from raw image pixels to a linguistic representation defined by the skip-gram language model [3]. With Mul-DR, each test image is now projected into the semantic word space where the unseen labels and their combinations can be represented as data points without the need to collect any visual data. ZS-MLP aims to address the multi-label ZSL problem in this semantic word space. Specifically, we note that in this space any label combination can be synthesised. We thus exhaustively synthesise the power set of all possible prototypes (i.e., combinations of multi-labels) to be treated as if they were a set of labelled instances in the space. With this synthetic dataset, we are able to extend conventional multi-label algorithms, to propose two new multi-label algorithms – direct multi-label zero-shot prediction (DMP) and transductive multi-label zero-shot prediction (TraMP). However, since Mul-DR is learned using the auxiliary classes/labels, it may not generalise well to the unseen classes/labels. To overcome this problem, we further exploit self-training to adapt the Mul-DR to the test classes to improve its generalisation capability.

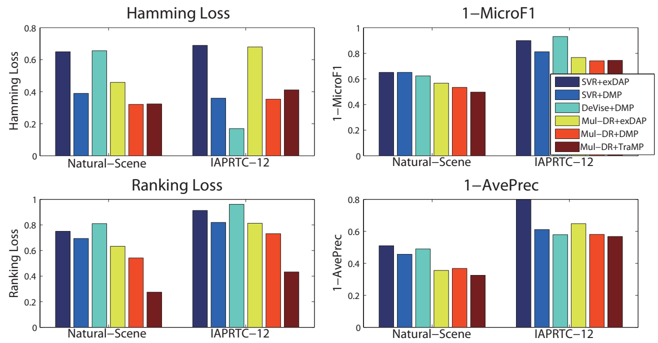

Figure 1: Comparing different zero-shot multi-label classification methods on Natural Scene and IAPRTC-12.So smaller values for all metrics are preferred.

Experiments. We conduct the experiments on several widely used benchmark multi-label datasets – Natural Scene and IAPRTC-12 are used to evaluate our framework. **Natural Scene** consists of 2000 natural scene images where each is labelled as any combinations of *desert*, *mountains*, *sea*, *sunset* and *trees*. We use a multi-class single label dataset – Scene dataset [4] (totally 2688 images) as the auxiliary dataset which are labelled with a non-overlapping set of labels such as *street*, *coast* and *highway*. **IAPRTC-12** consists of 20000 images and a total of 275 different labels. The labels are hierarchically organised into 6 main branches: humans, animals, food, landscape-nature, man-made and other. Our experiments consider the subset of landscape-nature branch (around 9500 images) and use the top 8 most frequent labels from this branch with over 30% of multi-label test images. For this dataset, we employ both Scene and Natural Scene as the auxiliary dataset.

The results in Fig 1 show the efficacy of our framework for for multi-label ZSL over a variety of baselines: (1)We first compare our Mul-DR with the alternative SVR [2] and DeViSE [1] model for learning the projection from raw images to the semantic word space. It is evident that our Mul-DR significantly improve the results on conventional SVR regression model (Mul-DR+DMP>SVR+DMP, Mul-DR+exDAP>SVR+exDAP). (2)Comparing to the DeViSE model (Mul-DR+DMP vs. DeViSE+DMP), our regression model is also clearly better using three of the four evaluation metrics, suggesting that direct and explicit mapping between the image space and the semantic word space is a better strategy. For more detailed discussions, please read our paper. All the data/codes can be downloaded from
`http://www.eecs.qmul.ac.uk/~yf300/multilabelZSL/`.

[1] Andrea Frome, Greg S. Corrado, Jon Shlens, Samy Bengio, Jeffrey Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model andrea. In *NIPS*, 2013.

[2] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[3] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[4] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42, 2001.