Queen Mary
University of London

Transductive multi-label zero-shot learning

Fu, Y; Yang, Y; Hospedales, T; Xiang, T; Gong, S

For additional information about this publication click this link.
http://qmro.qmul.ac.uk/jspui/handle/123456789/6424

Transductive Multi-label Zero-shot Learning

Yanwei Fu, Yongxin Yang
{y.fu,yongxin.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

Abstract

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, they have thus far been constrained to the single-label case, in contrast to the growing popularity and importance of more realistic multi-label data. In this paper, for the first time, we investigate and formalise a general framework for multi-label zero-shot learning, addressing the unique challenge therein: how to exploit multi-label correlation at test time with no training data for those classes? In particular, we propose (1) a multi-output deep regression model to project an image into a semantic word space, which explicitly exploits the correlations in the intermediate semantic layer of word vectors; (2) a novel zero-shot learning algorithm for multi-label data that exploits the unique compositionality property of semantic word vector representations; and (3) a transductive learning strategy to enable the regression model learned from seen classes to generalise well to unseen classes. Our zero-shot learning experiments on a number of standard multi-label datasets demonstrate that our method outperforms a variety of baselines.

1 Introduction

There are around 30,000 human-distinguishable basic object classes [1] and many more subordinate ones. A major barrier to progress in visual recognition is thus collecting training data for many classes. Zero-shot learning (ZSL) strategies have therefore gained increasing interest as a route to side-step this prohibitive cost, as well as enabling potential new categories emerging over time to be represented and recognised. To classify instances from a class with no examples, ZSL exploits knowledge transferred from a set of seen (auxiliary) classes to unseen (test) classes, typically via an intermediate semantic representation such as attributes. This has recently been explored at large scale on ImageNet [7, 28].

Prior zero-shot learning methods have assumed that class labels on each instance are mutually exclusive, i.e., multi-class single label classification. Nevertheless 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.

There is an even more acute need for zero-shot learning in the case of multi-label classification. This is because 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. However, 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 on 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 [22, 23] learned from linguistic knowledge bases. 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 [27], 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 [22, 23]. 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 [13, 17, 32, 34], 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.

2   Related Work

Multi-label classification   Multi-label classification has been widely studied – for a review of the field please see [32, 34]. Most previous studies assume plenty of training data. Recently efforts have been made to relax this assumption. Kong *et al.* [17] studied transductive multi-label learning with a small set of training instances. Hariharan *et al.* [13] explored the label correlations of auxiliary data via a multi-label max-margin formulation and bet-

ter incorporated such label correlations as prior for multi-class zero-shot learning problem. However, none of them addresses the multi-label zero-shot learning problem tackled in this work.

Zero-shot learning Multi-class single label zero-shot learning has now been widely studied using attribute-based intermediate semantic layers [3, 6, 10, 11, 19, 25] or data-driven [8, 9, 21, 30] representations. However attribute-based strategies have limited ability to scale to many classes because the attribute ontology has to be manually defined. To address this limitation, Socher *et al.* [30] first employed a linguistic model [15] as the intermediate semantic representation. However, this does not model the syntactic and semantic regularities in language [23] which allows vector-oriented reasoning. Such a reasoning is critical for our ZS-MLP 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 $Vec(\text{``Russia''})$ or $Vec(\text{``capital''})$ 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 [22, 23]. Frome *et al.* [7] also used the skip-gram language model. They learned a visual-semantic embedding model – DeViSE model for single label zero-shot learning by projecting both visual and semantic information of auxiliary data into a common space. However there are a number of fundamental differences between their work and ours: (1) Comparing the DeViSE model with our Mul-DR, the learning of the mapping between images and the semantic word space by Mul-DR is more explicit and direct. We show in our experiments that this leads to better projections and thus better classification performance. (2) Our Mul-DR can generalise better to the unseen test classes thanks to our self-training based transductive learning strategy. (3) Most critically, we address the multi-label ZSL problem whilst they only focused on the single label ZSL problem. Additionally, zero-shot learning can be taken as the generalisation of class-incremental learning (C-IL) [4, 35] or life-long learning [26].

Our Contributions Overall, we make following contributions: (1) As far as we know this is the first work that addresses the multi-label zero-shot learning problem. (2) Our multi-output deep regression framework exploits correlations across dimensions while learning the direct mapping from images to intermediate skip-gram linguistic word space. (3) Within the linguistic space, two algorithms are proposed for multi-label ZSL. (4) We propose a simple self-training strategy to make the deep regression model generalise better to the unseen test classes. (5) Experimental results on benchmark multi-label datasets show the efficacy of our framework for multi-label ZSL over a variety of baselines.

3 Methodology

3.1 Problem setup

Suppose we have two datasets – source/auxiliary and target/test. The auxiliary dataset $S = \{X_S, Y_S, L_S, \mathcal{W}_S\}$ has n_S training instances and test dataset $T = \{X_T, Y_T, L_T, \mathcal{W}_T\}$ has n_T test instances. We use $\mathcal{S} = \{1, \cdots, n_S\}$ and $\mathcal{U} = \{n_S + 1, \cdots, n_T + n_S\}$ to denote the index set for instances in auxiliary and test dataset. $X_S = \{\mathbf{x}_1, \cdots, \mathbf{x}_{n_S}\}$ and $X_T = \{\mathbf{x}_{n_S+1}, \cdots, \mathbf{x}_{n_S+n_T}\}$ are the raw image data of all auxiliary and test instances respectively. $Y_S = [\mathbf{y}_1, \cdots, \mathbf{y}_{n_S}]$ and $Y_T = [\mathbf{y}_{n_S+1}, \cdots, \mathbf{y}_{n_S+n_T}]$ are the intermediate semantic representations of each auxiliary and test instance – in our case $\mathbf{y}_i$ is a 100 dimensional continuous word vector for instance i in the skip-gram language model [23] space. $L_s = [\mathbf{l}_1, \cdots, \mathbf{l}_{n_S}]$ and $L_T = [\mathbf{l}_{n_S+1}, \cdots, \mathbf{l}_{n_S+n_T}]$ are the label vectors for auxiliary and test dataset to be predicted respectively.

The possible *textual* labels for each instance in L_S and L_T are denoted $\mathcal{W}_S = \{w_1, \cdots, w_{m_S}\}$ and $\mathcal{W}_T = \{w_{m_S+1}, \cdots, w_{m_S+m_T}\}$ respectively, where m_S and m_T are the total number of classes/labels in each dataset. Given a label-space of m_T binary labels, an instance $\mathbf{x}_i$ can be tagged with any of the 2^{m_T} possible label subsets, $\mathbf{l}_i \in \{0,1\}^{2^{m_T}}$, where $\mathbf{l}_{ij} = 1$ means instance i has label j, and $\mathbf{l}_{ij} = 0$ means otherwise. Denoting the power sets of textual labels $\mathcal{W}_S$ and $\mathcal{W}_T$ as $\mathcal{P}(\mathcal{W}_S)$ and $\mathcal{P}(\mathcal{W}_T)$, for multi-label classification we need to find the optimal class label set column vector $\mathbf{l}_i$ for the $i-th$ test instance in the power set space $\mathcal{P}(\mathcal{W}_T)$. At training time $X_S, Y_S, L_S, \mathcal{W}_S$ are all observed. At test time only new class names $\mathcal{W}_T$ and images X_T are given, their representation Y_T and multi-label vectors L_T are to be predicted.

3.2 Learning a semantic word space

The semantic representations Y_S and Y_T are the projection of each instance into a linguistic word vector space $\mathcal{V}$. The semantic word vector space is learned by using the state-of-the-art skip-gram language model [22, 23] on all English Wikipedia articles[1]. The space $\mathcal{V}$ represents almost all available English vocabulary and thus is potentially much more effective than human annotators to measure subtle similarities and differences between any two textual labels. Furthermore, $\mathcal{V}$ encodes the syntactic and semantic regularities in language [23] which allows vector-oriented reasoning by its 'compositionality' property. This property enables the critical capability of synthesising the exhaustive set of test label combinations $\mathcal{P}(\mathcal{W}_T)$. Note that cosine distance is used in the space $\mathcal{V}$ because of its robustness against noise [22, 23]. We use $v : \mathcal{W} \to \mathcal{V}$ to represent the skip-gram projection from textual concepts (words) in $\mathcal{W}$ to vectors in $\mathcal{V}$. Such a semantic space thus captures the correlations between labels without any need to collect visual examples – the meaning of multiple labels for one instance can be inferred by the sum of the word vector projections of its individual labels. Formally, we have

$$Y_S = v(\mathcal{W}_S) \cdot L_S, \qquad Y_T = v(\mathcal{W}_T) \cdot L_T \qquad (1)$$

where $v(\mathcal{W}_S)$ and $v(\mathcal{W}_T)$ are the word vector projections of the label class sets in the auxiliary and test datasets respectively. The next section discusses how to learn a predictive model for Y_T given visual data X_T.

3.3 Multi-output deep regression

We design a multi-output deep regression (Mul-DR) model $f : \mathcal{X} \to \mathcal{V}$ to predict the semantic representation $Y_T \in \mathcal{V}$ from images $X_T \in \mathcal{X}$ where $\mathcal{X}$ is the space of raw image pixel intensity values. Our Mul-DR is inspired by the recent success of the deep convolutional neural network (CNN) features [18, 29] as well as the importance of modelling correlations within the semantic representation. The Mul-DR model is a neural network composed of nine layers: Layer $1-5$ are convolutional layers; Layer $6-8$ are fully connected layers; Layer 9 is the linear mapping layer with 100 least square regressors.

Two key components contribute to the effectiveness of Mul-DR. The first component (layers 1-7) provides state-of-the-art feature extraction for many computer vision tasks [27]. It directly maps the raw image to the powerful CNN features[2], avoiding the pitful of bad

---

[1]Only articles are used without any user talk/discussion. To 13 Feb. 2014, it includes 2.9 billion words and 4.33 million vocabulary (single and bi/tri-gram words).

[2]However, it has more than 148.3 millions parameters and thus to prevent overfitting on small auxiliary dataset, ImageNet with 1.2 million labelled instances are used to train this component [29].

performance due to "wrong selection" of features for a given dataset. The second component (layers 8-9) provides the multi-output neural network (NN) regressors. Different from [18, 29], where the 8-th layer is an output layer for classification, the 8-th layer in our model is a fully connected layer of 1024 neurons with Rectified Linear Units (ReLUs) activation functions. This soft-thresholding non-linearity has better properties for generalisation than the widely used tanh activation units. Such a fully connected layer helps explore correlations among the different dimensions in the semantic word space. The final (9-th) layer of least square regressors provide an estimation of the 100 dimensional semantic representation in the space $\mathcal{V}$.

To apply this neural network, we resize all images X_S and X_T to 231×231 pixels. The parameters of the first components are pre-trained using ImageNet [29] while the parameters of the second component are trained by gradient descendent with auxiliary data X_S and Y_S. At test time, Mul-DR predicts the semantic word vector $\hat{y}_i$ for each unseen image $\mathbf{x}_i \in X_T, i \in \mathcal{U}$. Here the hat operator indicates the variable is estimated.

3.4   Zero-shot multi-label prediction

Given the estimated semantic representation $\hat{Y}_T$, we need to infer the labels $\hat{L}_T$ of the test set. A straightforward solution is to decompose the multi-label classification problem into multiple independent binary classification problems which is equivalent [14] to directly solving Eq (1) by:

$$\hat{L}_T = \left[[v(\mathcal{W}_T)]^T v(\mathcal{W}_T) \right]^\dagger [v(\mathcal{W}_T)]^T \cdot \hat{Y}_T \qquad (2)$$

where † is the Moore-Penrose pseudo-inverse. Eq (2) directly predicts the labels of each instance by a linear transformation of the intermediate representation $\hat{Y}_T$. In a way, this can be considered as an extension of the 'Direct Attribute Prediction (DAP)' [19] to the case of multi-label and continuous representation. We thus term this method exDAP. However, this does not exploit the multi-label correlations and thus has very limited expressive power [5, 33]. Hence we propose two more principled multi-label zero-shot algorithms – Direct Multi-label zero-shot Prediction (DMP) and Transductive Multi-label zero-shot Prediction(TraMP).

Direct Multi-label zero-shot Prediction (DMP)   Thanks to the compositionality property of $\mathcal{V}$, label-correlation can be explored by synthesising the representation of every possible multi-label annotations in $\mathcal{V}$: that is the power set of label vector matrix $P = v(\mathcal{P}(\mathcal{W}_T))$ where $P = [\mathbf{p}_1, \cdots, \mathbf{p}_{2^{m_T}}]$. Thus Eq (2) is replaced by a nearest neighbour (NN) classifier using all the synthesised instances as training data. The label set $\mathbf{l}_i$ of instance $i \in \mathcal{U}$ with representation $\hat{y}_i = f(\mathbf{x}_i)$ is then assigned as $\mathbf{p}_a \in v(\mathcal{P}(\mathcal{W}_T))$, where a is the index computed by

$$a = \operatorname*{argmin}_{j} \| \hat{y}_i - \mathbf{p}_j \| \qquad (3)$$

where $\| \cdot \|$ refers to the cosine distance.

Transductive Multi-label zero-shot Prediction (TraMP)   DMP can explore label correlations but only insofar as encoded by the compositionality of the prototypes in $\mathcal{V}$. It would be more desirable if the manifold structure of $\hat{Y}_T$ given test instances X_T could be used to improve multi-label zero-shot learning, i.e. via transductive learning. We therefore propose TramMP, which can be viewed as an extension the TRAM model in [17] for zero-shot learning, or a semi-supervised generalisation of Eq (3). The key idea is to use the power set of prototypes P as a known label set and to perform transductive label propagation from P

to the inferred semantic representations $\hat{Y}_T$. We denote the index of the power set proto-types as $\mathcal{L} = \{n_S + n_T + 1, \cdots, n_S + n_T + 2^{m_T}\}$ and its corresponding class label set as L_P. Specifically, we define a k-nearest neighbour (kNN) graph among the test instances $\hat{Y}_T$ and prototypes P. For any two instances i and z, where $i, z \in \{\mathcal{U}, \mathcal{L}\}$,

$$\omega_{iz} = \begin{cases} \frac{1}{Z_i} \exp\left(-\frac{\|\hat{y}_i - \hat{y}_z\|^2}{2\sigma^2}\right), & if\ z \in NN_k\left(\hat{y}_i, [\hat{Y}_T, P]\right) \\ 0 & otherwise \end{cases} \quad (4)$$

where $\sigma \approx \underset{i,z=1,\cdots,|\{\mathcal{U},\mathcal{L}\}|}{\text{median}} \| \hat{y}_i - \hat{y}_z \|^2$. $NN_k\left(\hat{y}_i, [\hat{Y}_T, P]\right)$ indicates the index set of k-nearest neighbors of $\hat{y}_i$ from $[\hat{Y}_T, P]$. $Z_i = \sum_{z \in NN_k(\hat{y}_i, [\hat{Y}_T, P])} \exp\left(-\frac{\|\hat{y}_i - \hat{y}_z\|^2}{2\sigma^2}\right)$ is the normalisation term to make sure $\sum_z \omega_{iz} = 1$. We define $A = I - \omega$ and partition the matrix A into blocks, $A = \begin{bmatrix} A_{\mathcal{L}\mathcal{L}} & A_{\mathcal{L}\mathcal{U}} \\ A_{\mathcal{U}\mathcal{L}} & A_{\mathcal{U}\mathcal{U}} \end{bmatrix}$ and the label set of test instances can be inferred by the following closed form solution [17],

$$\hat{L}_T = -A_{\mathcal{U}\mathcal{U}}^{-1} A_{\mathcal{U}\mathcal{L}} L_P. \quad (5)$$

3.5 Generalisation of multi-output deep regression

As described above, our framework consists of two key steps: applying the multi-output deep regression (Mul-DR) model to obtain the estimated semantic representation $\hat{Y}_T$, and followed by applying either DMP or TraMP to predict L_T. There is however an unsolved issue, that is, our Mul-DR is learned from the auxiliary data with a different set of labels from the target/test data. This projection model is thus not guaranteed to accurately project a test image to be near its ground truth label vector in the semantic word space. For example, if our Mul-DR is learned to project images of cat and dog to the word vector representation of "cat" and "dog" ($v(\text{"cat"})$ and $v(\text{"dog"})$), it may not accurately project an image with a person and a chair to its word vector representation of $v(\text{"person"}) + v(\text{"chair"})$ when both labels were not available for learning the Mul-DR model. Any regression model will have such a generalisation problem especially when the test data are distributed differently from the auxiliary data. To make the Mul-DR model generalise better to the target domain, we transductively exploit the predicted semantic representation $\hat{Y}_T$ to update the power set of label vector matrix P. In this way the target data would be better aligned with the synthesised label combination vectors in the semantic word space, thus helping generalise the Mul-DR to the target domain. This can be viewed as a semi-supervised learning (SSL) method starting from one instance for each label combination if the synthesised prototypes themselves are treated as instances. We therefore take a simple SSL strategy and perform one step of self-training [9] to refine each prototype of P,

$$\overline{p}_i = \frac{1}{k} \sum_{\hat{y}_T \in NN_k(p_i, \hat{Y}_T)} \hat{y}_T \quad (6)$$

where $\bar{P} = [\bar{p}_1, \cdots, \bar{p}_{2^{m_T}}]$ is the updated prototype matrix and k is the number of nearest neighbour[3] selected. We use the updated label vector matrix $\bar{P}$ to compute DMP (Eq (3)) and TramMP (Eqs (4) and (5)) in our framework.

---

[3]Note that k is not necessarily with the same k value in Eq (4).

4 Experiments

Datasets Two popular multi-label datasets – Natural Scene [33] and IAPRTC-12 [12] are used to evaluate our framework. **Natural Scene** consists of 2000 natural scene images where each image can be labelled as any combinations of *desert*, *mountains*, *sea*, *sunset* and *trees* and over 22% of the whole dataset is multi-labelled. For multi-label zero-shot learning on Natural Scene, we use a multi-class single label dataset – Scene dataset [24] (totally 2688 images) as the auxiliary dataset which have been 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 zero-shot classification on this dataset, we employ both Scene and Natural Scene as the auxiliary dataset.

4.1 Experimental setup

Evaluation metrics (a) **Hamming Loss**: it measures the percentage of mismatches between estimated and ground-truth labels; (b) **MicroF1** [16]: it evaluates both micro average of Precision (Micro-Precision) and micro average of Recall (Micro-Recall) with equal importance; (c) **Ranking Loss**: given the ranked list of predicted labels, it measures the number of label pairs that are incorrectly ordered by comparing their confidence scores with the ground-truth labels; (d) **Average precision**: given a ranked list of classes, it measures the area under precision-recall curve. These four criteria evaluate very different aspects of multi-label classification performance. Usually very few algorithms can achieve the best performance on all metrics. High values are preferred for MicroF1 and AP and vice-versa for Ranking and Hamming loss. For ease of interpretation we present $1-$MicroF1 and $1-$AP; so smaller values for all metrics are preferred.

Competitors Our full framework includes two main novel components: Mul-DR and DMP/TraMP. To evaluate the effectiveness of these two components, we define several competitors by replacing each component with possible alternatives. (1) **SVR+exDAP**: Support Vector Regression (SVR)[4] [2] is used to learn $f : \mathcal{X} \to \mathcal{V}$ and infer the representation of each test instance. Using exDAP (Eq (2)) is a straightforward generalisation of [19, 20] to multi-label zero-shot learning. (2) **SVR+DMP**: SVR replaces Mul-DR and we further use DMP (Eq (3)) for classification; thus it serves as a reference to compare DMP with exDAP. (3) **DeViSE+DMP**: We use DeViSE [7] to learn the visual-semantic embedding into which the power set P is projected. And we use Eq (3) for final labelling in the embedding space, i.e., DMP. Thus it corresponds to the extension of [7] to multi-label zero-shot learning problems. (4) **Mul-DR+exDAP**: Our Mul-DR is used to learn the visual-semantic embedding, with exDAP for multi-label classification; thus it can be used to compare Multi-DR with SVR. (5) **Mul-DR+DMP/TraMP**: Our method with either of the two proposed ZSL algorithms used. For fair comparison, all results use self-training strategy in Eq (6) to update the prototypes.

4.2 Results

Our Mul-DR model vs. alternatives The results obtained by various competitors on Natural-Scene and IAPRTC-12 are shown in Fig. 1. We first compare our Mul-DR with the alter-

---

[4]For fair comparison, we use the CNN features output by the first component (Layer 1-7) of our Mul-DR framework as the low-level feature for linear SVR used with the cost parameter set to 10.
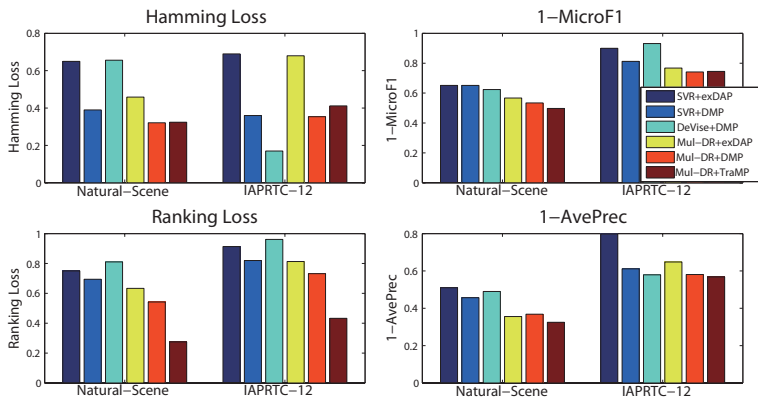
Figure 1: Comparing different zero-shot multi-label classification methods on Natural Scene and IAPRTC-12.

native SVR and DeViSE 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 [19, 20] regression model (Mul-DR+DMP>SVR+DMP, Mul-DR+exDAP>SVR+exDAP). This is because that SVR treats each of the 100 semantic word space dimensions independently, whilst our multi-output regression model, as well as the DeViSE model [7] capture the correlations between different dimensions. Comparing to the DeViSE model [7] (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. The only case where a better result is obtained by DeViSE+DMP is on the IAPCTC-12 dataset with Hamming Loss. But this result is worth further discussion. In particular, we note that Hamming Loss treats the false alarm and missing prediction errors equally. However, for multi-label classification problem, the distribution of labels is very unbalanced and each image usually has only a small portion of labels compared to the whole label set. This is particularly the case for IAPCTC-12. The good result of DeViSE on IAPCTC-12 with better Hamming loss but worse MicroF1 and Ranking Loss is an indication that it is mostly predicting no label, and biased against making any predictions. This explains the qualitative results of DeViSE shown in Table 1.

Our DMP/TraMP vs. exDAP   Given the same regression model, we compared our DAP against the alternative exDAP. The results (SVR+DMP>SVR+exDAP, Mul-DR+DMP>Mul-DR+exDAP) show that our algorithm, which is based on synthesising the label combinations in order to encode the multi-label correlations, is superior to exDAP which treats each label independently and decomposes the multi-label classification problem as multiple single label classification problems. Comparing the two proposed algorithms – DMP and TraMP, the main difference is that TraMP transductively exploits the manifold structure of the test data for label prediction. Figure 1 shows that this tranductive label prediction algorithm is better overall. Specifically, TraMP has much better Micro-F1, Ranking Loss and AP than DMP. The NN classifier (Eq (3)) used in DMP is directly minimising the Hamming Loss. This explains why TraMP is slightly worse than DMP on IAPCTC-12 on Hamming Loss.

Effectiveness of the self-training step   In this experiment we compare the results of our DMP and TraMP with and without the self-training step in Eq (6). We use '-' and '+' to indicate algorithms without and with self-training respectively. Both DMP and TraMP use
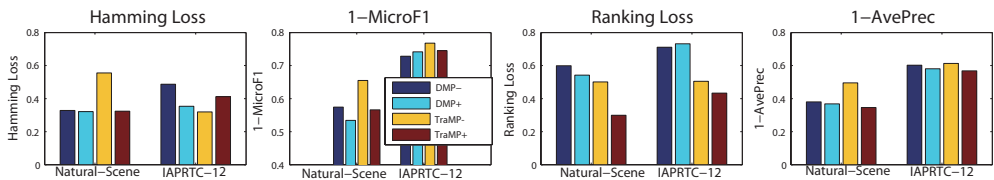
Figure 2: Effectiveness of self-training on DMP and TraMP.

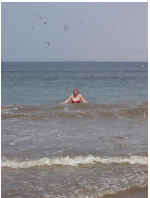|				
Groundtruth	sand-beach, mountain,sky	landscape-nature, mountain, sky	grass	sand-beach, sky
Mul-DR+DMP	sand-beach, sky	landscape-nature, mountain, sky	grass	sand-beach, sky
Mul-DR+TraMP	sand-beach, mountain, sky	landscape-nature, mountain, sky	grass, ground, landscape-nature	ground, sky, sand-beach
DeViSE+DMP	sky	–	–	sky

Table 1: Examples of multi-label zero-shot predictions on IAPRTC-12 dataset. Top 8 most frequent labels of landscape-nature branch are considered.

Mul-DR to infer the word vector $\hat{Y}_T$. As shown in Fig. 2, the self-training step clearly has a positive influence on the multi-label prediction performance. This result suggests that this simple step is helpful in making the learned Mul-DR model from the auxiliary data generalise better to the target data.

Qualitative results Table 1 gives a qualitative comparison of multi-label annotation by our DMP and TraMP with DeViSE on IAPCTC-12. As discussed, DeViSE is too conservative on this dataset and assigns no label to most instances.

5   Conclusion and future work

We have for the first time generalised zero-shot learning from the single label to the multi-label setting. It is somewhat surprising that it turns out to be possible to exploit label correlation at test time in the zero shot case – since there is no dataset of examples to learn co-occurance statistics in the conventional way. We achieve this via introducing novel strategies to exploit the compositionality of the semantic word space, and by transductively exploiting the unlabelled test data.

Besides the proposed tailor-made multi-label algorithms – DMP and TraMP, our strategy could potentially help other existing multi-label algorithms to generalise to the multi-label zero-shot learning problem. Finally, we note that many prototypes of the power set P actually have an extremely low chance to occur in the test dataset. They should not be considered in the same way as the other more likely prototypes. Thus another line of ongoing research is to investigate how to prune low-probability prototypes from the power set P.

References

[1] I. Biederman. Recognition by components - a theory of human image understanding. *Psychological Review*, 1987.

[2] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[3] Ke Chen, Shaogang Gong, Tao Xiang, and Chen Chang Loy. Cumulative attribute space for age and crowd density estimation. In *CVPR*, 2013.

[4] Qing Da, Yang Yu, and Zhi-Hua Zhou. Learning with augmented class by exploiting unlabeled data. In *AAAI*, 2014.

[5] Andre Elisseeff and Jason Weston. A kernel method for multi-labelled classification. In *NIPS*, 2001.

[6] V. Ferrari and A. Zisserman. Learning visual attributes. In *NIPS*, December 2007.

[7] 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.

[8] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Attribute learning for understanding unstructured social activity. In *ECCV*, 2012.

[9] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong. Learning multimodal latent attributes. *TPAMI*, 2013.

[10] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Zhengyong Fu, and Shaogang Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *ECCV*, 2014.

[11] Yanwei Fu, Timothy M. Hospedales, Tao Xiang, Shaogang Gong, and Yuan Yao. Interestingness prediction by robust learning to rank. In *ECCV*, 2014.

[12] Michael Grubinger. *Analysis and Evaluation of Visual Information Systems Performance*. PhD thesis, School of Computer Science and Mathematics, Faculty of Health, Engineering and Science, Victoria University, 2007.

[13] Bharath Hariharan, S. V. Vishwanathan, and Manik Varma. Efficient max-margin multi-label classification with applications to zero-shot learning. *Mach. Learn.*, 2012.

[14] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York Inc., 2009.

[15] Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *ACL*, 2012.

[16] Feng Kang, Rong Jin, and Rahul Sukthankar. Correlated label propagation with application to multi-label learning. In *CVPR*, 2006.

[17] Xiangnan Kong, M.K. Ng, and Zhi-Hua Zhou. Transductive multilabel learning via label set propagation. *Knowledge and Data Engineering, IEEE Transactions on*, 2013.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

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

[20] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE TPAMI*, 2013.

[21] Ryan Layne, Timothy M. Hospedales, and Shaogang Gong. Re-id: Hunting attributes in the wild. In *BMVC*, 2014.

[22] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representation in vector space. In *Proceedings of Workshop at ICLR*, 2013.

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

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

[25] Mark Palatucci, Geoffrey Hinton, Dean Pomerleau, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[26] Anastasia Pentina and Christoph H. Lampert. A pac-bayesian bound for lifelong learning. In *ICML*, 2014.

[27] Ali Sharif Razavian, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf : an astounding baseline for recognition. *arXiv:1403.6382v1*, 2014.

[28] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2012.

[29] Pierre Sermanet, David Eigen, Xiang Zhang, Michael Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *ICLR*, 2014.

[30] Viktoriia Sharmanska, Novi Quadrianto, and Christoph H. Lampert. Augmented attribute representations. In *ECCV*, 2012.

[31] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[32] Le Wu and Min-Ling Zhang. Multi-label classification with unlabeled data: An inductive approach. In *ACML*, pages 197–212, 2013.

[33] Min-Ling Zhang and Zhi-Hua Zhou. Ml-knn: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007.

[34] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, page 1, 2013.

[35] Zhi-Hua Zhou and Zhao-Qian Chen. Hybrid decision tree. *Knowledge-Based Systems*, 15(8):515 – 528, 2002.