Max-Margin Zero-Shot Learning for Multi-class Classification

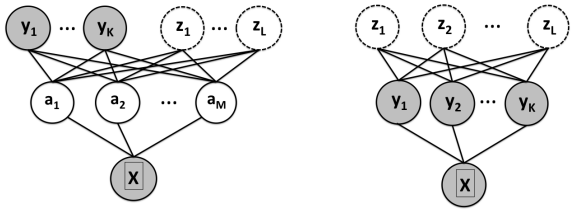
Xin Li and **Yuhong Guo**
Department of Computer and Information Sciences
Temple University, Philadelphia, PA 19122, USA
{xinli, yuhong}@temple.edu

Abstract

Due to the dramatic expanse of data categories and the lack of labeled instances, zero-shot learning, which transfers knowledge from observed classes to recognize unseen classes, has started drawing a lot of attention from the research community. In this paper, we propose a semi-supervised max-margin learning framework that integrates the semi-supervised classification problem over observed classes and the unsupervised clustering problem over unseen classes together to tackle zero-shot multi-class classification. By further integrating label embedding into this framework, we produce a dual formulation that permits convenient incorporation of auxiliary label semantic knowledge to improve zero-shot learning. We conduct extensive experiments on three standard image data sets to evaluate the proposed approach by comparing to two state-of-the-art methods. Our results demonstrate the efficacy of the proposed framework.

1 Introduction

Traditionally, learning multi-class classification models notoriously requires a large amount of labeled training instances. However, it is either costly or impractical to prepare a sufficient amount of annotated training instances for every single class given that a real world application can encounter a large number of categories. Moreover, it is important to reduce the cost of collecting new annotations whenever there is an expanse over the label categories with the dramatic increase of the data set. Due to these needs in real word applications, some new learning schemes such as few-shot, one-shot learning [Bart and Ullman, 2005, Fei-Fei et al., 2006, Krause et al., 2014, Lake et al., 2013] and the most challenging case, zero-shot learning [Rohrbach et al., 2013, Da et al., 2014, Mensink et al., 2014, Lampert et al., 2014] become increasingly popular.



(a) (b)

Figure 1: Two major types of zero-shot learning methods. Dark nodes indicate being observed. Y and Z nodes represent observed classes and unseen classes respectively. Left is the graphical representation of attribute-based methods with an attribute layer (a nodes) and right is of similarity-based methods.

Zero-shot learning addresses the problem of transferring knowledge from *observed classes*, in which one has labeled training instances, to automatically classify instances into proper *unseen classes*, in which one does not have labeled training instances. A good zero-shot learning approach can effectively reduce the labeling cost required to annotate new instances whenever a data set expands to include new label categories. It can also relieve the requirement of preparing annotated instances for all classes for learning effective multi-class classification models. Existing zero-shot learning methods typically bridge the gap between the observed classes and the unseen classes by using additional sources of information to build semantic links between the observed and unseen classes. Based on the way of building such semantic links, current zero-shot learning approaches can be grouped

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

into two major types, attribute-based methods and similarity-based methods. The methods of the former type build a latent layer of attributes (or features) as a common subspace representation for the seen and unseen classes [Farhadi et al., 2009, Lampert et al., 2009, Kankuekul et al., 2012, Yu and Aloimonos, 2010, Akata et al., 2013, Rohrbach et al., 2013, Palatucci et al., 2009], whereas the similarity-based methods utilize the semantic relationship between the seen and unseen classes [Rohrbach et al., 2011, Mensink et al., 2013, 2014]. Both types are illustrated in Figure 1. However, the attribute-based methods require not only the instance-level (usually images) annotations but also the attribute-level annotations for each instance, which is time and human effort consuming. Moreover, they introduce an intermediate problem of attribute classification. The similarity-based methods on the other hand heavily rely on the seen classes to express the unseen classes. They usually are only used for predicting unseen classes. Moreover, most of these existing methods, attribute-based or similarity-based, do not exploit unlabeled data which are typically adequate in amount and economic to collect.

In this work, we propose a semi-supervised max-margin classification framework that exploit both labeled and unlabeled data for zero-shot learning. Specifically, our framework integrates the semi-supervised classification problem on observed classes and the unsupervised clustering problem on unseen classes into a unified max-margin multi-class classification formulation. This framework treats the observed classes and unseen classes in an equal way and the classifier produced can be applied to categorize an instance to the most suitable class among all these classes, which overcomes the drawback of the similarity based methods. By further integrating a label embedding idea into this framework, we derive a kernelized dual formation of the max-margin model, which permits leverage of auxiliary linguistic resources or expert-specified knowledge to improve zero-shot learning performance. We conduct experiments on three standard image data sets for zero-shot learning to evaluate the proposed approach by comparing to two state-of-the-art zero-shot learning methods. Our results demonstrate the efficacy of the proposed framework.

2 Related Work

Zero-shot learning addresses the challenging problem of knowledge transfer from observed classes to unseen classes (i.e., novel classes) in which there are no labeled training instances at all. Existing zero-shot learning methods can be grouped into two major groups, attribute-based methods and similarity-based methods, based on the different ways of building semantic links between the observed and unseen classes.

Attribute-based methods build an intermediate layer of attributes or features across the observed and unseen classes to bridge the semantic gap. Many such works have been studied on image classification problems. [Farhadi et al., 2009] and [Lampert et al., 2009] belong to the first few attribute-based works that address zero-shot image classification tasks. Farhadi et al. [2009] introduced a feature selection method for learning attributes that generalize well across categories. Lampert et al. [2009] proposed a *direct attribute prediction* (DAP) method that learns binary attribute classifiers and then use the output of these attribute classifiers as image representation for image-level classification. Kankuekul et al. [2012] applied the self-organizing and incremental neural networks as the learning mechanism, and their model can learn new attributes and update existing attributes in an online incremental manner. Yu and Aloimonos [2010] extended the author-topic model for attribute prediction based on the analogy of document-author relation and attribute-class relation. Akata et al. [2013] casted the problem of attribute-based classification as a label embedding problem to avoid attribute classifier learning. Their method however requires the association measure between each class and each attribute. Rohrbach et al. [2011] utilized linguistic knowledge bases to build the attribute inventory automatically and discover the association between attributes and object classes. Though they exploit knowledge from linguistic sources such as *WordNet* and *Wikipedia*, their model does not utilize unlabeled data in the training process. Rohrbach et al. [2013] proposed a *propagated semantic transfer* (PST) method that incorporates the mid-layer of semantic attributes and linguistic information. This method shares similarity with our proposed approach as we both exploit knowledge from auxiliary resources and unlabeled data. However, the PST method is a graph-based learning algorithm while our proposed approach provides a discriminative max-margin learning framework. Moreover the PST method incorporates a mid-level layer which requires performing intermediate inference, while our proposed method does not have such an issue. In addition to these methods developed on image classifications, Palatucci et al. [2009] presented a model that utilizes an intermediate set of features derived from a semantic knowledge base for zero-shot learning on fMRI data of neural activities. This work again does not exploit unlabeled data.

Similarity-based methods perform zero-shot classification on unseen classes using the classifiers trained for observed classes and the relationships between the observed and unseen classes. Rohrbach et al. [2011] proposed a similarity-based model by representing the

novel classes using the observed ones with the assistance of linguistic knowledge, such as Wikipedia and WordNet. Mensink et al. [2013] proposed a similarity-based classifier, named nearest class mean, for large-scale zero-shot learning. Mensink et al. [2014] extensively explored various metrics to measure co-occurrence between different classes for zero-shot classification. Recently, Norouzi et al. [2014] took semantic label embeddings into account, and assumed the unseen classes can be predicted as convex combinations of the observed classes in the given semantic embedding space. These works however only performed zero-shot classifications on unseen classes, and they may not be much usable when the testing labels and training labels are not disjoint.

3 Proposed Approach

In this section, we present a semi-supervised max-margin classification framework to address multi-class zero-shot learning. For simplicity, the following notations are used in the presentation: We use $\mathbf{1}$ to denote a column vector with all 1 values, assuming its length can be determined from the context; use $\mathbf{1}_k$ to denote a column vector with all zeros but a single 1 in its k-th entry. We use I_t to denote an identity matrix with size t, use $O_{r,c}$ to denote a $r \times c$ matrix with all zeros, use X_i to denote the i-th row of matrix X, and use $\|\cdot\|_F$ to denote the Frobenius matrix norm.

3.1 Semi-supervised Max-margin Classification Framework

We assume a training set $\mathcal{D} = (X, Y)$ for a L-class classification problem, which contains t_ℓ labeled instances $X^\ell \in \mathbb{R}^{t_\ell \times d}$ with an observed label matrix $Y^\ell \in \{0,1\}^{t_\ell \times L}$, and t_u unlabeled instances $X^u \in \mathbb{R}^{t_u \times d}$ with an unknown label matrix $Y^u \in \{0,1\}^{t_u \times L}$, such that $X = [X^\ell; X^u]$, $Y = [Y^\ell; Y^u]$ and $t = t_\ell + t_u$. For each row label vector Y^ℓ_i, there will be a single 1 entry which denotes the class membership of the instance. Furthermore, without loss of generality, we assume the observed labels only appear in the first L_ℓ classes and hence the last $(L - L_\ell)$ classes are unseen classes, while the unlabeled instances can belong to any of the total L classes. We assume there are a sufficient number of unlabeled instances such that their latent labels will cover all the unseen classes. We aim to learn a L-class classifier on this training set.

The learning problem is apparently a zero-shot learning problem. In particular, we treat the problem as a standard semi-supervised learning problem over the first L_ℓ observed classes, and an unsupervised clustering problem over the $(L - L_\ell)$ unknown classes. We propose to integrate these two parts in a latent max-margin multi-classification framework below:

$$\begin{aligned} \min_{Y,W,\boldsymbol{\xi}} \quad & \frac{\beta}{2}\|W\|_F^2 + \mathbf{1}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & \mathrm{diag}\big((Y - \mathbf{1}\mathbf{1}_k^\top) W X^\top\big) \geq (\mathbf{1} - Y\mathbf{1}_k) - \boldsymbol{\xi},\ \forall k \in \mathcal{Y} \\ & Y \in \{0,1\}^{t \times L}, \quad BY = Y^\ell, \\ & Y\mathbf{1} = \mathbf{1}, \quad a\mathbf{1} \leq Y^\top \mathbf{1} \leq b\mathbf{1} \end{aligned} \tag{1}$$

where $W \in \mathbb{R}^{L \times d}$ is the model parameter matrix, $\boldsymbol{\xi}$ is a length t vector that captures the hinge loss over all the t instances; the set $\mathcal{Y} = \{1, \cdots, L\}$ is the class index set, and $B = [I_{t_\ell}, O_{t_\ell, t_u}]$ is a selector matrix that selects the first t_ℓ rows of Y. The constraint $a\mathbf{1} \leq Y^\top \mathbf{1} \leq b\mathbf{1}$ is introduced to avoid degenerated clustering, where the positive constants a and b are used to set a lower bound and an upper bound respectively on the number of instances assigned into each of the L classes. Note if the label matrix Y is fully observed, the optimization problem in (1) will be equivalent to the standard multi-class SVM model in [Crammer and Singer, 2001].

We expect the discriminative framework above can build a good foundation for zero-shot learning. But to produce an effective zero-shot learning model, we still need additional knowledge to bridge the gap between observed classes and unseen classes. Below we will extend the framework above by integrating a label embedding idea, which provides a natural form to incorporate auxiliary label relatedness knowledge.

For many real world multi-class classification problems, the class labels have semantic meanings. For example, for an image data set with classes {*people, street, car,* $\cdots$}, these class labels express more information than simple class index numbers. Hence instead of simplifying each class label into a class index, we propose to incorporate a label embedding idea into our zero-shot learning framework and express each class label as a semantic vector $\mathbf{u} \in \mathbb{R}^h$. For all the L classes, we form a label embedding matrix $U \in \mathbb{R}^{L \times h}$, whose each row contains the embedding vector for the corresponding class label. Given U, the label embedding vector for the i-th training instance can be written as $Y_i U$, where the label indicator vector is used as a selection vector. By incorporating this label embedding matrix, we extend the semi-supervised max-margin formulation in (1) into the following max-margin co-embedding formulation:

$$\begin{aligned} \min_{Y,U,W,\boldsymbol{\xi}} \quad & \frac{\beta}{2}\|W\|_F^2 + \frac{\beta}{2}\|U\|_F^2 + \mathbf{1}^\top \boldsymbol{\xi} \\ \text{s.t.} \quad & \mathrm{diag}\big((Y - \mathbf{1}\mathbf{1}_k^\top) U W X^\top\big) \geq (\mathbf{1} - Y\mathbf{1}_k) - \boldsymbol{\xi},\ \forall k \in \mathcal{Y} \\ & Y \in \{0,1\}^{t \times L}, \quad BY = Y^\ell, \\ & Y\mathbf{1} = \mathbf{1}, \quad a\mathbf{1} \leq Y^\top \mathbf{1} \leq b\mathbf{1} \end{aligned} \tag{2}$$

where the model parameter matrix $W \in \mathbb{R}^{h \times d}$ maps

the input feature vectors into the label embedding space to form an input-label co-embedding [Mirzazadeh et al., 2014, Weston et al., 2011].

Proposition 1 *The max-margin co-embedding problem in (2) can be equivalently formulated into the following dual optimization problem:*

$$\begin{aligned}\min_{Y,U}\max_{M}\quad & \frac{\beta}{2}tr(UU^\top) - tr(MY) - \\ & \frac{1}{2\beta}tr\Big((Y^\top - M)K(Y^\top - M)^\top UU^\top\Big) \qquad (3)\\ s.t.\quad & M \geq 0,\ \ M^\top\mathbf{1} = \mathbf{1},\quad Y \in \{0,1\}^{t\times L},\\ & BY = Y^\ell,\ \ Y\mathbf{1} = \mathbf{1},\ \ a\mathbf{1} \leq Y^\top\mathbf{1} \leq b\mathbf{1}\end{aligned}$$

where M is the $L \times t$ dual variable matrix, and $K = XX^\top$ denotes the kernel matrix on the input data. The primal matrix W can be recovered from M via

$$W = \frac{1}{\beta}U^\top(Y^\top - M)X \qquad (4)$$

This proposition can be proved by introducing Lagrangian multipliers and deriving the standard dual formulation of the primal max-margin multi-class optimization problem over W and $\boldsymbol{\xi}$ [Xu and Schuurmans, 2005], while assuming Y and U fixed.

Following the primal parameter matrix recovery equation in (4), we can predict the class index of a new instance $\mathbf{x} \in \mathbb{R}^d$ in terms of the dual model parameters, such as

$$y = \arg\max_{k\in\mathcal{Y}} U_k W\mathbf{x} = \arg\max_{k\in\mathcal{Y}} \frac{1}{\beta}U_k U^\top(Y^\top - M)X\mathbf{x} \qquad (5)$$

3.1.1 Incorporating Auxiliary Label Semantic Similarity Information

The dual formulation in (3) provides a general framework for incorporating auxiliary label semantic similarity information. Note the label embedding matrix U only appears in a product form $UU^\top$ in the objective function of (3). Let $Q = UU^\top$. Then the $L \times L$ matrix Q is a label covariance matrix which encodes the semantic similarity of each pair of labels. Information about such a label semantic relatedness matrix can be extracted from different auxiliary resources. For example, in some image classification problems, the attribute based representation of the class labels can be available [Kemp et al., 2006, Patterson and Hays, 2012, Farhadi et al., 2009]. In this case, a fixed label similarity matrix $\widehat{Q}$ can be computed by measuring the similarity of the attribute-based label vectors. In more general cases where one does not have meaningful attribute-based label representations available, auxiliary semantic information can still be extracted from a large free text corpus such as Wikipedia through methods such as explicit semantic analysis (ESA) [Gabrilovich and Markovitch, 2007]. Using ESA, one can represent a label phrase as a vector which contains the statistical appearance information of the phase in the large number of articles of the text corpus. A label similarity matrix $\widehat{Q}$ can then be computed based on the similarity measures of the label semantic representation vectors. Instead of learning the label embedding matrix U, we can directly incorporate the label similarity matrix into our learning problem (3) by replacing $UU^\top$ with the pre-computed $\widehat{Q}$. By further relaxing the integer constraints over Y, we have the following min-max optimization problem:

$$\begin{aligned}\min_{Y}\max_{M}\quad & -\mathrm{tr}(MY) - \frac{1}{2\beta}\mathrm{tr}\Big((Y^\top - M)K(Y^\top - M)^\top\widehat{Q}\Big) \qquad (6)\\ \text{s.t.}\quad & M \geq 0,\ \ M^\top\mathbf{1} = \mathbf{1},\quad Y \geq 0,\\ & BY = Y^\ell,\ \ Y\mathbf{1} = \mathbf{1},\quad a\mathbf{1} \leq Y^\top\mathbf{1} \leq b\mathbf{1}\end{aligned}$$

3.2 Training Algorithm

The min-max optimization problem (6) is a non-convex optimization problem. Though a simple alternating optimization procedure over Y and M can provide an intuitive solution, such an algorithm is not guaranteed to converge to local optima or stationary points. In this section, we instead present a first-order local conditional gradient descent algorithm to solve the optimization problem (6).

Let $\mathcal{L}(Y, M)$ denote the objective function of (6). We first re-express this min-max optimization problem as a minimization problem over a non-smooth and non-convex objective function $\mathcal{F}(Y)$,

$$\min_{Y\in\Omega_Y}\quad \mathcal{F}(Y) \qquad (7)$$

$$\text{with}\qquad \mathcal{F}(Y) = \mathcal{L}(Y, M_Y^*) = \max_{M\in\Omega_M}\ \mathcal{L}(Y, M) \qquad (8)$$

$$M_Y^* = \arg\max_{M\in\Omega_M}\ \mathcal{L}(Y, M). \qquad (9)$$

Here Ω_Y and Ω_M denote the feasible sets defined by the constraints over Y and M respectively, such that $\Omega_Y = \{Y \in \mathbb{R}^{t\times L} :\ BY = Y^\ell,\ Y \geq 0,\ Y\mathbf{1} = \mathbf{1},\ a\mathbf{1} \leq Y^\top\mathbf{1} \leq b\mathbf{1}\}$ and $\Omega_M = \{M \in \mathbb{R}^{L\times t} : M \geq 0,\ M^\top\mathbf{1} = \mathbf{1}\}$. Note M_Y^* denotes the optimal solution for the inner maximization problem over M given the fixed Y matrix. With a fixed Y, the inner maximization problem over M in (9) is a standard quadratic optimization problem and can be efficiently solved using a standard quadratic solver.

Then we develop a local conditional gradient descent procedure to iteratively solve the minimization problem in (7). We first randomly initialize a feasible matrix $Y_{(0)}$ and then repeatedly make updates in each

Algorithm 1: Local Conditional Gradient Descent

Input: K, $\widehat{Q}$, Y^ℓ, β, a, b
Initialize $Y_{(0)}$, and set $r = 0$.
Repeat
1. Compute subgradient $\nabla_Y \mathcal{F}(Y_{(r)})$ at the current point using Eq. (10).
2. Solve linear programming in Eq. (11) for $\widehat{Y}$.
3. Conduct backtracking line search to select an optimal step-size η^*:

$$\eta^* = \arg\min_{0 \le \eta \le 1} \mathcal{F}((1-\eta)Y_{(r)} + \eta\widehat{Y})$$

4. Set $Y_{(r+1)} = (1-\eta^*)Y_{(r)} + \eta^*\widehat{Y}$, $Y^* = Y_{(r+1)}$
5. **if** the difference between $Y_{(r+1)}$ and $Y_{(r)}$ is small enough **then** break out **endif**
6. Set $r = r + 1$.

Until max-iters are reached

iteration. At the $(r+1)$-th iteration, we first compute the subgradient of $\mathcal{F}(Y)$ at the current point $Y_{(r)}$:

$$\nabla_Y \mathcal{F}(Y_{(r)}) = -M^*_{(r)} - \frac{1}{\beta} K (Y_{(r)}^\top - M^*_{(r)})^\top \widehat{Q} \quad (10)$$

where $M^*_{(r)}$ is a simplification of $M^*_{Y_{(r)}}$. Next we compute an intermediate point using the first order method by solving a convex linear programming:

$$\widehat{Y} = \arg\min_{Y \in \Omega_Y} \ \mathrm{tr}(Y^\top \nabla_Y \mathcal{F}(Y_{(r)})) \quad (11)$$

Finally we find an optimal step-size η^* by performing backtracking line search over $0 \le \eta \le 1$ to minimize the objective $\mathcal{F}((1-\eta)Y_{(r)} + \eta\widehat{Y})$, and determine the next point as

$$Y_{(r+1)} = (1-\eta^*)Y_{(r)} + \eta^*\widehat{Y}. \quad (12)$$

The overall algorithm is presented in Algorithm 1. After solving for the optimal solution Y^* which contains continuous values between 0 and 1, we round Y^* back to class indicator values in a heuristic way by setting the entry with the largest value in each row to 1 and all other entries to 0s.

4 Experiments

4.1 Experimental Setting

Data sets. We evaluated the proposed approach on three standard zero-shot image data sets, *Animal with Attributes* introduced in [Kemp et al., 2006], *Attribute of Pascal* introduced in [Farhadi et al., 2009] and *SUN Attributes* introduced by [Patterson and Hays, 2012]. The *Animal with Attribute (AwA)* data set has 30,475 images across 50 mammal classes (e.g. lion, fox) and it also provides 85 binary semantic attributes (e.g. furry, spots). For each image of this data set, PHOG feature vectors [Bosch et al., 2007] are extracted separately for all 21 cells of a 3-level spatial pyramids (1×1, 2×2, 4×4). On each cell, a 12-dimensional base histogram is extracted and concatenated to form a 252-dimensional (21×12) feature vector. *Attribute of Pascal (aPascal)* is a subset of the *aPascal-aYahoo* data set. *aPascal* consists a 12,695-image subset of the PASCAL VOC 2008 data set with 20 object classes. Each image in this data set has been annotated with 64 binary attributes that characterize the visible objects. The feature vector of each image provided within this data set is a 9,751-dimensional vector, including local texture, HOG, edge and color descriptors. We performed dimension reduction on the long vectors using PCA and obtained 500-dimensional vectors for all the images. The *SUN Attributes (SUNA)* data set is a subset of the *SUN Database* [Xiao et al., 2010] for fine-grained scene categorization. It consists of 14,340 images from 717 classes (20 image per class). Each image is annotated with 102 binary attributes that describe the scenes' material and surface properties as well as lighting conditions, functions, affordances, and general image layout. In our experiment, we randomly selected a 50-class subset out of 717 classes and used the GIST feature vectors [Oliva and Torralba, 2001] for image representation. Table 1 summarizes the characteristics of all the three data sets used in our experiments.

Table 1: Statistical properties of the preprocessed data sets used in the experiments.

Dataset	Images	Classes	Attributes	Feature
AwA	30,475	50	85	252
aPascal	12,695	20	64	500
SUNA	1,000	50	102	512

Comparison methods. In order to evaluate the performance of the proposed method, we compared it with two recently published zero-shot learning approaches, Direct Attribute Prediction (DAP) [Lampert et al., 2014] and Propagated Semantic Transfer (PST) [Rohrbach et al., 2013], both of which are attribute-based methods and can be tested on data with both seen and unseen classes. The *DAP* method assumes that the class-attribute relationship is given and trains a probabilistic classifier for each attribute. MAP prediction of the classes can then be conducted for the test instances according to the attribute probabilities. The *PST* method performs label propagation on a graph structure over all (training and testing) instances and the graph is constructed based on their attribute representations. The results of both methods are repro-

Table 2: Test results (mean±std) in terms of classification accuracy on all three data sets with different numbers of unseen classes. The test results on unseen categories, seen categories and all categories (mixed) are all reported. Bold font indicates the best results.

Dataset	Unseen	Test	Proposed	DAP	PST
AwA	5	Unseen	**0.416 ± 0.022**	0.382 ± 0.036	0.375 ± 0.024
		Seen	**0.729 ± 0.025**	0.683 ± 0.031	0.704 ± 0.019
		Mix	**0.596 ± 0.024**	0.557 ± 0.037	0.553 ± 0.021
	10	Unseen	**0.382 ± 0.023**	0.371 ± 0.039	0.357 ± 0.021
		Seen	**0.741 ± 0.026**	0.703 ± 0.033	0.732 ± 0.024
		Mix	**0.591 ± 0.025**	0.553 ± 0.035	0.550 ± 0.023
aPascal	5	Unseen	**0.275 ± 0.016**	0.254 ± 0.015	0.261 ± 0.019
		Seen	**0.601 ± 0.013**	0.534 ± 0.017	0.554 ± 0.013
		Mix	**0.455 ± 0.014**	0.408 ± 0.016	0.427 ± 0.014
	10	Unseen	**0.187 ± 0.021**	0.168 ± 0.018	0.162 ± 0.021
		Seen	**0.637 ± 0.018**	0.559 ± 0.016	0.582 ± 0.012
		Mix	**0.419 ± 0.019**	0.373 ± 0.017	0.381 ± 0.015
SUNA	5	Unseen	**0.252 ± 0.018**	0.241 ± 0.018	0.247 ± 0.021
		Seen	**0.501 ± 0.019**	0.487 ± 0.016	0.492 ± 0.023
		Mix	**0.405 ± 0.019**	0.398 ± 0.017	0.399 ± 0.022
	10	Unseen	**0.189 ± 0.025**	0.181 ± 0.020	0.183 ± 0.024
		Seen	**0.512 ± 0.023**	0.501 ± 0.019	0.506 ± 0.024
		Mix	**0.371 ± 0.024**	0.359 ± 0.019	0.363 ± 0.024

duced using the code released by the authors.

Implementation. On each data set, we randomly selected k classes to be the unseen classes, where k equals to 5 or 10 in our experiments. For the *AwA* and *aPascal* data sets, we use 50% of the instances for training and the rest for testing. That is, 15,238 instances in *AwA* and 6,348 instances in *aPascal* are randomly selected as training examples. Training set is further divided into labeled, unlabeled and validation sets, where both labeled and validation sets are instances from known classes but unlabeled set contains instances from either known or unseen classes. The numbers of labeled, unlabeled and validation instances are 6942, 4130, 4166 respectively on *AwA* and 2668, 2079 and 1601 on *aPascal*. On the *SUNA* data set, we randomly selected 362 images as labeled ones, 221 as unlabeled ones and 217 as validation set, while the rest 200 instances are used for testing. Parameter selection is conducted by evaluating the classification performance of each approach on the validation set. For our proposed approach, we set a=ceil$(0.5 \times t/L)$ and b=ceil$(2 \times t/L)$ on each data set, while selecting the trade-off parameter β from $\{.001, .01, .01, .1, 1, 10, 100, 1000\}$. For the DAP method, its parameter C is selected from $\{.001, .01, .01, .1, 1, 10, 100, 1000\}$. For the PST method, its parameter δ is set to 0.15 and k is set to 50, and the trade-off parameter α is selected from $\{.001, .01, .01, .1, 1, 10, 100, 1000\}$. As all comparison methods can use an input kernel matrix, we used a RBF kernel with free parameter 1 in all experiments. The results reported in this section are averages of 5 repeated runs and each individual run involves model training, parameter selection and model testing for each comparison method.

4.2 Classification results

We first compared our proposed approach with the two related methods, DAP and PST, in terms of classification accuracy on the three image data sets, where class-attribute relationships are included in the data sets. The two comparison methods are attribute-based methods and they require the knowledge of attribute-based class representations on each data set. In order to provide a fair comparison, we hence incorporated the same auxiliary knowledge in our proposed approach, by leveraging the class-attribute relation (CAR) knowledge to compute a label similarity matrix $\widehat{Q}$ with the cosine similarity measure.

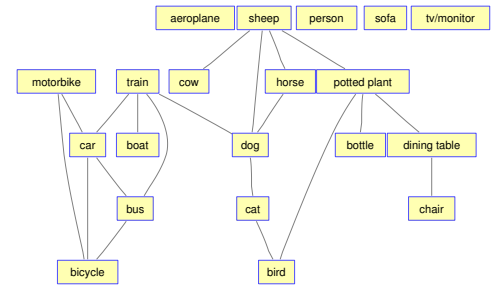
We conducted experiments on the three data sets with two different numbers of unseen classes, 5 and 10, while using the remaining classes as observed classes. In each case, we recorded the test accuracy results on unseen classes, seen classes, and the mixed all classes. The average classification results for each setting are reported in Table 2. We can see that with the number of unseen classes increasing from 5 to 10, the test accuracies decrease in general for all methods on unseen classes, but increase on seen classes due to better discrimination capability between fewer observed classes. Neverthe-

less, the proposed method produces the best results consistently across all the test cases, and it outperforms the other two methods with remarkable margins not only on unseen classes but also on seen classes. For example, on AwA our method improves the test accuracy in unseen classes by about 3% comparing to DAP and by about 4% comparing to PST when 5 classes are set unseen. Similar results can be observed on the other two data sets as well. These results demonstrate the efficacy of the proposed max-margin framework.

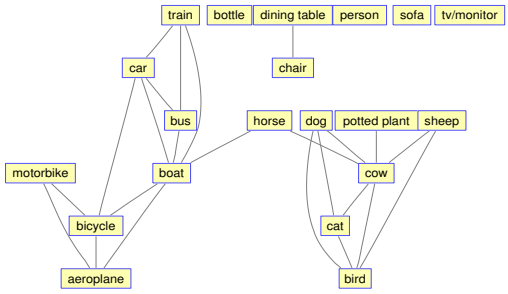
On the other hand, since all the comparison methods utilize class-attribute relationships to guide the model learning, one may wonder what makes the proposed method more effective. One possible reason is that both of DAP and PST involve an intermediate step of attribute inference. Such intermediate inference of attributes can introduce noise into image-level classification. Our method on the other hand incorporates the class-attribute relationship seamlessly into the max-margin framework in the form of label correlation matrix. Moreover, our semi-supervised max-margin model provides a discriminative framework for exploiting both labeled and unlabeled data.

4.3 Impact of External Knowledge

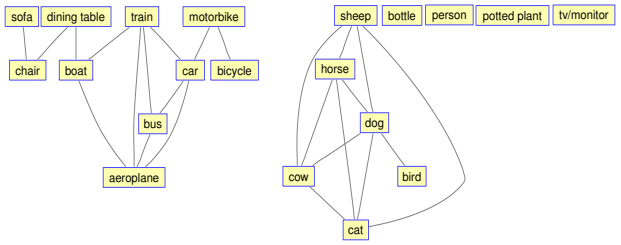
Different from the attribute-based methods, our proposed model can incorporate external auxiliary knowledge beyond the class-attribute relation knowledge included in the data sets. We next investigated the capacity of our proposed model in exploiting different auxiliary label relation knowledge. In this study, besides the class-attribute relation (CAR), we have also explored two other types of external knowledge, Explicit Semantic Analysis [Gabrilovich and Markovitch, 2007] and Word Embedding [Collobert et al., 2011]. ESA and WE provide linguistic knowledge learned from Wikipedia. Explicit Semantic Analysis (ESA) represents an input word by a vector of its appearance records in a set of concepts in Wikipedia. Word Embedding (WE) is trained by neural networks using an earlier dump of Wikipedia. It represents each input word using a 50-dimensional vector. We can use ESA and WE to provide semantic vector representations for the class labels in our data sets. Then the label similarity matrix $\widehat{Q}$ can be computed over the label vectors using a cosine similarity measure. Figure 2 uses graphs to illustrate the label similarity matrices obtained on the *aPascal* data set using CAR, ESA and WE respectively. To avoid cluttered graphs, we applied a threshold on the matrices. The thresholds are selected to preserve the top 20% edges for each graph. An edge in each graph indicates that the label nodes connected have close semantic relationship. From the three graphs, we can observe some interesting patterns such as transportation tools are connected with each other and animals are connected with each other. On the other hand, the three graphs differ from each other in some details. For example, the label "sofa" in ESA and WE is an isolated node after thresholding, but it has a connection with "chair" in CAR. We looked into the attributes of "sofa" and "chair", and found they have 10 attributes in common, including "has leg", "has back", and "has arm". CAR provides richer information than the other two types of external knowledge because it is directly provided by human experts based on their interpretations of the label concepts while ESA and WE representations are learned automatically from free textual documents.



(a) ESA



(b) WE



(c) CAR

Figure 2: Graphical illustration of Semantic Relatedness between Classes in the aPascal data set. An edge indicates that two nodes has semantic relationship.

Table 3: Test results (mean±std) in terms of classification accuracy on all three data sets for the proposed approach with different external knowledge. The test results on unseen categories, seen categories and all categories (mixed) are all reported. Bold font indicates the best results.

Dataset	Unseen	Test	Proposed+CAR	Proposed+ESA	Proposed+WE	Proposed+ø
AwA	5	Unseen	**0.416 ± 0.022**	0.401 ± 0.025	0.351 ± 0.026	0.264 ± 0.022
		Seen	**0.729 ± 0.025**	0.712 ± 0.021	0.698 ± 0.020	0.682 ± 0.023
		Mix	**0.596 ± 0.024**	0.579 ± 0.022	0.543 ± 0.022	0.481 ± 0.023
	10	Unseen	**0.382 ± 0.023**	0.376 ± 0.023	0.347 ± 0.021	0.152 ± 0.025
		Seen	**0.741 ± 0.026**	0.732 ± 0.021	0.720 ± 0.024	0.706 ± 0.025
		Mix	**0.591 ± 0.025**	0.576 ± 0.022	0.549 ± 0.023	0.437 ± 0.025
aPascal	5	Unseen	**0.275 ± 0.016**	0.259 ± 0.017	0.239 ± 0.016	0.210 ± 0.016
		Seen	**0.601 ± 0.013**	0.598 ± 0.016	0.587 ± 0.015	0.581 ± 0.016
		Mix	**0.455 ± 0.014**	0.438 ± 0.016	0.422 ± 0.015	0.401 ± 0.016
	10	Unseen	**0.187 ± 0.021**	0.177 ± 0.020	0.148 ± 0.021	0.126 ± 0.022
		Seen	**0.637 ± 0.018**	0.632 ± 0.019	0.615 ± 0.021	0.613 ± 0.020
		Mix	**0.419 ± 0.019**	0.414 ± 0.020	0.394 ± 0.021	0.382 ± 0.021
SUNA	5	Unseen	**0.252 ± 0.018**	0.247 ± 0.016	0.245 ± 0.020	0.203 ± 0.025
		Seen	**0.501 ± 0.019**	0.494 ± 0.019	0.492 ± 0.020	0.486 ± 0.025
		Mix	**0.405 ± 0.019**	0.399 ± 0.018	0.397 ± 0.020	0.359 ± 0.025
	10	Unseen	**0.189 ± 0.025**	0.185 ± 0.019	0.184 ± 0.021	0.103 ± 0.024
		Seen	**0.512 ± 0.023**	0.508 ± 0.020	0.505 ± 0.020	0.491 ± 0.023
		Mix	**0.371 ± 0.024**	0.367 ± 0.020	0.364 ± 0.020	0.309 ± 0.023

We compared the classification performance of the proposed approach with three different auxiliary knowledge, CAR, ESA and WE. We have also compared these three variants of the proposed approach to a baseline variant which uses the proposed model without any auxiliary knowledge by setting $\widehat{Q}$ to an identity matrix. The comparison results of the four variant methods are reported in Table 3. From the table, we can see all the three variants with auxiliary knowledge outperform the baseline with empty auxiliary knowledge. But even without any bridge between the seen and unseen classes, the baseline performs much better than random guesses, except on *SUNA*, where the training set is small and external knowledge becomes more critical. The proposed method with CAR outperforms the variant methods with ESA or WE auxiliary knowledge across all data sets. This is reasonable since CAR provides human expert knowledge while ESA and WE are from free linguistic resources. Even so, the difference between the results of *Proposed+CAR* and *Proposed+ESA* are quite small. Moreover, if we compare the results across Table 2 and Table 3, we can see that *Proposed+ESA* outperforms both *DAP* and *PST* across almost all the evaluation cases. These results suggest that it is helpful to use either expert-specified knowledge or external knowledge from auxiliary sources in context of zero-shot learning, while our proposed max-margin approach provides an effective framework to utilize different auxiliary resources for zero-shot learning.

5 Conclusion

In this paper, we developed a semi-supervised max-margin classification framework to tackle the challenging problem of zero-shot learning. Specifically, our framework integrates the semi-supervised classification problem over the observed classes and the unsupervised clustering problem over the unseen classes into a unified max-margin multi-class classification formulation, which exploits both labeled and unlabeled data. We further integrated the label embedding idea into this framework to produce a kernelized dual classification model, which provides the capacity of leveraging external linguistic or expert-specified knowledge to assist zero-shot learning. To evaluate the performance of the proposed model, we performed extensive experiments on multiple standard image data sets, by comparing the proposed approach to two state-of-the-art zero-shot learning methods. The experimental results demonstrated the efficacy of the proposed model and its superiority over the comparison methods. We have also investigated the capacity of the proposed model on integrating different auxiliary knowledge. The results showed that our model produced good performance even with free available linguistic resources.

Acknowledgements

This research was supported in part by NSF grant IIS-1422127 and IIS-1065397.

References

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings of CVPR*, 2013.

E. Bart and S. Ullman. Single-example learning of novel classes using representation by similarity. In *Proceedings of BMVC*, 2005.

A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of CIVR*, 2007.

R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537, 2011.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001.

Q. Da, Y. Yu, and Z. Zhou. Learning with augmented class by exploiting unlabeled data. In *Proceedings of AAAI*, 2014.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of CVPR*, 2009.

Li Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *TPAMI*, 28(4):594–611, 2006.

E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, 2007.

P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa. Online incremental attribute-based zero-shot learning. In *Proceedings of CVPR*, 2012.

C. Kemp, J. Tenenbaum, T. Griffiths, T. Yamada, and N. Ueda. Learning systems of concepts with an infinite relational model. In *Proceedings of AAAI*, 2006.

E. Krause, M. Zillich, T. Williams, and M. Scheutz. Learning to recognize novel objects in one shot through human-robot interactions in natural language dialogues. In *Proceedings of AAAI*, 2014.

B. Lake, R. Salakhutdinov, and J. Tenenbaum. One-shot learning by inverting a compositional causal process. In *Proceedings of NIPS*, 2013.

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

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

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *TPAMI*, 35(11):2624–2637, 2013.

T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of CVPR*, 2014.

F. Mirzazadeh, Y. Guo, and D. Schuurmans. Convex co-embedding. In *Proceedings of AAAI*, 2014.

M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. In *Proceedings of ICLR*, 2014.

A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, 2009.

G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceeding of CVPR*, 2012.

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

M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Proceedings of NIPS*, 2013.

J. Weston, S. Bengio, and N. Usunier. WSABIE: Scaling up to large vocabulary image annotation. In *Proceedings of IJCAI*, 2011.

J. Xiao, J. Hays, K. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Proceedings of CVPR*, 2010.

L. Xu and D. Schuurmans. Unsupervised and semi-supervised multi-class support vector machines. In *Proceedings of AAAI*, 2005.

X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of ECCV*, 2010.