Multi-Label Zero-Shot Learning with Transfer-Aware Label Embedding Projection

Meng Ye[1], **Yuhong Guo**[2],
[1] Computer and Information Sciences, Temple University, USA
[2] School of Computer Science, Carleton University, Canada
meng.ye@temple.edu, yuhong.guo@carleton.ca

Abstract

Zero-shot learning transfers knowledge from seen classes to novel unseen classes to reduce human labor of labelling data for building new classifiers. Much effort on zero-shot learning however has focused on the standard multi-class setting, the more challenging multi-label zero-shot problem has received limited attention. In this paper we propose a transfer-aware embedding projection approach to tackle multi-label zero-shot learning. The approach projects the label embedding vectors into a low-dimensional space to induce better inter-label relationships and explicitly facilitate information transfer from seen labels to unseen labels, while simultaneously learning a max-margin multi-label classifier with the projected label embeddings. Auxiliary information can be conveniently incorporated to guide the label embedding projection to further improve label relation structures for zero-shot knowledge transfer. We conduct experiments for zero-shot multi-label image classification. The results demonstrate the efficacy of the proposed approach.

1 Introduction

Despite the advances in the development of supervised learning techniques such as deep neural network models, the conventional supervised learning setting requires a large number of labelled instances for each single class to perform training, and hence induce substantial annotation costs. It is important to develop algorithms that enable the reduction of annotation cost for training classification models. Zero-shot learning (ZSL) which transfers knowledge from annotated *seen* classes to predict *unseen* classes that have no labeled data, hence has received a lot of attention [Lampert *et al.*, 2009; Akata *et al.*, 2015; Romera-Paredes and Torr, 2015; Zhang and Saligrama, 2015; Changpinyo *et al.*, 2017].

One primary source deployed in zero-shot learning for bridging the gap between seen and unseen classes is the *attribute* description of the class labels [Lampert *et al.*, 2009; Lampert *et al.*, 2014; Romera-Paredes and Torr, 2015; Fu *et al.*, 2015]. The attributes are typically defined by domain experts who are familiar with the common and specific characteristics of different category concepts, and hence are able to carry transferable information across classes. Nevertheless human labor is still involved in defining the attribute-based class representations. This propels the research community to exploit more easily accessible free information sources from the Internet, including textual descriptions from Wikipedia articles [Qiao *et al.*, 2016; Akata *et al.*, 2015], word embedding vectors trained from large text corpus using natural language processing (NLP) techniques [Akata *et al.*, 2015; Frome *et al.*, 2013; Xian *et al.*, 2016; Zhang and Saligrama, 2015; Al-Halah *et al.*, 2016], co-occurrence statistics of hit-counts from search engine [Rohrbach *et al.*, 2010; Mensink *et al.*, 2014], and WordNet hierarchy information of the labels [Rohrbach *et al.*, 2010; Rohrbach *et al.*, 2011; Li *et al.*, 2015b]. These works demonstrated impressive results on several standard zero-shot datasets. However, majority research effort has concentrated on multi-class zero-shot classifications, while the more challenging multi-label zero-shot learning problem has received very limited attention [Mensink *et al.*, 2014; Zhang *et al.*, 2016; Lee *et al.*, 2017].

In this work we propose a novel transfer-aware label embedding projection method to tackle multi-label zero-shot learning, as shown in Figure 1. Label embeddings have been exploited in standard multi-label classification to capture label relationships. We exploit the word embeddings [Pennington *et al.*, 2014] produced from large corpus with NLP techniques as the initial semantic label embedding vectors. These semantic embedding vectors have the nice property of catching general similarities between any pair of label phrases/words, but may not be optimal for multi-label classification and information transfer across classes. Hence we project the label embedding vectors into a low-dimensional semantic space in a transfer-aware manner to gain transferable label relationships by enforcing similarity between seen and unseen class labels and separability across unseen labels. We then simultaneously co-project the labeled seen class instances into the same semantic space under a max-margin multi-label classification framework to ensure the predictability of the embeddings. Moreover, we further incorporate auxiliary information to guide the label embedding projection for suitable inter-label relationships. To investigate the proposed approach, we conduct ZSL experiments on two standard multi-label image classification datasets, the PASCAL VOC2007 and VOC2012. The empirical results demonstrate

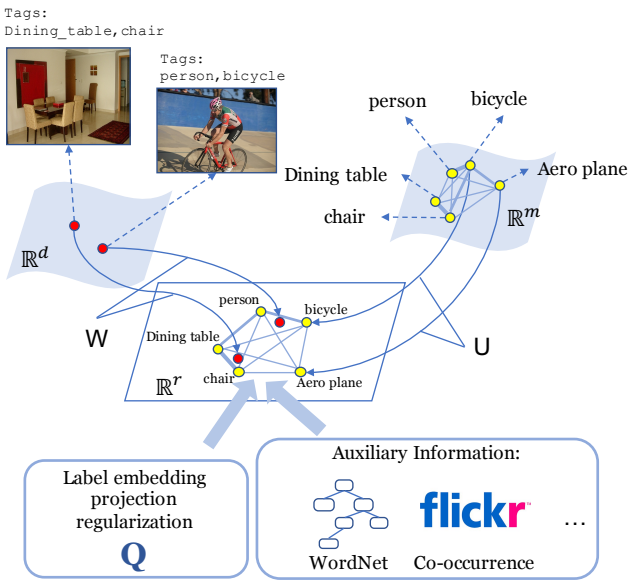

Figure 1: Illustration of the proposed multi-label ZSL framework. Red dots represent images in their visual feature space $\mathbb{R}^d$. They are mapped into a semantic space $\mathbb{R}^r$ by a visual projection matrix W. Yellow dots represent labels in the word embedding space $\mathbb{R}^m$ and they are mapped into the same $\mathbb{R}^r$ by a semantic projection matrix U. The projection matrices are learnt under a max-margin multi-label learning framework based on the matching scores of the images and labels in the projected semantic space. Embedding regularization and auxiliary information are leveraged to facilitate the knowledge transfer from seen classes to unseen classes on the projected common semantic space.

the effectiveness of the proposed approach by comparing to a number of related ZSL methods.

2 Related Work

Multi-label Classification Multi-label classification is relevant in many application domains, where each data instance can be assigned into multiple classes. Many multi-label learning works developed in the literature have centered on exploiting the correlation/interdependency information between the multiple labels, including the max-margin learning methods with pairwise ranking loss [Elisseeff *et al.*, 2001], weighted approximate pairwise ranking loss (WARP) [Weston *et al.*, 2011], and calibrated separation ranking loss (CSRL) [Guo and Schuurmans, 2011]. Moreover, incomplete labels are frequently encountered in many multi-label applications due to noise or crowd-sourcing, where only a subset of true labels are provided on some training instances. Multi-label learning methods with missing labels have largely depended on observed label correlations to overcome the label incompleteness of the training data [Bucak *et al.*, 2011; Yu *et al.*, 2014; Yang *et al.*, 2016]. These methods however assumed that all the labels are at least observed on a subset of training data and they cannot handle the more challenging zero-shot learning setting where some labels are completely missing from the training instances.

Zero-shot Learning There have been a significant number of works in multi-class zero-shot image classification, including the ones that explore different transferring embedding strategies [Romera-Paredes and Torr, 2015; Frome *et al.*, 2013; Norouzi *et al.*, 2013; Xian *et al.*, 2016] or different information sources [Akata *et al.*, 2015; Mensink *et al.*, 2014]. Many methods represent labels in a semantic attribute space [Lampert *et al.*, 2009] or word embedding space [Mikolov *et al.*, 2013; Pennington *et al.*, 2014]) to perform zero-shot learning by computing similarities between the instances and labels. [Romera-Paredes and Torr, 2015] proposed a simple approach to learn a projection matrix that maps image features into the attribute space, while [Frome *et al.*, 2013] used a CNN architecture followed by a transformation matrix to map images into the word embedding vector space. [Norouzi *et al.*, 2013] also took advantage of CNNs but they expressed image embeddings as convex combinations of seen class embeddings. [Akata *et al.*, 2015] considered learning a bilinear compatibility function for image features and output label embeddings. They evaluated attributes, word embedding vectors, as well as WordNet hierarchy and online text information, for producing label embeddings. In [Xian *et al.*, 2016], the authors proposed to use tensors as nonlinear latent embedding functions. [Li *et al.*, 2015a] learned the projection matrix by minimizing max-margin loss in a semi-supervised way. [Zhang and Saligrama, 2015] proposed to embed both image features and attribute signature of labels into a common semantic space which has the seen classes as bases. More recently, [Changpinyo *et al.*, 2017] proposed a method to generate visual examplars from semantic attributes, and then use them as optimized class prototypes for prediction on test instances. This work also projects both semantic and visual feature vectors into an intermediate space. Nonetheless, all theses methods are designed for multi-class zero-shot learning problems.

Despite the many works above on multi-class ZSL, to the best of our knowledge, there has not been much work on multi-label ZSL with the following exceptions. In [Fu *et al.*, 2014], the authors proposed to address multi-label zero-shot learning by mapping images into the semantic word space. However in testing phase it needs to consider all possible combinations of the outputs, which is the power set of unseen tags/classes. This prevents it from being applied on larger datasets. The authors of [Mensink *et al.*, 2014] proposed to express unseen class classifiers as weighted sums of seen class classifiers, while the weights are estimated from different kinds of co-occurrence statistics. This approach however treats the unseen class classifiers separately, without considering the correlations/dependencies among the classes. [Zhang *et al.*, 2016] proposed a fast zero-shot image tagging algorithm, which learns the principal direction of each image to separate tags into positive and negative ones. Their approach however uses fixed pre-given label embeddings which may not be the best for capturing useful class correlations between seen and unseen classes towards information transfer. More recently, in [Gaure *et al.*, 2017] the authors adopted a generative probabilistic framework to leverage the co-occurrence statistics of the seen labels for multi-label zero-shot prediction. This method however heavily depends on the auxil-

iary resource for gaining quality label co-occurrence statistics. [Lee *et al.*, 2017] proposed to construct a knowledge graph based on WordNet hierarchy for modeling label relations, and then propagate confidence scores from the seen to unseen labels through the graph. Its performance largely relies on the quality of the knowledge graph. By contrast, our proposed approach can project existing label embeddings into a more suitable low-dimensional semantic space to automatically retrieve better label relations for knowledge transfer between seen and unseen classes, while flexibly exploiting auxiliary information for additional help.

3 Proposed Approach

3.1 Problem Definition and Notations

We consider multi-label zero-shot learning in the following setting. Assume we have a set of n labeled training images $D = (X, Y)$, where $X \in \mathbb{R}^{n\times d}$ denotes the d-dim visual features extracted using CNNs for the n images, and $Y \in \{0,1\}^{n\times L^s}$ denotes the corresponding label indicator matrix across a set of seen classes, $\mathcal{S} = \{1, 2, ..., L^s\}$: "1" indicates the presence of the corresponding label (i.e., positive labels) and "0" indicates the absence of the corresponding label (i.e., negative labels). For multi-label classification, each row of Y can have multiple "1" values. Moreover, we also assume there are a set of L^u unseen classes, $\mathcal{U} = \{L^s + 1, ..., L\}$ such that $L = L^s + L^u$, and the labels for the unseen classes are completely missing in our labeled training data. In addition, we assume the word embeddings of the seen classes and unseen classes are both given: $M = [M^s; M^u] \in \mathbb{R}^{L\times m}$, where $M^s \in \mathbb{R}^{L^s\times m}$ are the seen class embeddings, $M^u \in \mathbb{R}^{L^u\times m}$ are the unseen class embeddings, and their concatenation M is for all the classes. We aim to learn a multi-label prediction model from the training data that allows us to perform multi-label classification on the unseen classes.

We use the following general notations in the presentation below. For any matrix, e.g., X, we use X_i to denote its i-th row vector. We use $\|\cdot\|_F$ to denote the Frobenius norm of a matrix and use tr$(\cdot)$ to denote the trace of a matrix. For Y_i, we use $\bar{Y}_i$ to denote its complement such that $\bar{Y}_i = 1 - Y_i$. We also reuse the notation Y_i to denote a set of indices of its non-zero values within proper contexts. We use $\|\cdot\|$ to denote the Euclidean norm and denote the rectified operator as $[\cdot]_+ = \max(\cdot, 0)$. We use $\mathbf{1}$ to denote a column vector of all 1s, assuming its size can be determined in the context, and use I to denote an identity matrix. We use $\mathbf{0}_{a,b}$ to denote a $a \times b$ matrix with all 0s and use $\mathbf{1}_{a,b}$ to denote a $a \times b$ matrix with all 1s.

3.2 Max-margin Multi-label Learning with Semantic Embedding Projection

Instead of entirely relying on the pre-given label embeddings in M obtained from word embeddings to facilitate cross-class information adaptation, we propose to co-project the input image visual features and the label embeddings into a more suitable common low-dimensional semantic space such that the similarity matching scores of each image with its positive labels in this semantic space will be higher than that with its negative labels. Specifically, we want to learn a projection function $\theta : \mathbb{R}^d \to \mathbb{R}^r$ that maps an instance X_i from the visual feature space $\mathbb{R}^d$ into a semantic space $\mathbb{R}^r$; assuming a linear projection we have $\theta(X_i) = X_i W$, where W is a $d \times r$ projection matrix. Simultaneously, we learn another linear projection function $\phi : \mathbb{R}^m \to \mathbb{R}^r$ such that $\phi(M_c) = M_c U$, where U is a $m \times r$ projection matrix, which maps a class c from the original word embedding space $\mathbb{R}^m$ into the same semantic space $\mathbb{R}^r$. Then the similarity matching score between an instance X_i and the c-th class label can be computed as the inner product of their project representations in the common semantic space:

$$F(i, c) = \theta(X_i)\phi(M_c)^\top = X_i W U^\top M_c^\top \tag{1}$$

To encode the assumption that the similarity score $F(i, c)$ between an instance X_i and any of its positive label $c \in Y_i$ should be higher than the similarity score $F(i, \bar{c})$ between instance X_i and any of its negative label $\bar{c} \in \bar{Y}_i$, i.e., $F(i, c) \succ F(i, \bar{c})$, we formulate the projection learning problem within a max-margin multi-label learning framework:

$$\min_{W,U:U^\top U=I} \sum_{i=1}^{n} \mathcal{L}(W, U; X_i, Y_i) + \mathcal{R}(W) \tag{2}$$

where $\mathcal{L}(\cdot)$ denotes a max-margin ranking loss and $\mathcal{R}(W)$ is a model regularization term. In this work we adopt a calibrated separation ranking loss:

$$\mathcal{L}(W, U; X_i, Y_i) = \left\{ \begin{array}{l} \max_{c\in Y_i} \left[1 + F(i, 0) - F(i, c)\right]_+ \\ + \max_{\bar{c}\in \bar{Y}_i} \left[1 + F(i, \bar{c}) - F(i, 0)\right]_+ \end{array} \right\} \tag{3}$$

where $F(i, 0) = X_i W_0$ can be considered as the matching score for an auxiliary class 0, which produces a separation threshold score on the i-th instance such that the scores for positive labels should be higher than it and the scores for negative labels should be lower than it, i.e., $F(i, c) \succ F(i, 0) \succ F(i, \bar{c})$, to minimize the loss.

We assume the project matrix U has orthogonal columns to maintain a succinct label embedding projection. For the regularization term over W, we consider a Frobenius norm regularizer, $\mathcal{R}(W) = \frac{\beta}{2}\left(\|W\|_{\mathrm{F}}^2 + \|W_0\|^2\right)$, where W_0 can be considered as an auxiliary column to W, and β is a trade-off weight parameter.

3.3 Transfer-Aware Label Embedding Projection

Employing the ranking loss to minimize classification error on seen classes can ensure the predictability of the projected label embedding. However for ZSL our goal is to predict labels from the unseen classes. This requires a label embedding representation that can encode suitable inter-class label relations to facilitate information transfer from seen to the unseen classes such that the similarity score $F(i, c)$ can well reflect the relative prediction scores on an unseen class c under the learned model parameters W and U. Our intuition is that classification or ranking on the target unseen class labels would be easier if they are well separated in the projected embedding space and knowledge transfer would be easier if unseen classes and seen classes have high similarities in the projected label embedding space. We hence propose to guide the

label embedding projection learning by encoding this intuition through a transfer-aware regularization objective $\mathcal{H}(U)$ such that:

$$\mathcal{H}(U) = \frac{\gamma}{2L^u(L^u-1)} \sum_{i,j\in\mathcal{U}, i\neq j} M_i U U^\top M_j^\top - \frac{\gamma}{2L^s L^u} \sum_{i\in\mathcal{S}, j\in\mathcal{U}} M_i U U^\top M_j^\top$$

which can be equivalently expressed in a more compact form:

$$\mathcal{H}(U) = \frac{\gamma}{2}\text{tr}\left(U^\top M^\top Q M U\right) \quad (4)$$

where γ is a balance parameter for $\mathcal{H}(\cdot)$, and

$$Q = \begin{bmatrix} \mathbf{0}_{L^s,L^s} & \frac{-1}{2L^sL^u}\mathbf{1}_{L^s,L^u} \\ \frac{-1}{2L^sL^u}\mathbf{1}_{L^u,L^s} & \frac{1}{L^u(L^u-1)}(\mathbf{1}_{L^u,L^u} - I_{L^u}) \end{bmatrix} \quad (5)$$

Here we use the inner product of a pair projected label embedding vectors as the similarity value for the corresponding pair of classes, and aim to maximize the similarities across seen and unseen classes and minimize the similarities between unseen classes. By incorporating this regularization objective into the framework in Eq.(2), we obtain the following Transfer-Aware max-margin Embedding Projection (TAEP) learning problem:

$$\min_{\substack{W,W_0,\boldsymbol{\xi},\boldsymbol{\eta},\\ U:\,U^\top U=I}} \mathbf{1}^\top\boldsymbol{\xi} + \mathbf{1}^\top\boldsymbol{\eta} + \frac{\beta}{2}(\|W\|_{\text{F}}^2 + \|W_0\|^2) + \mathcal{H}(U) \quad (6)$$

$$\text{s.t.} \quad F(i,c) - F(i,0) \geq 1 - \boldsymbol{\xi}_i, \forall c \in Y_i, \forall i;\ \boldsymbol{\xi} \geq 0;$$
$$F(i,0) - F(i,\bar{c}) \geq 1 - \boldsymbol{\eta}_i, \forall \bar{c} \in \bar{Y}_i, \forall i;\ \boldsymbol{\eta} \geq 0$$

The objective learns W and U by enforcing positive labels to rank higher than negative labels, while incorporating the regularization term $\mathcal{H}(U)$ to refine the label embedding structure in the semantic space. $\mathcal{H}(U)$ can help produce better inter-class relationship structure for cross-class knowledge transfer. The regularization form $\mathcal{H}(U)$ also has a nice property — it allows a closed-form solution for U to be derived and hence simplifies the training procedure.

Note after learning the projection matrices W and U, it will be straightforward to rank all unseen labels for instance i based on the prediction scores $F(i,c)$ for all $c \in \mathcal{U}$.

3.4 Integration of Auxiliary Information

In addition to explicit word embeddings, similarity information about the class labels can be derived from some external resources. We propose to leverage such auxiliary information to further improve label embedding projection.

In general, we can assume there is some auxiliary source in terms of a similarity matrix R over the seen and unseen labels; i.e., $R_{ij}, i,j \in \{1,2,...,L\}$ defines the similarity between a label pair (i,j). Then $Q^{\mathcal{A}} = I - D^{-1/2}RD^{-1/2}$, where $D = diag(R\mathbf{1})$, is the normalized Lapalacian matrix of R. We use a manifold regularization term to enforce the projected label embeddings to be better aligned with the inter-class affinity R:

$$\mathcal{A}(U) = \frac{\lambda}{2}\text{tr}\left(U^\top M^\top Q^{\mathcal{A}} M U\right) \quad (7)$$

where λ is a balance parameter for $\mathcal{A}(\cdot)$. This regularization form has the following advantages. First, it can be conveniently integrated into the learning framework in Eq.(6) by simply updating the regularization function $\mathcal{H}(U)$ to:

$$\mathcal{H}(U) = \frac{\gamma}{2}\text{tr}\left(U^\top M^\top (Q + \frac{\lambda}{\gamma}Q^{\mathcal{A}}) M U\right) \quad (8)$$

Second, it is convenient to exploit different auxiliary resources by simply replacing R (or $Q^{\mathcal{A}}$) with the one computed from the specific resource. In this work we study two different auxiliary information resources, WordNet [Miller, 1995] hierarchy and web co-occurrence statistics.

WordNet: WordNet [Miller, 1995] is a large lexical database of English. Words are grouped into a hierarchical tree structure based on their semantic meanings. Since words are organized based on ontology, their semantic relationships can be reflected by their connection paths. We find the shortest path between any two words based on "is-a" taxonomy, and then define the similarity between two labels i and j as the reciprocal of the path length between the corresponding words, i.e., $R_{ij} = \frac{1}{path\_len(i,j)+1}$.

Co-occurrence statistics: Many researchers have exploited the usage of online data, for example Hit-Count, to compute similarity between labels [Rohrbach *et al.*, 2010; Mensink *et al.*, 2014]. The Hit-Count $HC(i,j)$ denotes how many times in total i and j appear together in the auxiliary source – for example, the number of records returned by a search engine. It is the co-occurrence statistics of i and j in the scale of the entire World Wide Web. Following previous works, we use the Flickr Image Hit Count to compute the dice-coefficient as similarity between two labels, i.e., $R_{ij} = \frac{HC(i,j)}{HC(i)+HC(j)}$.

3.5 Dual Formulation and Learning Algorithm

With the orthogonal constraint on U and the appearance of U in both the objective function and the linear inequality constraints, it is difficult to perform learning directly on Eq.(6). We hence deploy the standard Lagrangian dual formulation of the max-margin learning problem for fixed U. This leads to the following equivalent dual formulation of Eq.(6):

$$\min_{U:U^\top U=I} \max_{\Psi}\ \text{tr}\left(\Psi^\top(2Y - \mathbf{1}\mathbf{1}^\top)\right) + \frac{\gamma}{2}(U^\top M^\top Q M U) - \frac{1}{2\beta}\text{tr}\left(\Psi^\top X X^\top \Psi\left(M^s U U^\top M^{s\top} + \mathbf{1}\mathbf{1}^\top\right)\right) \quad (9)$$

$$\text{s.t.} \quad \Psi_i \text{diag}(Y_i) \geq 0, \quad \Psi_i Y_i^\top \leq 1,\ \forall i;$$
$$\Psi_i \text{diag}(Y_i - 1) \geq 0, \quad \Psi_i (Y_i - 1)^\top \leq 1,\ \forall i$$

where the primal W and W_0 can be recovered from the dual variables Ψ by $W = \frac{1}{\beta}X^\top \Psi M^s U$ and $W_0 = \frac{-1}{\beta}X^\top \Psi \mathbf{1}$.

One nice property about the dual formulation in Eq.(9) is that it allows a convenient closed-form solution for U. To solve this min-max optimization problem, we develop an iterative alternating optimization algorithm to perform training. We start from an infeasible initialization point by setting both U and Ψ as zeros. Then in each iteration, we perform the following two steps, which will quickly move into the feasible region after one iteration.

Step 1: Given the current fixed U, the inner maximization over Ψ is a linear constrained convex quadratic programming. Though we can solve it directly using a quadratic solver, it subjects to a scalability problem– the Hessian matrix over Ψ will be very large whenever the data size n or the label size L^s is large. Hence we adopt a coordinate descent method to iteratively update each row of Ψ given other rows fixed, since the constraints over each row of Ψ can be separated. The maximization over the i-th row Ψ_i can be equivalently written as the following simple quadratic minimization problem:

$$\begin{aligned} \boldsymbol{z}^* = \arg\min_{\boldsymbol{z}} \quad & \frac{1}{2}\boldsymbol{z}^\top H\boldsymbol{z} + \boldsymbol{f}^\top \boldsymbol{z} \\ \text{s.t.} \quad & \text{diag}(Y_i)\boldsymbol{z} \geq 0, \quad \text{diag}(Y_i - 1)\boldsymbol{z} \geq 0, \\ & Y_i\boldsymbol{z} \leq 1, \quad (Y_i - 1)\boldsymbol{z} \leq 1 \end{aligned} \tag{10}$$

where $H = \frac{1}{\beta} X_i X_i^\top (M^s U U^\top M^{s\top} + \mathbf{1}\mathbf{1}^\top)$ and $\boldsymbol{f} = \mathbf{1} - 2Y_i^\top + \frac{1}{\beta}(M^s U U^\top M^{s\top} + \mathbf{1}\mathbf{1}^\top)\Psi^\top X X_i^\top$. After obtaining the optimal solution $\boldsymbol{z}^*$, we can update Ψ with $\Psi \leftarrow \Psi + \mathbf{1}_i \boldsymbol{z}^{*\top}$, where $\mathbf{1}_i$ denote a one-hot vector with a single 1 in its i-th entry and 0s in all other entries.

Step 2: After updating each row in Ψ, we fix the value Ψ and perform minimization over U. By taking a negative sign from Eq.(9), we have the following maximization problem:

$$\max_{U:U^\top U=I} \text{tr}\Big(U^\top \big(\frac{1}{2\beta} M^{s\top}\Psi^\top X X^\top \Psi M^s - \frac{\gamma}{2} M^\top Q M\big) U\Big) \tag{11}$$

which has a closed-form solution. Let $S = \frac{1}{2\beta} M^{s\top}\Psi^\top X X^\top \Psi M^s - \frac{\gamma}{2} M^\top Q M$. Then the solution for U is the top-r eigenvectors of S.

4 Experiments

To investigate the empirical performance of the proposed method, we conducted experiments on two standard multi-label image classification datasets to test its performance on multi-label zero-shot classification and generalized multi-label zero-shot classification.

4.1 Experimental Setting

Datasets In our experiments we used two standard multi-label datasets: The PASCAL VOC2007 dataset and VOC2012 dataset. The PASCAL VOC2007 dataset contains 20 visual object classes. There are 9963 images in total, 5011 for training and 4952 for testing. The VOC2012 dataset contains 5717 and 5823 images from 20 classes for training and validation. We used the validation set for test evaluation.

Detailed settings For each image, we used VGG19 [Simonyan and Zisserman, 2014] pre-trained on ImageNet to extract the 4096-dim visual features. For the label embeddings, we used the 300-dim word embedding vectors pre-trained by GloVe [Pennington *et al.*, 2014]. All image feature vectors and word embedding vectors are l_2 normalized. To determine the hyper-parameters, we further split the seen classes into two disjoint subsets with equal number of classes for training and validation. We train the model on the training set and choose hyper-parameters based on the test performance on the validation set. For the proposed model, we choose β, γ and λ from $\beta \in \{1, 2, ..., 10\}$ and $\gamma, \lambda \in \{0.01, 0.1, 1, 10\}$ respectively. After parameter selection, the training and validation data are put back together to train the model for the final evaluation on unseen test data.

Evaluation metric We used four different multi-label evaluation metrics: MiAP, micro-F1, macro-F1 and Hamming loss. The Mean image Average Precision (MiAP) [Li *et al.*, 2016] measures how well are the labels ranked on a given image based on the prediction scores. The other three standard evaluation metrics for multi-label classification measure how well the predicted labels match with the ground truth labels on the test data.

4.2 Multi-label Zero-shot Learning Results

Comparison methods We compared the proposed method with four related multi-label ZSL methods, *ConSE, LatEm-M*, *DMP* and *Fast0Tag*, which also adopted the visual-semantic projection strategy. The first two methods are the multi-label adaptations of two standard ZSL approaches, the convex combination of semantic embedding (ConSE) [Norouzi *et al.*, 2013] and the latent embedding (LatEm) method [Xian *et al.*, 2016]. For LatEm, we adopted a multi-label ranking objective to replace the original one of LatEm and denote this variant as Latent Embedding Multi-label method (*LatEm-M*). The direct multi-label zero-shot prediction method (DMP) [Fu *et al.*, 2014] and the fast tagging method (Fast0Tag) [Zhang *et al.*, 2016] are specifically developed for mulit-label zero-shot learning. For our proposed transfer-aware max-margin embedding projection (*TAEP*) method, we also provide comparisons for two TAEP variants with different types of auxiliary information: *TAEP-H* uses WordNet Hierarchy as auxiliary information, and *TAEP-C* uses Flickr Image Hit-Count as auxiliary information.

Zero-shot multi-label learning results. We divided the datasets into two subsets of equal number of classes, and then use them as seen and unseen classes respectively. All methods use seen class instances in the training set to train their models and make predictions on the unseen class instances in test set. We selected the hyper-parameters for the comparison methods based on grid search. With selected fixed parameters, for each approach we repeated 5 runs and reported its mean performance in Table 1. We can see the direct multi-label prediction method, DMP, outperforms both ConSE and LatEm-M on the two datasets in terms of almost all measures. This shows that the specialized multi-label ZSL method, DMP, does have advantage over extended multi-class ZSL methods. Fast0Tag is a bit less effective than DMP, but still consistently outperforms ConSE. The proposed TAEP on the other hand consistently outperforms all the four comparison methods across all measures and with notable improvements on both datasets. By integrating auxiliary information, the proposed TAEP-C and TAEP-H further improve the performance of the proposed model TAEP, while TAEP-C achieves the best results in terms of all measures. These

Table 1: Average comparison results (%) over five runs on zero-shot multi-label image tagging. Smaller values indicate better results in terms of Hamming loss, while larger values indicate better results in terms of the remaining measures.

Methods	VOC2007				VOC2012			
	MiAP	micro-F1	macro-F1	Hamm.	MiAP	micro-F1	macro-F1	Hamm.
ConSE	49.98	30.80	27.57	28.12	49.95	33.48	28.83	27.13
LatEm-M	52.45	35.32	36.69	26.28	51.44	35.74	36.33	26.21
DMP	53.52	36.70	40.44	25.72	52.92	35.73	41.04	26.12
Fast0Tag	52.39	35.01	36.76	26.53	52.29	34.23	35.38	26.41
TAEP	57.42	38.48	42.33	24.98	54.39	37.63	41.58	25.25
TAEP-C	**59.22**	**39.84**	**43.77**	**24.01**	**57.13**	**39.30**	**42.97**	**24.27**
TAEP-H	57.62	38.95	43.29	24.46	56.10	38.89	42.23	24.44

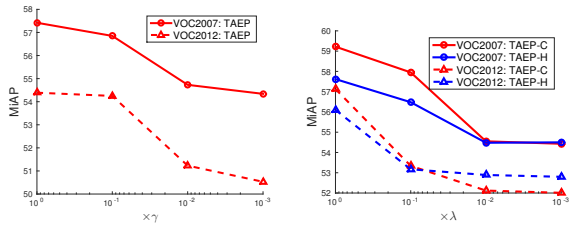

Figure 2: Impacts of the $\mathcal{H}(U)$ term and the auxiliary information. Note x-axis shows relative scaling factors on γ or λ. By gradually diminishing the regularization term (via γ, on the left) or the auxiliary information (via λ, on the right), the performance drops.

results verified the efficacy of the proposed model. They also demonstrated the usefulness of auxiliary information and validated the effective information integration mechanism of our proposed model.

Generalized multi-label zero-shot learning results. Although zero-shot learning has often been evaluated only on the unseen classes in the literature, it is natural to evaluate multi-label zero-shot learning on all the classes, which is referred to as generalized multi-label zero-shot learning. Hence we conducted experiments to test the generalized zero-shot classification performance of the comparison methods. Each method is still trained on the same seen classes $\mathcal{S}$, but the test set now contains all the seen and unseen labels, i.e., $\mathcal{S} \cup \mathcal{U}$. The average comparison results on the two datasets are reported in Table 2. We can see that the two specialized multi-label zero-shot learning methods, DMP and Fast0Tag, outperform the adapted methods ConSE and LatEm-M in terms of most measures on both VOC2007 and VOC2012, while TAEP achieves competitive performances with them. By further incorporating the auxiliary information, the proposed methods, TAEP-C and TAEP-H, not only consistently outperform all the three comparison methods on both datasets in terms of all the evaluation metrics, they also consistently outperform the base model TAEP. TAEP-C again produced the best results in most cases. These results suggest our proposed model provides an effective framework on learning transfer-aware label embeddings for generalized multi-label zero-shot learning, and it also provides the effective mechanism on incorporating free auxiliary information.

4.3 Impact of Label Embedding Regularization

In this section we study the impact of label embedding projection regularization term $\mathcal{H}(U)$, i.e., the transfer-aware part of the proposed model. For TAEP, we firstly set the parameters to the same values, γ_0, as those that generate Table 1, and then reduce γ by a factor of 10 each time to repeat the experiments. That is, we try γ=$\gamma_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$. Since γ is the weight for the regularization term $\mathcal{H}(U)$, by doing this we are actually reducing the contribution of the embedding projection regularization term. The results in terms of MiAP are presented in Figure 2. Similarly, we also tested the impact of auxiliary information through the regularization term $\mathcal{H}(U)$ for TAEP-H and TAEP-C by reducing λ by factors of $\{10^0, 10^{-1}, 10^{-2}, 10^{-3}\}$. From Figure 2 we can see that, as γ decreases, the performance of TAEP decreases on both datasets. This suggests that the label embedding projection regularization term $\mathcal{H}(U)$ is a necessary and useful component. By regularizing the label embeddings to induce better inter-label relationships, the cross-class information transfer can be facilitated in zero-shot learning. Similarly, we also observe that when λ decreases, the performance of TAEP-C and TAEP-H decreases as well on both datasets. This again verifies the usefulness of auxiliary information and the effectiveness of auxiliary integration mechanism of the proposed transfer-aware embedding projection method.

5 Conclusion

In this paper we proposed a transfer-aware label embedding approach for multi-label zero-shot image classification. This approach projects both images and labels into the same semantic space to rank the similarity scores of the images with positive and negative labels under a max-margin learning framework, while guiding the label embedding projection with a transfer-aware regularization objective to achieve a suitable inter-label relations for information adaptation. The regularization framework also allows convenient incorporations of auxiliary information. We conducted experiments to compare our approach with a few related ZSL methods on multi-label image classification tasks. The results demonstrated the efficacy of the proposed approach.

Table 2: Average comparison results (%) on generalized multi-label zero-shot Learning. Smaller values indicate better results in terms of Hamming loss, while larger values indicate better results in terms of the remaining measures.

Methods	VOC2007				VOC2012			
	MiAP	micro-F1	macro-F1	Hamm.	MiAP	micro-F1	macro-F1	Hamm.
ConSE	64.10	42.11	32.29	12.78	62.85	41.17	35.72	13.04
LatEm-M	66.46	43.11	32.37	12.56	63.06	39.95	32.35	13.31
DMP	67.79	43.97	34.13	12.37	64.24	41.29	32.39	13.02
Fast0Tag	67.34	43.54	33.31	12.49	64.63	41.28	32.46	12.97
TAEP	68.16	43.61	35.29	12.01	64.67	40.60	34.07	12.75
TAEP-C	**69.87**	**44.75**	**35.62**	**11.98**	**65.33**	**42.10**	**36.74**	**12.53**
TAEP-H	69.74	44.55	35.56	12.00	65.10	41.39	35.95	12.94

References

[Akata *et al.*, 2015] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015.

[Al-Halah *et al.*, 2016] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen. Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning. In *CVPR*, 2016.

[Bucak *et al.*, 2011] S. Bucak, J. Rong, and A. Jain. Multi-label learning with incomplete class assignments. In *Proc. of CVPR*, 2011.

[Changpinyo *et al.*, 2017] S. Changpinyo, W.-L. Chao, and F. Sha. Predicting visual exemplars of unseen classes for zero-shot learning. In *ICCV*, 2017.

[Elisseeff *et al.*, 2001] A. Elisseeff, J. Weston, et al. A kernel method for multi-labelled classification. In *NIPS*, 2001.

[Frome *et al.*, 2013] A. Frome, G. S Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013.

[Fu *et al.*, 2014] Y. Fu, Y. Yang, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-label zero-shot learning. In *BMVC*, 2014.

[Fu *et al.*, 2015] Y. Fu, T. M Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *TPAMI*, 37(11):2332–2345, 2015.

[Gaure *et al.*, 2017] A. Gaure, A. Gupta, V. Kumar Verma, and P. Rai. A probabilistic framework for zero-shot multi-label learning. In *UAI*, 2017.

[Guo and Schuurmans, 2011] Y. Guo and D. Schuurmans. Adaptive large margin training for multilabel classification. In *AAAI*, 2011.

[Lampert *et al.*, 2009] C. H Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009.

[Lampert *et al.*, 2014] C. H Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *TPAMI*, 36(3):453–465, 2014.

[Lee *et al.*, 2017] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. Frank Wang. Multi-label zero-shot learning with structured knowledge graphs. *arXiv preprint arXiv:1711.06526*, 2017.

[Li *et al.*, 2015a] X. Li, Y. Guo, and D. Schuurmans. Semi-supervised zero-shot classification with label representation learning. In *ICCV*, 2015.

[Li *et al.*, 2015b] X. Li, S. Liao, W. Lan, X. Du, and G. Yang. Zero-shot image tagging by hierarchical semantic embedding. In *SIGIR*. ACM, 2015.

[Li *et al.*, 2016] X. Li, T. Uricchio, L. Ballan, M. Bertini, C. GM Snoek, and A. Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)*, 49(1):14, 2016.

[Mensink *et al.*, 2014] T. Mensink, E. Gavves, and C. GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *CVPR*, 2014.

[Mikolov *et al.*, 2013] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[Miller, 1995] G. A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[Norouzi *et al.*, 2013] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[Pennington *et al.*, 2014] J. Pennington, R. Socher, and C. D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014.

[Qiao *et al.*, 2016] R. Qiao, L. Liu, C. Shen, and A. van den Hengel. Less is more: zero-shot learning from online textual documents with noise suppression. In *CVPR*, 2016.

[Rohrbach *et al.*, 2010] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010.

[Rohrbach *et al.*, 2011] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011.

[Romera-Paredes and Torr, 2015] B. Romera-Paredes and P. HS Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015.

[Simonyan and Zisserman, 2014] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[Weston *et al.*, 2011] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[Xian *et al.*, 2016] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele. Latent embeddings for zero-shot classification. In *CVPR*, 2016.

[Yang *et al.*, 2016] H. Yang, Joey T. Zhou, and J. Cai. Improving multi-label learning with missing labels by structured semantic correlations. In *Proc. of ECCV*, 2016.

[Yu *et al.*, 2014] H. Yu, P. Jain, P. Kar, and I. Dhillon. Large-scale multi-label learning with missing labels. In *Proc. of ICML*, 2014.

[Zhang and Saligrama, 2015] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015.

[Zhang *et al.*, 2016] Y. Zhang, B. Gong, and M. Shah. Fast zero-shot image tagging. In *CVPR*, 2016.