Adaptive Graph Guided Embedding for Multi-label Annotation

Lichen Wang[†], **Zhengming Ding**[†], **Yun Fu**[†‡]
[†] Department of Electrical & Computer Engineering, Northeastern University, Boston, USA
[‡] College of Computer & Information Science, Northeastern University, Boston, USA
wanglichenxj@gmail.com, allanding@ece.neu.edu, yunfu@ece.neu.edu

Abstract

Multi-label annotation is challenging since a large amount of well-labeled training data are required to achieve promising performance. However, providing such data is expensive while unlabeled data are widely available. To this end, we propose a novel Adaptive Graph Guided Embedding (AG^{2}E) approach for multi-label annotation in a semi-supervised fashion, which utilizes limited labeled data associating with large-scale unlabeled data to facilitate learning performance. Specifically, a multi-label propagation scheme and an effective embedding are jointly learned to seek a latent space where unlabeled instances tend to be well assigned multiple labels. Furthermore, a locality structure regularizer is designed to preserve the intrinsic structure and enhance the multi-label annotation. We evaluate our model in both conventional multi-label learning and zero-shot learning scenario. Experimental results demonstrate that our approach outperforms other compared state-of-the-art methods.

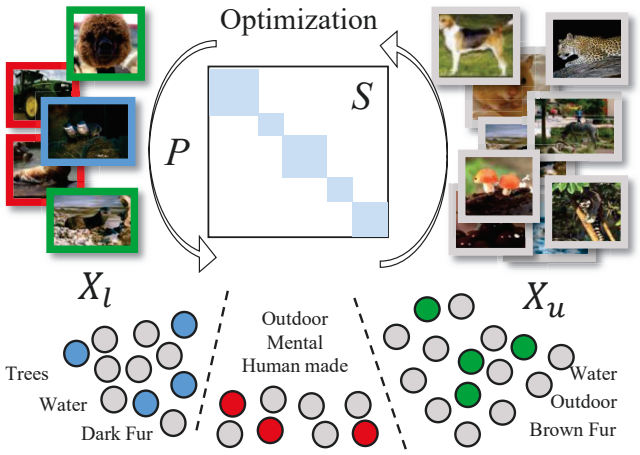

Figure 1: Framework of our model, where an adaptive affinity graph S connects pair-wise relations across labeled and unlabeled samples. A pre-defined graph fully provides local structure information and accelerates optimization process. A projection P projects data into a common and distinctive space which also eliminates interruptions from noise and outliers. P, S, and label matrix F are simultaneously updated to achieve accurate and robust performance.

1 Introduction

In the real-world scenarios, each individual object could contain tens or hundreds of semantic descriptions, such as colors, materials and shapes. Different from single-label learning, multi-label learning assigns multiple labels for each sample [Liu *et al.*, 2017], which is much more challenging compared with single-label scenario. First, the relevant datasets [Lampert *et al.*, 2009; Patterson and Hays, 2012; Wah *et al.*, 2011] are small due to high labeling cost. Second, the labels follow a long-tailed distribution, that means some labels show up more frequently than others. The situation makes label recovery dominated by the major labels, and it might hurt the label recovery performance. Third, labels such as stressful, cold and warm are subjectively assigned, and different people hold different standards, and thus the noise and outliers significantly appear in datasets compared with single label scenario.

Although the number of well-labeled data is limited, related unlabeled data are widely accessible. Thus, it is practical to utilize unlabeled samples to improve the learning performance. Consequently, semi-supervised learning [Zhu *et al.*, 2003] especially graph-based approach [Zha *et al.*, 2009] has attracted great attention. However, they still have limitations that these methods highly depend on the pre-constructed graph but rarely optimize it online. In addition, most methods construct graph directly on feature space, which is sensitive to noise and outliers. Several work utilize adaptive graphs to handle the problem [Nie *et al.*, 2012; 2016]. However, these methods mainly focus on single-label classification instead of multi-label scenario.

To this end, we propose a novel Adaptive Graph Guided Embedding (AG^{2}E) for multi-label learning in semi-supervised fashion. Figure 1 shows the framework of AG^{2}E, whose core idea is learning a semi-supervised label propagation and an effective embedding simultaneously to seek a latent space, and thus unlabeled images can be well recovered. Our main contributions are summarized as follows:

- We seek an adaptive graph to automatically capture the latent structure of the data. Moreover, a pre-defined

locality-constrained graph is also utilized to preserve the intrinsic structure and guide the adaptive graph learning.

- A linear projection is jointly learned to obtain more effective and distinctive feature representations for better label propagation. It enhances the accuracy and robustness of our approach.
- Non-trivially, we propose an efficient optimization strategy to solve the model. Experimental results on five benchmarks demonstrate the effectiveness and the efficiency of our model.

2 Related Works

Related work including Multi-label learning and Semi-supervised learning are introduced in this section.

Multi-label learning learns patterns, which compose instances associating with multiple labels. It widely exists in real-world applications, such as visual annotation [Boutell *et al.*, 2004] and image retrieval [Liu *et al.*, 2018]. It is challenging since the possible label sets number is tremendous. The intuitive solution is to consider the task as several single-label problems. [Boutell *et al.*, 2004] learns several classifiers responding for each label. However, since latent connections exist between labels, ignoring the connections would limit the learning performances. [Godbole and Sarawagi, 2004] proposes to leverage the correlations across labels by adding a contextual fusion step. [Liu and Tsang, 2015] explores metric learning paradigm to improve accuracy. However, most methods are in supervised learning scenario. Obtaining sufficient labeled data to achieve acceptable performance is costing which limits their practical applications.

Semi-supervised learning achieves well-trained models by using a few labeled data as well as a large number of unlabeled data. A comprehensive survey can be found in [Zhu, 2005]. Graph-based methods achieve high performance by constructing an affiliate graph to recover labels. [Zhu *et al.*, 2003] proposes Gaussian random fields and harmonic function to obtain semi-supervised learning. [Sindhwani and Belkin, 2005] gives a semi-supervised kernel that is not limited to unlabeled points, but defined over all input spaces. [Nie *et al.*, 2012] actively selects training sets to make the model be independent to initialization process. [Wang *et al.*, 2018] aims to transfer well-label source video information to boost clustering performance on unlabeled target domain. However, these methods are highly depended on the pre-defined affiliate graph. It is difficult and tedious to tune the graph to an optimized structure. Moreover, real-world datasets always contain noise and outliers, which could impair the final performance. [Liu *et al.*, 2006] designs a new scheme to generate an adaptive similarity graph. [Nie *et al.*, 2016] proposes a graph optimization strategy on unsupervised feature selection scenario. [Nie *et al.*, 2017] designs an optimal graph in clustering and classification settings. However, the graphs are optimized in unsupervised manner, which is hard to involve supervision information to enhance the learning performance. Moreover, it is still purely based on the similarity measurement in feature space, which is easily affected by noise and outliers.

Different from previous work, we deploy an adaptive graph for semi-supervised multi-label learning. Specifically, adaptive graph is more flexible to capture the intrinsic data structure and accurately predict the labels. Meanwhile, an effective embedding is jointly learned to align the different distribution data in a low-dimensional but distinctive common space.

3 The Proposed Approach

3.1 Notations

Assume we have labeled data X_l and Y_l. $X_l \in \mathbb{R}^{d\times n_l}$ is the labeled feature matrix, where each column x_i is an instance, n_l is the sample number, d is the feature dimension. $Y_l \in \mathbb{R}^{d_l\times n_l}$ is the label matrix, where d_l denotes the label dimension. $X_u \in \mathbb{R}^{d\times n_u}$ represents unlabeled feature matrix, where n_u is the instance number. Our approach aims to utilize X_l, Y_l, and X_u to jointly seek an effective transformation for better feature extraction and recover the multiple labels of X_u.

3.2 AG^{2}E Approach

AG^{2}E approach enlarges the multi-label distribution by utilizing the labeled data to the unlabeled data through an adaptive graph. Previous methods usually achieve label propagation based on pre-defined adjacency matrix [Guo *et al.*, 2016]. This strategy assumes samples which are close in feature space shall share the similar labels. The objective function is shown below:

$$\min_{F} \quad \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij}, \text{ s.t. } F_l = Y_l, \tag{1}$$

where f_i and f_j are the corresponding labels of i-th and j-th instances. $F = [F_l, F_u]$, where $F_l \in \mathbb{R}^{d_l\times n_l}$, $F_u \in \mathbb{R}^{d_l\times n_u}$ are the recovered label matrices of X_l and X_u. We set $F_l = Y_l$ since F_l is expected to be the same as the ground truth Y_l. $S \in \mathbb{R}^{n\times n}$ is the similarity matrix, where $n = n_l + n_u$. Each entry s_{ij} is the similarity metric between feature points x_i and x_j. Several methods are proposed to obtain S such as [Ng *et al.*, 2001; Ding and Fu, 2014; Wang *et al.*, 2018], with the definition as follows:

$$s_{ij} = \begin{cases} e^{-\|x_i - x_j\|_2^2/2\delta^2}, & \text{if } x_i \in \mathcal{N}_K(x_j) \\ & \text{or } x_j \in \mathcal{N}_K(x_i), \\ 0, & \text{otherwise,} \end{cases} \tag{2}$$

where $\mathcal{N}_K(x_i)$ denotes the K-nearest neighbors of x_i. And $1 \le (i,j) \le n$. The quality of S affects learning performance significantly. If S is obtained directly in feature space, it is challenging for S to reveal the intrinsic structure within the data since the noise and outliers are high. To this end, we propose an adaptive graph instead of a fixed graph. Our approach obtains the similarity matrix and simultaneously recovers the labels to achieve the best results. Compared with fixed graph, adaptive graph could be more robust and accurate. We extend Eq. (1) and utilize label correlation information to learn an adaptive graph in learning process. The objective function is shown as below:

$$\begin{aligned} &\min_{F,S} \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij} + \mu \sum_{i,j} \|x_i - x_j\|_2^2 s_{ij}, \\ &\text{s.t. } F_l = Y_l, S \ge 0. \end{aligned} \tag{3}$$

where $\|x_i - x_j\|_2^2 s_{ij}$ constrains the graph optimization that similar features correspond to high similarities and vice versa. μ is a trade-off parameter to balance weights between feature space and label space. F and S are optimized simultaneously.

However, directly learning S in feature space would involve errors due to high-level noise and label outliers. Therefore, we jointly seek a linear projection $P \in \mathbb{R}^{r\times d}$ to project original features into a low-dimensional common space, where r regulates the space dimension [Ding and Fu, 2014]. By this way, the graph quality would be improved, which would help the label recovery performance. The expression is below:

$$\begin{aligned}&\min_{F,P,S} \sum_{i,j} \|f_i - f_j\|_2^2 s_{ij} + \mu \sum_{i,j} \|Px_i - Px_j\|_2^2 s_{ij} \\ &\quad = \text{tr}(FL_SF^\top) + \mu\text{tr}(PXL_SX^\top P^\top), \\ &\text{s.t. } F_l = Y_l, S \geq 0,\end{aligned} \tag{4}$$

where $\text{tr}(\cdot)$ indicates the trace of a matrix. L_S is graph Laplacian matrix [Merris, 1994], where $L_S = D_S - \frac{S+S^\top}{2}$. And $D_S \in \mathbb{R}^{n\times n}$ is a diagonal matrix and each element $D_{S_{ii}} = \sum_j \frac{s_{ij}+s_{ji}}{2}$. μ is the trade-off parameter. Simply learning S cannot obtain clear local structure information of the data. Thus, we propose to further utilize a pre-defined graph $\bar{S}$ associating with a structure regularizer to pull S be close to $\bar{S}$. By this strategy, our model can obtain the detailed locality structure from $\bar{S}$ and still learn an accurate and robust graph S at the same time. Moreover, $\bar{S}$ guides the optimization process which could reduce the computational cost. The objective function is below:

$$\begin{aligned}&\min_{F,P,S} \text{tr}(FL_SF^\top) + \mu\text{tr}(PXL_SX^\top P^\top) + \lambda\|S - \bar{S}\|_{\text{F}}^2, \\ &\text{s.t. } F_l = Y_l, S \geq 0.\end{aligned} \tag{5}$$

Eq. (5) contains a simple solution that only the points of nearest data are assigned as 1 which could eliminate the learning performance. Thus, we further supplement constraints. First, $S\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is a all-one vector. It means that the sum of S entries in each row is 1. Second, $PXHX^\top P^\top = \text{I}$ (I is an identity matrix), and it implies and introduces additional data discriminating ability into our model. Third, $H = \text{I} - \frac{1}{n}\mathbb{I}$ where $\mathbb{I} \in \mathbb{R}^{n\times n}$ is all-one matrix. Then, the objective function can be written as follow:

$$\begin{aligned}&\min_{F,P,S} \text{tr}(FL_SF^\top) + \mu\text{tr}(PXL_SX^\top P^\top) + \lambda\|S - \bar{S}\|_{\text{F}}^2, \\ &\text{s.t. } F_l = Y_l,\ PXHX^\top P^\top = \text{I},\ S\mathbf{1} = \mathbf{1},\ S \geq 0.\end{aligned} \tag{6}$$

3.3 Optimization

Since we have three variables to be optimized in Eq. (6), we adopt the popular method, i.e., Alternative directions method of multipliers (ADMM) [Ding and Fu, 2014; Boyd *et al.*, 2011], to obtain our solution. Specifically, we iteratively optimize each variable by fixing other variables. In the optimization process, other variables are fixed and update one each time until it converges.

Update F: While other variables are fixed, the equation can be rewritten as follows:

$$\min_F \text{tr}(FL_SF^\top), \ \text{ s.t. } \ F_l = Y_l. \tag{7}$$

The differentiate of Eq. (7) is shown below:

$$\begin{aligned}FL_S = 0 \ &\Rightarrow [F_l \quad F_u]\begin{bmatrix}L_{ll} & L_{lu}\\ L_{ul} & L_{uu}\end{bmatrix} \\ &\Rightarrow \begin{cases} F_lL_{ll} + F_uL_{ul} = 0 \\ F_lL_{lu} + F_uL_{uu} = 0.\end{cases}\end{aligned} \tag{8}$$

Then we have $F_u = -F_lL_{lu}L_{uu}^{-1}$. Since we set $F_l = Y_l$, thus F can be updated as $F_u = -Y_lL_{lu}L_{uu}^{-1}$.

Update S: Since S is difficult to optimized directly. We fix other variables and optimize S row by row. The equation can be written as follow:

$$\begin{aligned}&\min_S \text{tr}(FL_SF^\top) + \mu\text{tr}(PXL_SX^\top P^\top) + \lambda\|S - \bar{S}\|_{\text{F}}^2, \\ &\text{s.t. } S\mathbf{1} = \mathbf{1},\ S \geq 0.\end{aligned} \tag{9}$$

We separately discuss the equations. The first term can be written as follow:

$$\sum_i \|f_i - f_j\|_2^2 s_{ij} = \sum_i a_i\mathbf{s}_i^\top, \tag{10}$$

where $a_i = \{a_{ij}, 1 \leq j \leq n\} \in \mathbb{R}^{1\times n}$ with $a_{ij} = \|f_i - f_j\|_2^2$, $\mathbf{s}_i$ is the i-th row of S. We obtain the same format of the second term as follow:

$$\sum_i \mu\|Px_i - Px_j\|_2^2 s_{ij} = \mu \ \sum_i b_i\mathbf{s}_i^\top, \tag{11}$$

where $b_i = \{b_{ij}, 1 \leq j \leq n\} \in \mathbb{R}^{1\times n}$ with $b_{ij} = \|Px_i - Px_j\|_2^2$. For the third term, we can write the format as follow:

$$\begin{aligned}\lambda\|S - \bar{S}\|_{\text{F}}^2 \ &= \lambda\text{tr}((S - \bar{S})(S - \bar{S})^\top) \\ &= \lambda\text{tr}(SS^\top - 2\bar{S}S^\top + \bar{S}\bar{S}^\top).\end{aligned} \tag{12}$$

To minimize $\|S - \bar{S}\|_{\text{F}}^2$, we have the following format:

$$\begin{aligned}&\min_S \lambda\|S - \bar{S}\|_{\text{F}}^2 = \min_S \lambda\text{tr}(SS^\top - 2\bar{S}S^\top), \\ &\text{s.t. } S\mathbf{1} = \mathbf{1},\ S \geq 0.\end{aligned} \tag{13}$$

Following the same strategy, we can convert the term in Eq. (13) into the following format:

$$\lambda\text{tr}(SS^\top - 2\bar{S}S^\top) = \lambda\sum_i \left(\mathbf{s}_i\mathbf{s}_i^\top - 2\lambda\bar{\mathbf{s}}_i\mathbf{s}_i^\top\right). \tag{14}$$

Then we can transform the objective function into the following format:

$$\begin{aligned}&\min_{\mathbf{s}_i} \sum_i (a_i\mathbf{s}_i^\top + \mu b_i\mathbf{s}_i^\top + \lambda(\mathbf{s}_i\mathbf{s}_i^\top - 2\bar{\mathbf{s}}_i\mathbf{s}_i^\top)) \\ &\quad = \sum_i (\lambda\mathbf{s}_i\mathbf{s}_i^\top + (a_i + \mu b_i - 2\lambda\bar{\mathbf{s}}_i)\mathbf{s}_i^\top), \\ &\text{s.t. } \ \mathbf{s}_i\mathbf{1} = 1, \mathbf{s}_i \geq 0.\end{aligned} \tag{15}$$

The optimization problem of Eq. (15) is simple and the accelerated projected gradient approach is utilized to linearly solve the problem. The core process of the projected gradient approach is solving the proximal equation shown below:

$$\min_{z\geq 0} \tfrac{1}{2}\|z - c\|_2^2, \ \text{ s.t. } \ z\mathbf{1} = 1. \tag{16}$$

The KKT approach is used to solve this proximal problem. After each $\mathbf{s}_i$ is solved, we concatenate the result together and obtain the updated graph S.

Update P: While other variables are fixed, we can obtain the equation shown below:

$$\begin{aligned} P &= \underset{P^\top XHX^\top P=\mathrm{I}}{\arg\min} \ \mu \ \mathrm{tr}(PX(I-S)X^\top P^\top) \\ &= \underset{P^\top XHX^\top P=\mathrm{I}}{\arg\min} \ \mathrm{tr}(P[\mu X(I-S)X^\top]P^\top), \end{aligned} \tag{17}$$

and the generalized Eigen-decomposition approach can be used to solve Eq. (17) as shown in Eq. (18).

$$\big(\mu X(I-S)X^\top\big)\rho = \gamma XHX^\top \rho, \tag{18}$$

in which γ is the eigenvalue corresponding to the eigenvector ρ for Eq. (18). Specifically, we could achieve p eigenvectors $\rho_i (i = 0, 1, \cdots, p-1)$, given by the minimum eigenvalue solutions to the generalized Eigen-decomposition problem. Thus, we have $P = [\rho_0, \cdots, \rho_{p-1}]^\top$.

3.4 Discussion

We iteratively optimize all variables until the objective function is convergent. Specifically, F_u is initialized through Eq. (1) and it reduces iteration times and avoids local optimal solution.

There are two time consuming optimization steps. The first is updating F, which uses Bartels Stewart algorithm and the complexity is $\mathbf{O}(n^3)$. The second is updating P and its Eigen-decomposition process costs $\mathbf{O}(d^3)$. These steps can be reduced to $\mathbf{O}(d^{2.37})$ using Coppersmith-Winograd algorithm [Coppersmith and Winograd, 1987]. Then the total computational complexity is $\mathbf{O}(td^{2.37} + tn^{2.37})$ where t is the iteration number. Since the initialization approach reduces the iteration times significantly, thus, our model is applicable to large-scale real-world applications.

4 Experiment

Evaluation datasets, experimental settings, results and discussions are introduced in this section.

4.1 Datasets

Three image, one acoustic and one emotion datasets are evaluated in our experiments. Brief introductions are as follows:

SUN Dataset [Patterson and Hays, 2012] is a large-scale scene multi-label dataset. There are $14,000$ images come from 700 classes. Each instance contains a 102-dimensional label vector with averagely 6.3 labels. The label entries are continuous values range in $[0, 1]$.

CUB Dataset [Wah *et al.*, 2011] contains bird $11,788$ images captured from 200 species. All images are extracted to 312-dimensional attribute labels. Each instance has roughly 28 annotations. The label entries are binary values (0 or 1).

AWA Dataset [Lampert *et al.*, 2014] contains more than $30,000$ images from 50 animals. Each label vector contains 85 continuous values range from 0 to 120 corresponding to 85 semantic attribute labels such as habits, colors, and shapes. Each instance has roughly 15 labels.

BIRD Dataset [Briggs *et al.*, 2013] is an acoustic dataset, which contains 645 ten-second audio recordings from 19 species of bird. Each recording is labeled by several experts along with their confidence.

EMO Dataset [Trohidis *et al.*, 2008] is designed to evaluate automated music emotion detection methods. It collects songs from 233 musical albums and conducted to a set of 30-seconds 593 songs with 6 clusters of music emotions. Sound clips were annotated by experts in music.

4.2 Experimental Setup

For image datasets, we utilize Very Deep Convolution Networks [Simonyan and Zisserman, 2014] to extract $4,096$-dimensional features. For BIRD dataset, we directly use the features provided by [Briggs *et al.*, 2013]. For EMO dataset, we utilize both Rhythmic and Timbre features provided by [Trohidis *et al.*, 2008]. In Multi-label annotation setting, we randomly and evenly split samples into labeled and unlabeled subset. We run our model five times with the randomly generated subsets and report the average performance. 5-fold cross-validation is utilized to select the parameters μ and λ. $r$is empirically set to 120. While since EMO dataset contains 72-dimensional features, we manually set $r = 50$ for EMO dataset. The parameter sensitivity will be discussed in the following sections. Our approach is compared with several state-of-the-art multi-label learning methods, with the brief introductions as follows:

Least Squares Regression (Regression) is a ridge regression method, partial subset of tags labels are utilized to learn a graph and recovery the tags.

Semi-Supervised Multi-Label Dimensionality Reduction (SSMLDR) [Guo *et al.*, 2016] designs a special label propagation method, which transfers the multi-label information across labeled and unlabeled data.

FastTag [Chen *et al.*, 2013] utilizes two linear mappings that are co-regularized in a joint convex loss function. It is able to infer the full list of incomplete tags.

Multi-Label with a Mixed Graph (ML-PGD) [Wu *et al.*, 2015] proposes a uniform approach of label dependencies by generating a graph based on hierarchy structure. This approach simultaneously considers the class co-occurrence as well as the sample-level similarity as non-directive edges. The hierarchy-free version is called ML-PGD.

4.3 Performance Comparison

We utilize the same metrics adopted in [Guillaumin *et al.*, 2009] for fair comparison. In our experiments, we report averaged performance across all instances. To evaluate the result easier, we calculate the F1-score where $F1 = 2\frac{P \times R}{P+R}$ and the non-zero recall number N-R. In all metrics, higher values indicate better performances.

The results shown in Table 1 indicate that the proposed AG^{2}E approach obtains the highest performance in most of the metrics. Our approach leads to a 3.5% improvement on precision, 6% on recall. The results demonstrate the robustness and high accuracy of our approach. Furthermore, our approach is general, which is not limited for image datasets. Our approach fails to achieve significant improvements in AWA dataset, where we consider that AWA dataset labels are continuous value range between 0 to more than 100 with different metrics across class labels. It allows our model to learn an inaccurate affinity graph. While least squares approach

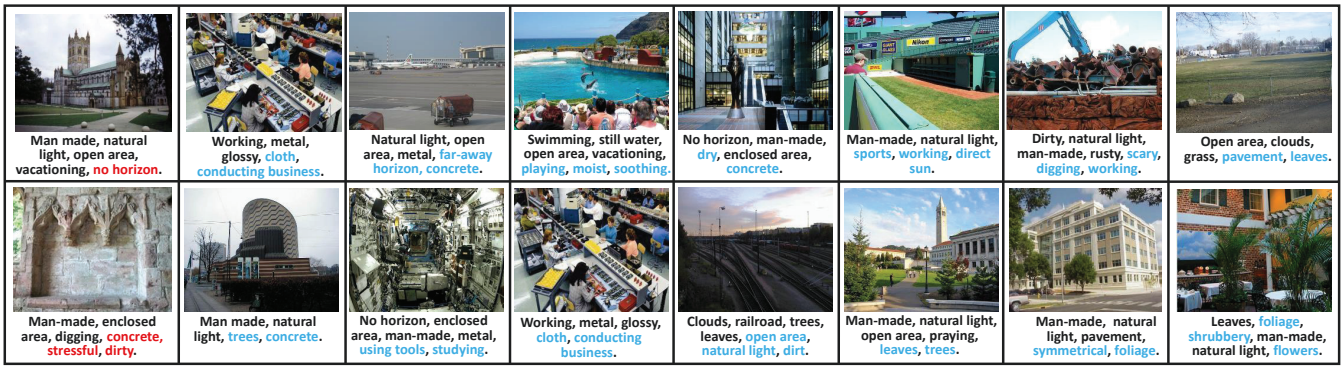

Figure 2: Sample images of recovered labels from SUN dataset. Each image contains several semantic labels. Black fonts denote correct labels. Red fonts denote incorrect labels and blue fonts denote true labels based on our judgments but don't exist in the ground truth labels.

Dataset	Method	Prec	Recall	F1	N-R
SUN	Regression	0.6318	0.1504	0.2429	101
	SSMLDR	0.5625	0.1239	0.2031	68
	FastTag	0.6187	0.1473	0.2379	101
	ML-PGD	0.7218	0.1521	0.2513	100
	AG^{2}E (Ours)	**0.7460**	**0.1625**	**0.2669**	**102**
CUB	Regression	0.2183	0.0247	0.0443	162
	SSMLDR	0.2162	0.0399	0.0674	164
	FastTag	0.3231	0.0496	0.0860	163
	ML-PGD	0.3029	0.0448	0.0781	132
	AG^{2}E	**0.3351**	**0.0525**	**0.0908**	**194**
AWA	Regression	**0.8198**	0.0819	0.1489	**75**
	SSMLDR	0.8085	0.0948	0.1698	74
	FastTag	0.7848	0.0857	0.1545	67
	ML-PGD	0.5283	0.0631	0.1127	45
	AG^{2}E	0.7745	**0.1285**	**0.2204**	72
EMO	Regression	0.3793	0.9114	0.5357	6
	SSMLDR	0.3556	0.8965	0.5093	6
	FastTag	0.3833	0.9459	0.5456	6
	ML-PGD	0.3784	0.9265	0.5373	6
	AG^{2}E	**0.3995**	**0.9714**	**0.5762**	**6**
BIRD	Regression	0.0764	0.3726	0.1268	13
	SSMLDR	0.0709	0.3465	0.1178	12
	FastTag	0.1005	0.3783	0.1601	16
	ML-PGD	0.0809	0.3883	0.1338	15
	AG^{2}E	**0.1021**	**0.4529**	**0.1653**	**17**

Table 1: Comparison of our approach with other methods

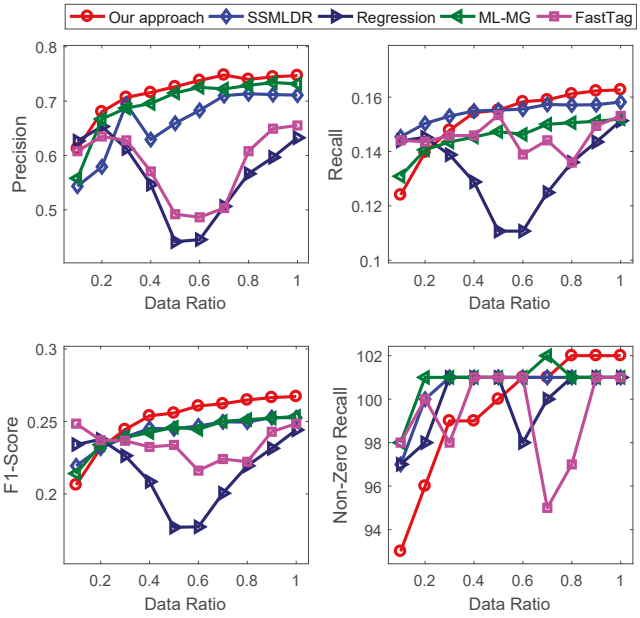

Figure 3: Label recovery performance based on partial of the training set. The results denote that our approach can achieve the best performance if more than 30% of training data are available.

finds a mapping to recover the label values rather than labels, thus it is less affected by this situation.

4.4 Training Data Analysis

To analyze the source data, we randomly remove partial labeled samples and train our model based on 10%, 20% to 100% of the data. Each setting is tested 5 times and we report the average performance. The results are shown in Figure 3, where we notice that our approach cannot achieve the best performance if only 10% to 20% data is available. As the training sample increases, our approach stably improves the accuracy and achieves the best performance. Moreover, our approach still has potential improvements if more labeled data are available. We assume that since our model mainly depends on P and S, and it cannot perform well if any of the two variables are not well trained. Due to this reason, a minimal number of labeled samples are required to achieve accurate performance.

4.5 Image Sample Annotations

Figure 2 shows samples of SUN dataset and their corresponding recovered labels. Considering in some cases there are more than 15 labels from some instances, we only show the labels which have top highest scores in the recovered labels. The red font denotes incorrect labels. And the blue font denotes labels recovered by our approach. These labels don't exist in ground truth, however, they are still reasonable based on our judgments. From the result, we conclude that our approach is effective, which reliably recovers the vast majority of labels in high accuracy. Moreover, our model also recovers missing labels in the original datasets. This property is crucial and useful in real-world applications since missing labels and outliers always happen in existing datasets.

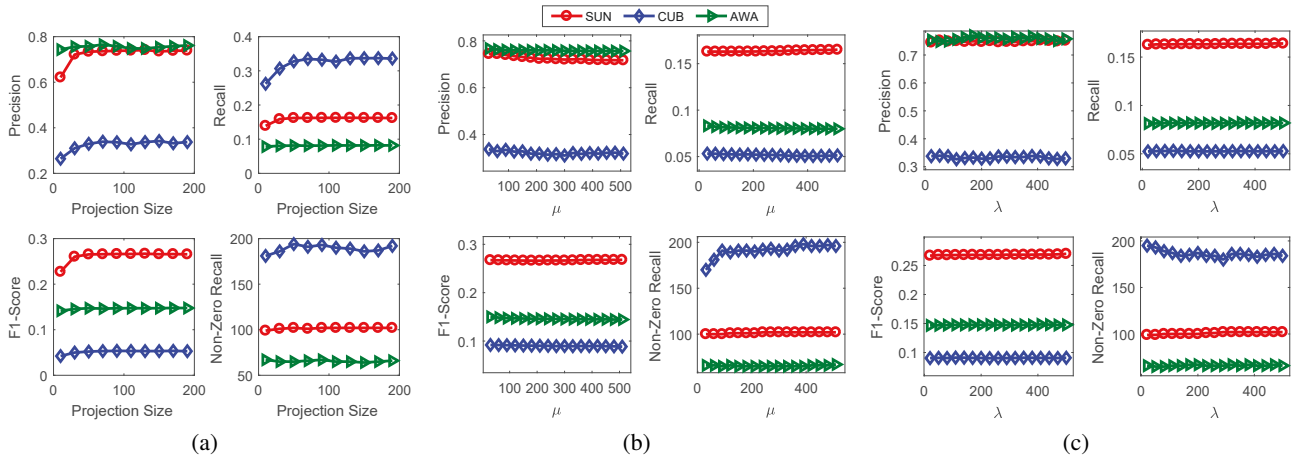

Figure 4: Parameter sensitivity evaluations on SUN dataset. (a) Annotation performance with different projection sizes r. (b) Annotation performance with different μ. (c) Annotation performance with different pre-defined graph constraint parameter λ. From the figures, we can see that our approach is parameter insensitive and the performance is stable when $r > 30$.

Approaches	SUN	CUB	AWA
Labeled data	65.20	27.24	52.31
Regression	65.00	27.21	52.33
SSMLDR	66.00	32.19	53.64
FastTag	64.00	27.18	54.32
ML-PGD	65.40	28.48	54.93
AG^{2}E (Ours)	**67.40**	**32.53**	**55.71**

Table 2: Zero-shot Classification Accuracy (%)

4.6 Zero-shot Learning

We extend our approach to Zero-Shot Learning (ZSL) scenario [Lampert *et al.*, 2009; Antol *et al.*, 2014; Ding *et al.*, 2017]. ZSL tries to recognize classes which do not exist in the training set. To achieve this goal, middle level semantic information is utilized to align unseen classes and visual features. It is a more challenging task due to the larger distribution gap across classes. In our experiments, each sample is assigned for one class label and several attribute labels. In experiments, we split the dataset into three subsets including labeled set, unlabeled set and test set. Test set contains non-overlapped classes compared with other two sets. Our approach recovers the labels of the test data and jointly trains a classifier based on the recovered labels to recognize the classes. We normalize the AWA feature vector y_i based on equation $z_i = y_i / \max(y_i)$ where z_i is the normalized label. In the implementation, we calculate $\bar{S}$ based on class labels. That means $\bar{s}_{ij} = 1$ if x_i and x_j belong to the same class, and otherwise $\bar{s}_{ij} = 0$. KNN is used to classify unseen classes.

The experimental results are shown in Table 2. We observe that the recovered labels can improve the ZSL performance except from linear regression. Since linear regression only recovers labels without considering the existence of missing labels or inner connections. Thus, the performance is as the same as only based on labeled data. Compared with other methods, our method obtains the highest performance in all three datasets. Since the visual feature distributions of seen and unseen classes have larger difference, the result denotes that our AG^{2}E model can obtain more general and compatible feature structures from limited labeled and unlabeled data. It is more accurate and robust than other methods.

4.7 Parameter Analysis

Our approach contains three major parameters, i.e., projection size r, trade-off parameters μ and λ. λ constraints the similarity level between S and $\bar{S}$. μ balances the weight of projected feature space and label space. We adopt different values to evaluate the performance on SUN dataset. The results are shown in Figure 4. We observe that our approach can achieve accurate results when $r \geq 50$, μ is in the range of $[100, 500]$ and λ is in the range of $[100, 500]$. $r \geq 50$ is required since the learned feature space needs to have enough dimension to make the samples distinctive enough to represent diverse samples. From the experimental results, we conclude that the parameter ranges are wide and our approach is robust and parameter insensitive.

5 Conclusion

In this work, we designed an Adaptive Graph Guided Embedding (AG^{2}E) approach in semi-supervised multi-label learning scenario. AG^{2}E utilized limited labeled data associating with unlabeled data to improve multi-label learning performance. In our model, a label propagation and an effective embedding were jointly learned to seek a latent space, where unlabeled information can be fully utilized. Moreover, a locality structure regularizer was explored to preserve intrinsic information and accelerate the optimization procedure. Extensive experimental results demonstrated that our approach was effective and outperformed other methods on several datasets. Our model was robust and parameter insensitive. In the future, more experiments for large-scale image annotation, image retrieval and other applications will be evaluated.

Acknowledgments

This research is supported in part by the NSF IIS award 1651902, ONR Young Investigator Award N00014-14-1-0484, and U.S. Army Research Office Award W911NF-17-1-0367.

References

[Antol *et al.*, 2014] Stanislaw Antol, C Lawrence Zitnick, and Devi Parikh. Zero-shot learning via visual abstraction. In *ECCV*, pages 401–416, 2014.

[Boutell *et al.*, 2004] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *PR*, 37(9):1757–1771, 2004.

[Boyd *et al.*, 2011] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122, 2011.

[Briggs *et al.*, 2013] F Briggs, Huang Yonghong, R Raich, et al. New methods for acoustic classification of multiple simultaneous bird species in a noisy environment. In *MLSP*, pages 1–8, 2013.

[Chen *et al.*, 2013] Minmin Chen, Alice Zheng, et al. Fast image tagging. In *ICML*, pages 1274–1282, 2013.

[Coppersmith and Winograd, 1987] Don Coppersmith and Shmuel Winograd. Matrix multiplication via arithmetic progressions. In *STC*, pages 1–6, 1987.

[Ding and Fu, 2014] Zhengming Ding and Yun Fu. Low-rank common subspace for multi-view learning. In *ICDM*, pages 110–119. IEEE, 2014.

[Ding *et al.*, 2017] Zhengming Ding, Ming Shao, and Yun Fu. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *CVPR*, pages 2050–2058, 2017.

[Godbole and Sarawagi, 2004] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In *PAKDD*, pages 22–30, 2004.

[Guillaumin *et al.*, 2009] Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, et al. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV*, pages 309–316, 2009.

[Guo *et al.*, 2016] Baolin Guo, Chenping Hou, Feiping Nie, and Dongyun Yi. Semi-supervised multi-label dimensionality reduction. In *ICDM*, pages 919–924, 2016.

[Lampert *et al.*, 2009] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

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

[Liu and Tsang, 2015] Weiwei Liu and Ivor Tsang. On the optimality of classifier chain for multi-label classification. In *NIPS*, pages 712–720, 2015.

[Liu *et al.*, 2006] Jing Liu, Mingjing Li, Wei-Ying Ma, et al. An adaptive graph model for automatic image annotation. In *ACM SIGMM*, pages 61–70, 2006.

[Liu *et al.*, 2017] Weiwei Liu, Ivor W Tsang, and Klaus-Robert Müller. An easy-to-hard learning paradigm for multiple classes and multiple labels. *JMLR*, pages 3300–3337, 2017.

[Liu *et al.*, 2018] Weiwei Liu, Donna Xu, Ivor Tsang, and Wenjie Zhang. Metric learning for multi-output tasks. *IEEE TPAMI*, 2018.

[Merris, 1994] Russell Merris. Laplacian matrices of graphs: a survey. *Linear algebra and its applications*, pages 143–176, 1994.

[Ng *et al.*, 2001] Andrew Y Ng, Michael I Jordan, Yair Weiss, et al. On spectral clustering: Analysis and an algorithm. In *NIPS*, volume 14, pages 849–856, 2001.

[Nie *et al.*, 2012] Feiping Nie, Dong Xu, and Xuelong Li. Initialization independent clustering with actively self-training method. *IEEE TSMC*, 42(1):17–27, 2012.

[Nie *et al.*, 2016] Feiping Nie, Wei Zhu, and Xuelong Li. Unsupervised feature selection with structured graph optimization. In *AAAI*, pages 1302–1308, 2016.

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. pages 2408–2414. AAAI, 2017.

[Patterson and Hays, 2012] Genevieve Patterson and James Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, pages 2751–2758, 2012.

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

[Sindhwani and Belkin, 2005] Vikas Sindhwani and Mikhail Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML*, pages 824–831, 2005.

[Trohidis *et al.*, 2008] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, pages 325–330, 2008.

[Wah *et al.*, 2011] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011.

[Wang *et al.*, 2018] Lichen Wang, Zhengming Ding, and Yun Fu. Learning transferable subspace for human motion segmentation. In *AAAI*, 2018.

[Wu *et al.*, 2015] Baoyuan Wu, Siwei Lyu, and Bernard Ghanem. Ml-mg: multi-label learning with missing labels using a mixed graph. In *ICCV*, pages 4157–4165, 2015.

[Zha *et al.*, 2009] Zheng-Jun Zha, Tao Mei, Jingdong Wang, Zengfu Wang, and Xian-Sheng Hua. Graph-based semi-supervised learning with multiple labels. *JVCIR*, pages 97 – 103, 2009.

[Zhu *et al.*, 2003] Xiaojin Zhu, Zoubin Ghahramani, et al. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.

[Zhu, 2005] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.