MULTI-LABEL ZERO-SHOT LEARNING WITH TRANSFER-AWARE LABEL EMBEDDING PROJECTION

Meng Ye* Yuhong Guo†
*Temple University, USA †Carleton University, Canada

ABSTRACT
Zero-Shot learning (ZSL) recently has drawn a lot of attention due to its ability to transfer knowledge from seen classes to novel unseen classes, which greatly reduces human labor of labelling data for building new classifiers. Much effort on ZSL 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 both standard zero-shot multi-label image classification and generalized zero-shot multi-label classification. The results demonstrate the efficacy of the proposed approach.

Index Terms— Zero-shot learning, multi-label classification, label embedding, auxiliary information

1. INTRODUCTION
Image categorization has been one essential task in computer vision analysis. Currently the most effective approaches for this task are based on deep convolutional neural networks (CNNs)[1, 2, 3, 4]. However, these supervised approaches require a large number of labelled images for each single class to perform training, and hence induce substantial annotation costs. Therefore it is very important to develop algorithms that enable the reduction of annotation cost for training classification models. Zero-shot learning (ZSL) transfers knowledge from annotated seen classes to predict unseen classes that have no labeled data, hence has received a lot of attention[5, 6, 7, 8, 9].

One primary source deployed in ZSL for bridging the gap between seen and unseen classes is the attribute description of the class labels [5, 10, 7, 11]. 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. Some other works, instead of using human-defined attributes, exploit more easily accessible free information sources from the Internet, including textual descriptions from Wikipedia articles [12, 6], word embedding vectors [13, 14] trained from large text corpus using natural language processing (NLP) techniques [6, 15, 16, 17, 8, 18], co-occurrence statistics of hit-counts from search engine [19, 20], and WordNet hierarchy information of the labels [19, 21, 22]. These works demonstrated impressive results on several standard zero-shot datasets. However, majority research effort has concentrated on multi-class ZSL, while the more challenging multi-label ZSL problem has received very limited attention [20, 23, 24, 25].

In this work we propose a novel transfer-aware label embedding projection method to tackle multi-label ZSL. Instead of using fixed pre-defined word vectors or attribute vectors as label embedding space, we propose to project them into a low-dimensional semantic space in a transfer-aware manner to enforce similarity between seen and unseen class la-
bels and separability across unseen labels. We then simultaneously co-project the seen class instances into the same semantic space under a max-margin multi-label classification framework. Moreover, we incorporate auxiliary information into the transfer regularization term in a convenient manner to guide the label embedding projection and gain a better inter-label relationship structure for multi-label ZSL knowledge transfer. With the learned co-projection of features and label embedding vectors, one can easily predict multiple labels from unseen classes for any new instance. Experimental results on two datasets under both standard ZSL and generalized ZSL setting demonstrate the effectiveness of the proposed approach.

2. PROPOSED APPROACH

2.1. Problem Definition and Notations

We consider multi-label ZSL in the following setting. Assume we have a set of $n$ labeled training images $D = \{X, Y\}$, where $X \in \mathbb{R}^{n \times d}$ is the $d$-dim feature matrix for the $n$ images, and $Y \in \{0, 1\}^{n \times L'}$ denotes the corresponding label indicator matrix across a set of seen classes, $\mathcal{S} = \{1, 2, ..., L_s\}$: 1/0 indicates the presence/absence of the corresponding label. 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. The goal is to learn a multi-label model from the training data that can perform prediction on the unseen classes. We use $X_i$ to denote the $i$-th row vector of matrix $X$ and $Y_i$ to denote the complement of $Y_i$ such that $Y_i = 1 - Y_i$. We also reuse $Y_i$ to denote a set of indices of its non-zero values within proper contexts. We use $\mathbf{I}$ to denote a column vector of all 1s, whose size can be determined in the context, and use $1_1$ to denote an identity matrix. and $O_{a,b}$ and $1_{a,b}$ to denote $a \times b$ matrix with all 0s and all 1s.

2.2. Max-margin Multi-label Learning with Semantic Embedding Co-projection

Instead of entirely relying on the pre-given label embedding vectors $M$, we propose to project them and the image features into a more suitable common low-dimensional semantic space. Specifically, we want to learn a projection function $\theta : \mathbb{R}^d \rightarrow \mathbb{R}^r$ and a function $\phi : \mathbb{R}^m \rightarrow \mathbb{R}^r$ that maps an instance $X_i$ and a label embedding vector $M_c$ into a semantic space $\mathbb{R}^r$; in linear projection form we have $\theta(X_i) = X_iW, W \in \mathbb{R}^{d \times r}$ and $\phi(M_c) = M_c U, U \in \mathbb{R}^{m \times r}$, respectively. 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 projected representations in the common semantic space:

$$F(i,c) = \theta(X_i)\phi(M_{c})^\top = X_iW U^\top M_{c}^\top$$

(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 greater than the similarity score $F(i,\bar{c})$ between instance $X_i$ and any of its negative label $\bar{c} \in \bar{Y}_i$, we formulate the projection learning problem within a max-margin multi-label learning framework:

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

(2)

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

$$\mathcal{L}(W;U;X_i, Y_i) = \max_{c \in Y_i} [1 + F(i,0) - F(i,c)]_+ + \max_{\bar{c} \in \bar{Y}_i} [1 + F(i,\bar{c}) - F(i,0)]_+$$

(3)

where $[\cdot]_+ = \max(\cdot, 0)$. $F(i,0) = X_iW_0$ 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., $F(i,c) > F(i,0) > F(i,\bar{c})$, to minimize the loss. We assume the project matrix $U$ has orthogonal columns to maintain a succinct label embedding projection. For the regularization over $W$, we consider a Frobenius norm regularizer, $\mathcal{R}(W) = \frac{\beta}{2} (\|W\|_F^2 + \|W_0\|_F^2)$, $\beta$ is a trade-off parameter.

2.3. Transfer-Aware Label Embedding Projection

The ranking loss in Eq.(3) minimizes classification error on seen classes. However the goal of ZSL is to predict labels from the unseen classes. Our intuition is that if unseen class embedding vectors are close to the seen class embedding vectors and, at the same time, well separated from each other in the projected space, knowledge transfer would be much easier. We hence propose a transfer-aware regularization objective $\mathcal{H}(U)$ such that:

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

$$= \frac{\gamma}{2} \text{tr}(U^\top M^\top Q MU)$$

where

$$Q = \begin{bmatrix}
0_{L^s \times L^s} & -\frac{1}{2L^s-1} 1_{L^s \times L^s} \\
-\frac{1}{2L^s-1} 1_{L^s \times L^s} & \frac{1}{2L^s-1} (1_{L^s \times L^s} - I_{L^s})
\end{bmatrix}$$

(4)
The first term in $H(U)$ encourages the separability among unseen label embedding vectors while the second term encourages their similarity to seen embedding vectors. Also note that this regularization ignores the separability for the seen classes. This is because seen classes can be well separated through classification loss over the labeled instances. Further encoding such information in this regularization objective may cause overfitting on seen classes and negatively impact zero-shot transfer to the unseen ones.

2.4. Integration of Auxiliary Information

In addition to Eq.(4), we propose to leverage auxiliary information, in a similar way, to further improve label embedding projection.

We 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 $Q^A = I - D^{-1/2}RD^{-1/2}$, where $D = \text{diag}(R1)$, is the normalized Laplacian matrix of $R$. We use a manifold regularization term to enforce the projected label embeddings to be better aligned with the inter-class affinity $R$:

$$A(U) = \frac{\lambda}{2} \text{tr} \left( U^\top M^\top Q^A U \right)$$  (5)

where $\lambda$ is a balance parameter for $A(\cdot)$. This regularization form has two advantages. First, it can be conveniently integrated into the framework by simply updating $H(U)$ to:

$$H(U) = \frac{\gamma}{2} \text{tr} \left( U^\top M^\top (Q + \frac{\lambda}{\gamma} Q^A) U \right)$$  (6)

Second, it is convenient to exploit different auxiliary resources by simply replacing $R$ (or $Q^A$). In this work we study two types of auxiliary information resources, WordNet [26] hierarchy and web co-occurrence statistics.

WordNet: WordNet [26] is a large lexical database of English. Words are grouped into a hierarchical tree structure based on their semantic meanings. We use the shortest path between any two words to define the similarity between two labels $i$ and $j$, such that $R_{ij} = \frac{1}{\text{path}(\text{w}(i), \text{w}(j)) + 1}$.

Co-occurrence statistics: 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. Following previous works [19, 20], we use the Flickr Image Hit Count to compute the dice-coefficient as similarity between two labels, i.e., $R_{ij} = \frac{HC(i, j)}{HC(i) + HC(j)}$.

2.5. TAEP for Multi-label ZSL

By incorporating $H(U)$ into the framework in Eq.(2), we obtain the following Transfer-Aware max-margin Embedding Projection (TAEP) learning problem:

$$\min_{W,W_0,\xi,\eta, U: U^\top U = I} \mathbf{1}^\top \xi + \mathbf{1}^\top \eta + \frac{\beta}{2} (||W||_F^2 + ||W_0||_F^2) + H(U)$$  (7)

s.t. $F(i, c) - F(i, 0) \geq 1 - \xi_i, \forall c \in Y_i, \forall i; \xi_i \geq 0$; $F(i, 0) - F(i, \tilde{c}) \geq 1 - \eta_i, \forall \tilde{c} \in \bar{Y}_i, \forall i; \eta_i \geq 0$

The objective learns $W$ and $U$ by enforcing positive labels to rank higher than negative labels, while incorporating the regularization term $H(U)$ to refine the label embedding structure in the semantic space. To overcome the difficulties of learning Eq.(7) with various constraints, we re-formulate it into a min-max optimization problem by deploying Lagrangian dual formulation. Then we develop an iterative alternating optimization algorithm to perform training. Detailed formulations can be found in supplementary material, we omit them for clarity of the main manuscript.

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 $c \in U$.

3. EXPERIMENTS

In this section we conduct experiments to investigate the empirical performance of the proposed method.

3.1. 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.

Implementation details We used VGG19 [2] pre-trained on ImageNet to extract the 4096-dim visual features and used the 300-dim word vectors pre-trained by GloVe [14] as label embedding vectors. Both of them are $l_2$ normalized. For parameter selection, we split the seen classes set into two subsets of classes: training and validation. Then we train the model on the training set and choose hyper-parameters, $\beta \in \{1, 2, ..., 10\}$ and $\gamma, \lambda \in \{0.01, 0.1, 1, 10\}$ based on the performances on the validation set. 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 three different multi-label evaluation metrics: MiAP, micro-F1 (mi-F1) and macro-F1 (ma-F1). The Mean image Average Precision (MiAP) [27] measures how well are the labels ranked on a given image, while the other two measure how well the predicted labels match with the ground truth labels.
3.2. Multi-label Zero-shot Learning Results

Comparison methods. We compared the proposed method with four related multi-label ZSL methods, ConSE [15], LatEm-M [17], DMP [23] and Fast0Tag [24]. ConSE and LatEm are the multi-label adaptations of two standard ZSL approaches: ConSE can be adapted to multi-label setting straightforwardly; 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). DMP and Fast0Tag are specifically developed for multi-label ZSL. For our proposed transfer-aware max-margin embedding projection (TAEP) method, we also provide two variants with different types of auxiliary information: TAEP-H uses WordNet hierarchy, and TAEP-C uses Flickr Image Hit-Count.

Standard multi-label ZSL results. We first conducted experiments under the standard ZSL setting. We divided the datasets into two subsets of equal number of classes, as seen and unseen classes respectively. All methods train their models on seen instances and make predictions on the unseen class instances. We selected the hyper-parameters for the comparison methods based on grid search. Average performance of 5 runs of each approach are reported in Table 1. We can see the specialized multi-label ZSL methods, DMP and Fast0Tag, consistently outperform ConSE. The proposed TAEP outperforms all the four comparison methods with notable improvements on both datasets. By integrating auxiliary information, TAEP-C and TAEP-H further improve the performance of TAEP. These 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 approach.

Generalized multi-label ZSL results. A more challenging task scenario is to evaluate a model on all the classes, which is referred to as generalized multi-label ZSL. Under this scenario, each method is still trained on the seen classes \( S \) but tested on all the seen and unseen labels, i.e., \( S \cup U \). The average comparison results on the two datasets are reported in Table 2. We can see that TAEP achieved competitive performance with DMP and Fast0Tag. By further incorporating the auxiliary information, TAEP-C and TAEP-H consistently outperform the base model TAEP and all the three comparison methods on both datasets. These results again suggest the effectiveness of our proposed framework.

3.3. 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 parameter to the original value \( \gamma \) that generate Table 1, and then reduce it by a factor of 10, such that \( \gamma = \gamma_0 \times \{10^1, 10^{-1}, 10^{-2}, 10^{-3}\} \). For TAEP-H and TAEP-C we change the value of \( \lambda \) so that \( \lambda = \lambda_0 \times \{10^0, 10^{-1}, 10^{-2}, 10^{-3}\} \). 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 (left for TAEP, right for TAEP-H and TAEP-C). We can see that, as \( \gamma \) decreases, the performance of TAEP also decreases on both datasets. Similarly, when \( \lambda \) decreases, the performance of TAEP-C and TAEP-H decreases as well on both datasets. This suggests that the label embedding projection regularization term \( \mathcal{H}(U) \) is a necessary and useful component to facilitate the cross-class information in ZSL.

4. CONCLUSION

In this paper, we proposed a transfer-aware label embedding approach for multi-label ZSL problem. The approach uses transfer-aware regularization objective for effective information adaptation and allows convenient incorporation of auxiliary information for further improvements.
5. REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E Hinton, “Imagenet classification with deep convolutional neural networks,” in NIPS, 2012.

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

[3] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, “Deep convolutional ranking for multilabel image annotation,” arXiv preprint arXiv:1312.4894, 2013.

[4] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, “CNN-RNN: A unified framework for multilabel image classification,” in CVPR, 2016.

[5] C. H. Lampert, H. Nickisch, and S. Harmeling, “Learning to detect unseen object classes by between-class attribute transfer,” in CVPR, 2009.

[6] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele, “Evaluation of output embeddings for fine-grained image classification,” in CVPR, 2015.

[7] B. Romera-Paredes and P. HS Torr, “An embarrassingly simple approach to zero-shot learning,” in ICML, 2015.

[8] Z. Zhang and V. Saligrama, “Zero-shot learning via semantic similarity embedding,” in CVPR, 2015.

[9] S. Changpinyo, W.-L. Chao, and F. Sha, “Predicting visual exemplars of unseen classes for zero-shot learning,” in ICCV, 2017.

[10] C. H. Lampert, H. Nickisch, and S. Harmeling, “Attribute-based classification for zero-shot visual object categorization,” TPAMI, vol. 36, no. 3, pp. 453–465, 2014.

[11] Y. Fu, T. M Hospedales, T. Xiang, and S. Gong, “Transductive multi-view zero-shot learning,” TPAMI, vol. 37, no. 11, pp. 2332–2345, 2015.

[12] 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.

[13] T. Mikolov, I. Sutskever, K. Chen, G. S Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” in NIPS, 2013.

[14] J. Pennington, R. Socher, and C. D Manning, “Glove: Global vectors for word representation.,” in EMNLP, 2014.

[15] 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.

[16] A. Frome, G. S Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al., “Derive: A deep visual-semantic embedding model,” in NIPS, 2013.

[17] Y. Xian, Z. Akata, G. Sharma, Q. Nguyen, M. Hein, and B. Schiele, “Latent embeddings for zero-shot classification,” in CVPR, 2016.

[18] Z. Al-Halah, M. Tapaswi, and R. Stiefelhagen, “Recovering the missing link: Predicting class-attribute associations for unsupervised zero-shot learning,” in CVPR, 2016.

[19] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele, “What helps where— and why? semantic relatedness for knowledge transfer,” in CVPR, 2010.

[20] T. Mensink, E. Gavves, and C. GM Snoek, “Costa: Co-occurrence statistics for zero-shot classification,” in CVPR, 2014.

[21] M. Rohrbach, M. Stark, and B. Schiele, “Evaluating knowledge transfer and zero-shot learning in a large-scale setting,” in CVPR, 2011.

[22] X. Li, S. Liao, W. Lan, X. Du, and G. Yang, “Zero-shot image tagging by hierarchical semantic embedding,” in SIGIR. ACM, 2015.

[23] Y. Fu, Y. Yang, T. M. Hospedales, T. Xiang, and S. Gong, “Transductive multi-label zero-shot learning,” in BMVC, 2014.

[24] Y. Zhang, B. Gong, and M. Shah, “Fast zero-shot image tagging,” in CVPR, 2016.

[25] 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.

[26] G. A Miller, “Wordnet: a lexical database for english,” Communications of the ACM, vol. 38, no. 11, pp. 39–41, 1995.

[27] 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), vol. 49, no. 1, pp. 14, 2016.