Boosted Zero-Shot Learning with Semantic Correlation Regularization

Te Pi[1], **Xi Li**[1,2*], **Zhongfei (Mark) Zhang**[1]
[1]Zhejiang University, Hangzhou, China
[2]Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Hangzhou, China
peterpite@zju.edu.cn; xilizju@zju.edu.cn; zhongfei@zju.edu.cn

Abstract

We study zero-shot learning (ZSL) as a transfer learning problem, and focus on the two key aspects of ZSL, model effectiveness and model adaptation. For effective modeling, we adopt the boosting strategy to learn a zero-shot classifier from weak models to a strong model. For adaptable knowledge transfer, we devise a Semantic Correlation Regularization (SCR) approach to regularize the boosted model to be consistent with the inter-class semantic correlations. With SCR embedded in the boosting objective, and with a self-controlled sample selection for learning robustness, we propose a unified framework, Boosted Zero-shot classification with Semantic Correlation Regularization (BZ-SCR). By balancing the SCR-regularized boosted model selection and the self-controlled sample selection, BZ-SCR is capable of capturing both discriminative and adaptable feature-to-class semantic alignments, while ensuring the reliability and adaptability of the learned samples. The experiments on two ZSL datasets show the superiority of BZ-SCR over the state-of-the-arts.

1 Introduction

Recent years have witnessed a widespread increase of interest in Zero-Shot Learning (ZSL), which aims at learning a classifier from the data of the seen classes $\mathcal{Y}_S$ for classifying the data from the unseen target classes $\mathcal{Y}_T$. A common solution of ZSL is to introduce some auxiliary information on labels (e.g., attributes [Akata *et al.*, 2013], word vector representations [Frome *et al.*, 2013]) to model the semantic relationships between $\mathcal{Y}_S$ and $\mathcal{Y}_T$ in a common structured embedding space. Hence, ZSL is typically solved as a transfer learning problem, which seeks to exploit the shared knowledge among classes and transfer these knowledge to adapt on $\mathcal{Y}_T$.

From the perspective of transfer learning, a zero-shot classifier is expected to be both discriminative for the data distribution of $\mathcal{Y}_S$ and with robust adaptation to the data distribution of $\mathcal{Y}_T$. Therefore, the classifier should not only capture the significant feature-to-class semantic alignments, but also reflect the semantic connections between $\mathcal{Y}_S$ and $\mathcal{Y}_T$. In brief, the two key principles of learning a zero-shot classifier are the model effectiveness and the model adaptation.

First, for effective modeling, we adopt the boosting strategy [Zhou, 2012] as the basis of the ZSL framework, by learning a zero-shot boosting classifier from weak models to a strong model. The superiority of boosting lies in its flexible piecewise approximation to the data distributions based on ensembles of weak hypotheses. Thus, a boosting ZSL model learns in an asymptotic way to sufficiently capture the discriminative patterns in the feature space and their alignments to the semantic patterns in the embedding space, and obtains the learning effectiveness. On the other hand, for a better adaptation to the target classes $\mathcal{Y}_T$, the model needs to explore the shared semantic patterns of $\mathcal{Y}_S$ and $\mathcal{Y}_T$ to reflect their semantic correlations. For this sake, it is reasonable to assume that a well-adapted classifier should generate a high compatibility score for a sample and a target class semantically close to the sample's true class, and a low score otherwise. Based on these analysis, we propose a Semantic Correlation Regularization (SCR) approach to impose this constraint, which encourages a negative correlation between a sample's scores on the target classes and the classes' semantic divergences to the sample's ground truth. In this way, SCR constrains the learned model to be consistent with the source-target semantic correlations, for a semantically adaptable model transfer.

Furthermore, in addition to the label relations, a feasible model adaptation also relies on a modeling of the sample relations for an exploration of the common geometric structures of the feature space and the semantic space. This requires a control or selection scheme for the samples to be learned, to distinguish the samples with robust patterns for classifying $\mathcal{Y}_S$, and with adaptable patterns for the transfer to the semantics of $\mathcal{Y}_T$. Therefore, a self-controlled sample selection of [Zhao *et al.*, 2015] is applicable for ZSL, which adaptively incorporates the samples into learning from easy ones to complex ones, inspired by the human learning process. Such sample selection smoothly controls the learning pace of ZSL by what the model has already learned. With this self-controlled learning pace, a zero-shot model is smoothly guided to focus on those both reliable and adaptable samples to explore the robust feature-semantic alignments, for learning a classifier with good generalization to the target classes.

*Corresponding author

Therefore, motivated by a simultaneous enhancement of the model effectiveness and model adaptation for ZSL as a transfer learning task, we propose a novel framework, Boosted Zero-shot classification with Semantic Correlation Regularization (BZ-SCR). With the SCR embedded into the boosted classification, the boosting process would favor models with not only accurate prediction for an effective classifier, but also with semantic consistency to the target classes for an adaptable model transfer. With a self-controlled sample selection, the BZ-SCR puts emphasis on the reliable and adaptable samples to explore the common feature-semantic structures for a robust model generalization. The proposed framework learns by effectively balancing boosted model selection and self-controlled sample selection. As a result, the proposed framework is capable of jointly considering the learning effectiveness, the cross-semantics adaptation, and the model robustness for learning a zero-shot classifier.

In mathematics, we formulate BZ-SCR as a max-margin boosting optimization with self-controlled sample selection. The contributions of this paper are summarized as follows:

1. We propose a boosted ZSL approach that theoretically seeks for effective knowledge transfer in both model learning effectiveness and cross-domain model adaptation, via max-margin boosting optimization with self-controlled sample selection. To the best of our knowledge, this work is innovative in ZSL by jointly considering model effectiveness, model adaptation, and sample selection within a boost learning framework.

2. We present a novel Semantic Correlation Regularization (SCR) for ZSL, which regularizes the learned boosting model to be consistent with the source-target semantic correlations. In principle, the proposed SCR is capable of effectively capturing the inter-class correlations by modeling the intrinsic geometrical structure properties.

2 Related Work

We review the related work on zero-shot learning, boost learning and self-controlled sample selection approaches.

Known as learning from disjoint training and test classes, ZSL requires the ability to transfer knowledge from classes with training data to classes without. The possible sources of side informations of classes include manually annotated attributes [Akata *et al.*, 2013], semantic class taxonomies [Miller, 1995], and unsupervised word representations [Mikolov *et al.*, 2013]. Based on the way of leveraging these informations, the existing ZSL approaches fall into two categories. One is to build intermediate attribute classifiers and use their probabilistic weightings to make class predictions [Lampert *et al.*, 2014; 2009]. The other group is to represent the semantic knowledge of labels in an embedding space and directly learn a mapping function between the feature space and the embedding space. Among them, CCA [Hastie *et al.*, 2001] maximizes the inter-domain statistical correlations; [Palatucci *et al.*, 2009] learns a linear mapping, ALE [Akata *et al.*, 2016], SJE [Akata *et al.*, 2015] and DeViSE [Frome *et al.*, 2013] learn a bilinear mapping, and [Socher *et al.*, 2013] models a nonlinear regression with neural networks. Efforts for directly tackling the absence of the training data are also made, such as semi-supervised transductive methods [Guo *et al.*, 2016] and the generation of labeled virtual data for unseen classes [Wang *et al.*, 2016].

Boosting is a family of supervised ensemble approaches that build a strong model by successively learning and combining multiple weak models [Zhou, 2012]. The main variation among boosting methods is their ways of weighting samples and weak learners, e.g., Adaboost [Freund and Schapire, 1997], SoftBoost [Rätsch *et al.*, 2007] and LPBoost [Demiriz *et al.*, 2002]. The superiority of boosting lies in its piecewise approximation of a nonlinear decision function to explore the data patterns [Schapire and Freund, 2012].

Proposed by [Kumar *et al.*, 2010], the self-controlled sample selection is inspired by the learning process of humans that gradually incorporates the training samples into learning from easy ones to complex ones. It is initially developed for avoiding the bad local minima of latent models, by learning the data in an order from easy to hard determined by the feedback of the learner itself. It is applied in different applications, such as multimedia reranking [Jiang *et al.*, 2014a], matrix factorization [Zhao *et al.*, 2015], and multiple instance learning [Zhang *et al.*, 2015]. [Meng and Zhao, 2015] provides a theoretical analysis of the robustness of this scheme, which reveals its consistency with the non-convex upper-bounded regularization.

3 Our Approach

3.1 Problem Formulation

Let $\{(x_i, y_i)\}_{i=1}^N$ be a set of N training samples with feature $x_i \in \mathcal{X} \subseteq \mathbb{R}^m$ and label $y_i \in \mathcal{Y}_S = \{1, \ldots, C_S\}$, where $\mathcal{Y}_S$ denotes the set of C_S seen labels. The goal of ZSL is to learn a classifier for a target label set $\mathcal{Y}_T = \{C_S + 1, \ldots, C\}$ disjoint from $\mathcal{Y}_S$. Let $\mathcal{Y} = \mathcal{Y}_S \cup \mathcal{Y}_T$. The side information of labels such as attributes is available to embed the labels into a structured embedding space, represented by an embedding function $\varphi : \mathcal{Y} \to \mathbb{R}^d$. As common in supervised learning, the classifier aims to learn a compatibility score function $F : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ with which the prediction is made:

$$\tilde{y}(x) = \underset{r \in \mathcal{Y}}{\operatorname{argmax}} F(x, r; w), \tag{1}$$

where $F(x, r; w)$ determines the compatibility of the pair (x, r) based on the label embedding $\varphi(r)$ and parameter w. The general max-margin formulation for zero-shot classification with loss L and regularization Ω is given by:

$$\min_w \sum_{i=1}^N \sum_{r \in \mathcal{Y}_S} L(\rho_{ir}) + \nu \Omega(w) \tag{2}$$
$$s.t.\ \forall i, r, \rho_{ir} = \delta F(x_i, r, y_i; w) + \Delta(y_i, r),$$

where $\delta F(x_i, r, y_i; w) = F(x_i, r; w) - F(x_i, y_i; w)$; ρ_{ir} is the score margin of x_i between class r and its ground truth y_i; $\Delta : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ defines the semantic divergence between two labels, as a penalty of predicting r for the true label y_i; $\nu > 0$ is a trade-off hyperparameter. Generally, the loss function $L : \mathbb{R} \to \mathbb{R}^+$ should be convex and monotonically increasing for a small ρ.

3.2 Semantic Correlation Regularization

In Section 3.1, we have considered to learn a max-margin classifier on $\mathcal{Y}_S$. However, due to the divergences of semantics and data distributions between $\mathcal{Y}_S$ and $\mathcal{Y}_T$, the learned models from $\mathcal{Y}_S$ may not necessarily adapt well on $\mathcal{Y}_T$. Thus, we aim to regularize the model based on the source-target semantic correlations for better cross-semantics adaptation.

First, we notice that the semantic correlations of a sample (x_i, y_i) to the classes of $\mathcal{Y}_T$ are not uniform, but specified by $\Delta(y_i, r)$. Hence, it is reasonable to assume that a sample (x_i, y_i) should be assigned a high score $F(x_i, r)$ for class $r \in \mathcal{Y}_T$ if r is semantically close to y_i ($\Delta(y_i, r)$ is low), and a low score $F(x_i, r)$ if r is semantically far from y_i ($\Delta(y_i, r)$ is high). In other words, for a pair of classes $r_1, r_2 \in \mathcal{Y}_T$, $F(x_i, r_1)$ is expected larger than $F(x_i, r_2)$ if $\Delta(y_i, r_1)$ is smaller than $\Delta(y_i, r_2)$, and vice versa. Thus, we consider the following term:

$$\sigma_i(r_1, r_2) \triangleq [\Delta(y_i, r_1) - \Delta(y_i, r_2)]\,[F(x_i, r_1) - F(x_i, r_2)]\,,$$

and penalizes $\max(0, \sigma_i(r_1, r_2))$ that is valid if $\sigma_i(r_1, r_2) > 0$, i.e., the magnitude relationship of the two scores are conflicted with that of their Δ divergences. Then, by the summation over $r_1, r_2 \in \mathcal{Y}_T$, and with a relaxation of a sum of hinge to a hinge of sum for ease of optimization, we have:

$$\begin{aligned}
&\sum\nolimits_{r_1, r_2 \in \mathcal{Y}_T} \max(0, \sigma_i(r_1, r_2)) \\
&\xrightarrow{\text{relax}} \max\Big(0, \sum\nolimits_{r_1, r_2 \in \mathcal{Y}_T} \sigma_i(r_1, r_2)\Big) \\
&= \max\left(0, 2|\mathcal{Y}_T|^2 \left\{\mathrm{E}[\Delta_i^t \odot F_i^t] - \mathrm{E}[\Delta_i^t]\mathrm{E}[F_i^t]\right\}\right) \\
&= 2|\mathcal{Y}_T|^2 \max\left(0, \mathrm{cov}\left[\Delta_i^t, F_i^t\right]\right) \triangleq 2|\mathcal{Y}_T|^2 \max(0, cov_i),
\end{aligned}$$

where $\Delta_i^t, F_i^t \in \mathbb{R}^{|\mathcal{Y}_T|}$ are the stacked vectors of $\Delta(y_i, r)$ and $F(x_i, r)$ for $r \in \mathcal{Y}_T$, respectively; $\mathrm{E}[\cdot]$ is a mean operator and $\mathrm{cov}[\cdot, \cdot]$ is a covariance operator on the elements of vectors.

Based on the above analysis, we propose the Semantic Correlation Regularization (SCR), to encourage the covariance cov_i to be lower than 0 such that the score distributions and the semantic divergence distributions on $\mathcal{Y}_T$ are negatively correlated. Instead of using $\max(0, cov_i)$, we use a smooth surrogate of the hinge function, the logistic function, for derivation convenience. The SCR is defined as:

$$R(\rho_i^t) \triangleq \ln(1 + e^{cov_i}),\ R : \mathbb{R}^{|\mathcal{Y}_T|} \to \mathbb{R}, \tag{3}$$

where $\rho_i^t \in \mathbb{R}^{|\mathcal{Y}_T|}$ is the stacked vector of ρ_{ir} for $r \in \mathcal{Y}_T$. Here, we define the SCR term as a function of ρ_i^t for the convenience of dual derivation in Section 3.4, since

$$\begin{aligned}
cov_i &= \mathrm{cov}\left[\Delta_i^t, F_i^t\right] = \mathrm{cov}\left[\Delta_i^t, F_i^t - F(x_i, y_i)\right] \\
&= \mathrm{cov}\left[\Delta_i^t, \rho_i^t - \Delta_i^t\right] = \mathrm{cov}\left[\Delta_i^t, \rho_i^t\right] - \mathrm{D}[\Delta_i^t].
\end{aligned}$$

With the SCR embedded into the classification objective Eq. (2), the learned model is regularized to be consistent with the source-target semantic correlations, which improves the model adaptation to the target classes.

3.3 BZ-SCR Framework

We formulate the BZ-SCR framework based on the boost learning, the SCR in Section 3.2, and the self-controlled sample selection, for an effective and adaptable zero-shot classifier. Specifically, for effective modeling, we adopt the boosting strategy to formulate F as an ensemble of weak classifiers $\{h_j \in \mathcal{H}\}_{j=1}^K$ in the space of weak models $\mathcal{H}$:

$$F(x, r; w) = \sum_{j=1}^{K} w_j h_j(x, r),\ w \geqslant 0, \tag{4}$$

where each $h_j : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a base score function; w is specified as the weight parameter to be learned.

On the other hand, the samples relations should also be modeled in addition to the labels relations for a robust model transfer. Thus, inspired by the adaptive scheme of [Kumar *et al.*, 2010] that learns a model smoothly from the easy/faithful samples to the hard/confusing ones, we reformulate the objective of Eq. (2) with a self-controlled sample selection procedure. Then, based on the boosted model Eq. (4) and the SCR Eq. (3), we have the formulation of BZ-SCR framework:

$$\min_{w,s} \sum_{i=1}^{N} \left\{ s_i \left[\sum_{r \in \mathcal{Y}_S} L(\rho_{ir}) + \beta R(\rho_i^t) \right] + g(s_i; \lambda) \right\} + \nu\Omega(w), \tag{5}$$

$$s.t.\ \forall i, r, \rho_{ir} = \delta F(x_i, r, y_i) + \Delta(y_i, r); w \geqslant 0; s \in [0, 1]^N,$$

where $s_i \in [0, 1]$ is the weight of sample x_i that indicates its learning "easiness"; $g(\cdot; \lambda) : [0, 1] \to \mathbb{R}$ is the function that specifies how the samples are selected (the reweighting scheme of s) controlled by the parameter $\lambda > 0$; and $\beta > 0$ is a trade-off hyperparameter between the classification loss $L(\cdot)$ for fitness on $\mathcal{Y}_S$ and the SCR $R(\cdot)$ for adaptation to $\mathcal{Y}_T$. Note that in Eq. (5), a weight s_i is assigned to each sample as a measure of its "easiness", which is tuned based on the currently learned confidence (including classification loss and SCR) and the function $g(s_i; \lambda)$ to adaptively select reliable and adaptable samples.

For the convenience of derivation, we specify the loss $L(\cdot)$ as a smooth loss function, the logistic loss, and specify the regularization $\Omega(\cdot)$ as the l_1-norm to impose a sparsity constraint on the model ensemble:

$$L(\rho) \triangleq \ln(1 + e^{\rho});\ \Omega(w) \triangleq \|w\|_1 = \sum\nolimits_j w_j. \tag{6}$$

3.4 Optimization

Following [Pi *et al.*, 2016], we use the alternating optimization to solve Eq. (5). For the optimization of s, we have

$$s_i^* = \operatorname*{argmin}_{s_i} s_i l_i + g(s_i; \lambda),\ s.t.\ s_i \in [0, 1], \tag{7}$$

where $l_i = \sum_{r \in \mathcal{Y}_S} L(\rho_{ir}) + \beta R(\rho_i^t)$ is a composite cost of model fitness on $\mathcal{Y}_S$ and model transfer utility on $\mathcal{Y}_T$. Based on the summarized general properties and candidates of the g function in [Jiang *et al.*, 2014a], we specify g and the corresponding s_i^* as the scheme for mixture weighting due to its

better overall performance in the experiments:

$$g\left(s_i;\lambda,\zeta\right)=-\zeta\ln\left(s_i+\zeta/\lambda\right),\ \lambda,\zeta>0,$$
$$s_i^*=\begin{cases}1, & l_i\leqslant\zeta\lambda/(\zeta+\lambda)\\ 0, & l_i\geqslant\lambda\\ \zeta/l_i-\zeta/\lambda, & \text{otherwise}\end{cases},\qquad(8)$$

which is a mixture of hard 0-1 weighting and soft real-valued weighting, with an extra parameter ζ. Note that s_i^* is monotonically decreasing with l_i and increasing with λ, so as to select easy samples with small losses under the tolerance λ.

For the optimization of w, we have

$$w^*=\underset{w}{\operatorname{argmin}}\sum_i s_i l_i(\rho_i)+\nu\|w\|_1,\qquad(9)$$
$$s.t.\ \forall i,r,\rho_{ir}=\delta F\left(x_i,r,y_i\right)+\Delta(y_i,r);\ w\geqslant 0.$$

To solve w in Eq. (9), we adopt the column generation method [Demiriz *et al.*, 2002] in the dual space of w to handle the potentially infinite candidate weak models in the $\mathcal{H}$ space. We check the dual problem of Eq. (9):

$$\min_Q\sum_{i,r\in\mathcal{Y}_S}H(Q_{ir})+\sum_i\beta R_{s_i}^\star\left(Q_i^t/\beta\right)-J(Q,\Delta),\quad(10)$$
$$s.t.\ \sum_{i,r}Q_{ir}\left(h^{(iy_i)}-h^{(ir)}\right)\leqslant\nu\mathbf{1}_K,$$

where $Q\in\mathbb{R}^{N\times C}$ is the Lagrangian multiplier of the equality constraints of Eq. (9); $Q_i^t\in\mathbb{R}^{|\mathcal{Y}_T|}$ is the stacked vector of Q_{ir} for $r\in\mathcal{Y}_T$; $R_{s_i}^\star:\mathbb{R}^{|\mathcal{Y}_T|}\to\mathbb{R}$ is the Fenchel dual function of $s_iR(\cdot)$; $H(Q_{ir})=Q_{ir}\ln Q_{ir}+(s_i-Q_{ir})\ln(s_i-Q_{ir})$; $J(Q,\Delta)=\sum_{i,r}Q_{ir}\Delta(y_i,r)$; $h^{(ir)}\in\mathbb{R}^K$ is the stacked vector of $h_j(x_i,r)$ for $j\in[1,K]$. The relation between the dual and the primal solution is:

$$Q_{ir}=\begin{cases}\frac{s_i}{1+\exp\left(-\rho_{ir}\right)}, & r\in\mathcal{Y}_S\\ \frac{\Delta(y_i,r)-\mathrm{E}[\Delta_i^t]}{|\mathcal{Y}_T|}\frac{\beta s_i}{1+\exp\left(-cov_i\right)}, & r\in\mathcal{Y}_T\end{cases}.\qquad(11)$$

Please refer to Appendix A for the dual derivation of Eqs. (10) and (11).

Based on the column generation, the set of weak models is augmented by a weak model $\hat{h}$ that most violates the current dual constraint in Eq. (10):

$$\hat{h}=\underset{h\in\mathcal{H}}{\operatorname{argmax}}\sum_{i,r}Q_{ir}\left\{h(x_i,y_i)-h(x_i,r)\right\}.\qquad(12)$$

Then the optimization continues with the new set of weak models, until the violation score (objective value of Eq. (12)) reaches a tolerance threshold.

Eq. (12) indicates that the matrix Q serves as the sample importance for learning a new weak model. From Eq. (11), we see that Q gives high weights to not only the misclassified samples on $\mathcal{Y}_S$ (with large ρ_{ir}) and those not aligned well with their semantic correlations on $\mathcal{Y}_T$ (with large positive cov_i), but also the reliable samples with high weights s_i. These reliable samples are those both discriminative on $\mathcal{Y}_S$ and adapt well on $\mathcal{Y}_T$ in the previous iteration. As a result, the future weak learners will emphasize on samples both insufficiently learned currently and easily learned previously, so as to obtain an effective, adaptable and robust zero-shot classifier for knowledge transfer to the target classes.

Algorithm 1: Boosted Zero-shot classification with Semantic Correlation Regularization (BZ-SCR)

Input : Training set $\{(x_i,y_i)\}_{i=1}^N$; label embeddings $\{\varphi(y)\}_{y\in\mathcal{Y}}$; ν; β; λ_{max}; T_{ES}; $\mu>1$; $\epsilon>0$.
Output : A set of K weak models $\{h_j\in\mathcal{H}\}_{j=1}^K$; w.

1 **Initialize**: $s^{(0)}$; (λ,ζ); $Q\leftarrow s^{(0)}\mathbf{1}_C^T$; $t\leftarrow 0$;
2 **repeat**
3 $\quad t\leftarrow t+1$;
4 $\quad$ Learn a new weak model: solve Eq. (12) to obtain h_t based on Q;
5 $\quad$ Update w: solve Eq. (9) for $w^{(t)}$ based on $s^{(t-1)}$;
6 $\quad$ Update Q: compute Q by Eq. (11) based on $w^{(t)}$ and $s^{(t-1)}$;
7 $\quad$ Update s: compute $s^{(t)}$ by Eq. (8) based on $w^{(t)}$;
$\quad$ **Validation:**
8 $\quad$ Test $\{h_j\}_{j=1}^t$ and $w^{(t)}$ on the validation set, to obtain the classification error rate $err^{(t)}$;
$\quad$ **Annealing:**
9 $\quad$ **if** $\lambda<\lambda_{max}$ **then**
10 $\quad\quad$ $(\lambda,\zeta)\leftarrow(\mu\lambda,\mu\zeta)$;
11 $\quad$ **end**
12 **until** $\sum_{i,r}Q_{ir}\left\{h_t(x_i,y_i)-h_t(x_i,r)\right\}<\nu+\epsilon$ ***or*** $t\geqslant T_{ES}$ ***and*** $err^{(t)}>\min_{1\leqslant\tau\leqslant t-1}err^{(\tau)}$;
13 $K\leftarrow\operatorname{argmin}_\tau err^{(\tau)}$;
Return : $\{h_j\}_{j=1}^K$, $w=w^{(K)}$.

Furthermore, the convergence of Algorithm 1 is guaranteed by the convexity of the two subproblems Eqs. (7) and (9) of the alternating optimization. This leads to the monotonic decreasing of the objective value of Eq. (5) after each iteration. Since the objective is bounded below, such optimization procedure is guaranteed to converge. Figure 1(a)(b) in the experiment empirically shows that the convergence is usually reached within several hundreds of iterations.

4 Experiments

We evaluate the performance of BZ-SCR on the two classic ZSL image datasets, *Animal With Attributes*[1] (*AWA*) and *Caltech-UCSD-Birds-200*[2] (*CUB200*). The comparative methods include Structured Joint Embedding (SJE) [Akata *et al.*, 2015], Attribute Label Embedding (ALE) [Akata *et al.*, 2016], Nonlinear Regression (NR) [Socher *et al.*, 2013], and Canonical Correlation Analysis (CCA) [Hastie *et al.*, 2001].

[1] http://attributes.kyb.tuebingen.mpg.de/

[2] http://www.vision.caltech.edu/visipedia/CUB-200-2011.html

Table 1: Statistics of the datasets

Dataset	*AWA*	*CUB200*
Samples	30475	11788
Classes (Split)	50 (40/10)	200 (150/50)
Feature (dim m)	PCAed VGG (512)	GoogleNet (1024)
Embedding (dim d)	Word2Vec (300)	Attributes (312)

4.1 Datasets and Experimental Settings

We summarize the statistics of the datasets in Table 1. For label embeddings, we use the 300-D GoogleNews Word2Vec[3] representations for *AWA*, and the 312-D annotated attributes for *CUB200*. For the class split, we adopt the default split (40/10 for train+val/test) for *AWA* and the same split as [Akata *et al.*, 2013] (150/50 for train+val/test) for *CUB200*.

We specify for *AWA* the semantic divergence Δ based on the wordnet hierarchy[4]:

$$\Delta(y_1, y_2) = 1 - (SPath(y_1, y_2) + 1)^{-1}, \qquad (13)$$

where $SPath(y_1, y_2)$ is the length of the shortest path between words y_1, y_2 in the wordnet hierarchy. For *CUB200*, we specify Δ based on the cosine similarities between the label embeddings (attributes):

$$\Delta(y_1, y_2) = \frac{1 - cos\left(\varphi(y_1), \varphi(y_2)\right)}{1 - \min_{r_1, r_2 \in \mathcal{Y}} cos\left(\varphi(r_1), \varphi(r_2)\right)}, \qquad (14)$$

where $cos(\cdot, \cdot)$ is the cosine similarity between two vectors, and the denominator is for a normalization to $[0, 1]$.

We specify the weak score function $h(x, y)$ as a rank-one bilinear model, for the convenience of learning a new $\hat{h}$:

$$h(x, y; u, v) = x^T u \cdot v^T \varphi(y), \ s.t. \ \|u\|_2 = \|v\|_2 = 1. \quad (15)$$

With h taking the form of Eq. (15), the learning of $\hat{h}$ (Eq. (12)) is equivalent to a simple SVD problem.

We adopt the strategy in [Jiang *et al.*, 2014b] for the annealing of the parameters (λ, ζ) (Line 9 to 11 in Algorithm 1). At each iteration, we sort the samples in the ascending order of their losses, and set (λ, ζ) based on the proportion of samples to be selected by now. We anneal the proportion of the selected samples instead of the absolute values of (λ, ζ), which is shown more stable in [Jiang *et al.*, 2014b].

We implement a grid search for the tuning of the hyperparameters ν, β, and report the best performances.

4.2 Experimental Results

Table 2 shows the $\bar{\Delta}$ (mean of Δ) and the ER (error rate) performances of BZ-SCR and the comparative methods, where "Boosting" is the boosting-only version of BZ-SCR (fixing all $s = \mathbf{1}_N$). The best results are shown in bold face. Since SJE and ALE also involve a Δ loss in learning, we show their ER results as the better one between two forms of Δ: the same as BZ-SCR (Eq. (13)/(14)) and the uniform 0-1 error. The shown results of BZ-SCR are achieved with $\nu/N = 10^{-4}$, $\beta/N = 0.4$ for *AWA*, and $\nu/N \in \{0.025, 0.05\}$, $\beta/N \in \{0.1, 0.2\}$ for *CUB200*. We see that BZ-SCR has a better overall performance than the comparative methods, since BZ-SCR jointly addresses the issues of model effectiveness and model adaptation for ZSL, based on a self-adaptive SCR-regularized boosting optimization.

[3]http://code.google.com/archive/p/word2vec/

[4]http://stevenloria.com/tutorial-wordnet-textblob/

Table 2: The $\bar{\Delta}$ and error rate performance of each method

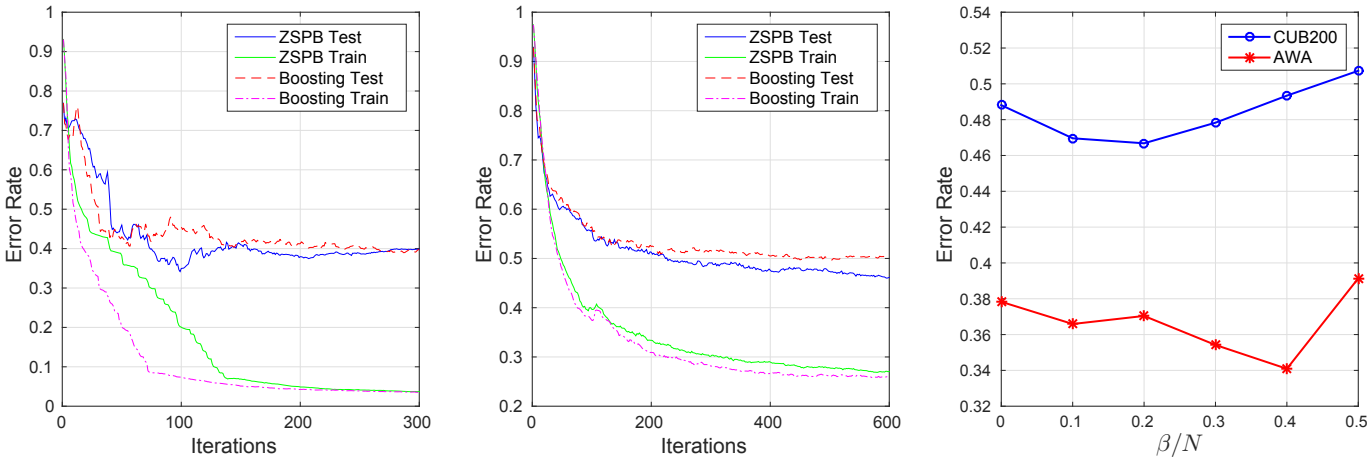
	AWA		*CUB200*	
	$\bar{\Delta}$	ER	$\bar{\Delta}$	ER
BZ-SCR	**0.2907**	**0.3409**	**0.1133**	**0.4664**
Boosting	0.3305	0.3704	0.1277	0.4966
SJE	0.3716	0.4239	0.1286	0.4988
ALE	0.3394	0.3875	0.1401	0.4954
NR	0.4191	0.5100	0.1661	0.6478
CCA	0.3945	0.4702	0.2614	0.6860

Furthermore, we show in Figure 1(a)(b) the change of the error rates on the training and the test set w.r.t. the learning iterations of BZ-SCR and boosting model only, with the same parameter settings. We see that BZ-SCR generally has a smaller gap between the training curves and the test curves, which shows that BZ-SCR achieves a more stable learning process and a more adaptable model transfer. This is due to the smooth learning pace of BZ-SCR based on a self-controlled sample selection from easy reliable samples to hard confusing ones, instead of learning from the whole data batch as boosting does.

We investigate the efficacy of the semantic correlation regularization (SCR) for our model, and show the error rate results w.r.t. different β values in Figure 1(c). From the figure we see that the integration of SCR is beneficial for the model performance. We empirically find that a better overall performance is generally achieved with β/N around $0.1|\mathcal{Y}_S|/|\mathcal{Y}_T|$.

5 Conclusion

In this work, we study zero-shot learning (ZSL) as a transfer learning problem, and focus on its two key aspects, model effectiveness and model adaptation. We propose a unified framework, Boosted Zero-shot classification with Semantic Correlation Regularization (BZ-SCR), that simultaneously addresses these two aspects. We adopt the boosting strategy to learn an effective ensemble classifier, and devise a Semantic Correlation Regularization (SCR) to regularize the model with the inter-class semantic correlations for learning a semantically adaptable zero-shot model. Moreover, we embed a self-controlled sample selection into the framework to model the samples relations for a robust model generalization. By effectively balancing the SCR-regularized boosted model selection and the self-controlled sample selection, the BZ-SCR framework is capable of capturing both discriminative and adaptable feature-to-class semantic alignments, while ensuring the reliability and adaptability of the samples involved in learning. Thus, the proposed framework jointly considers the learning effectiveness, the cross-semantics adaptation, and the model robustness for learning a zero-shot classifier. The experiments on two classic ZSL image datasets verify the superiority of the proposed framework.



(a) *AWA* (b) *CUB200* (c) Effect of SCR

Figure 1: (a)(b) The error rates on the training and test set of BZ-SCR and boosting model w.r.t. the iterations, with $\nu/N = 10^{-4}, \beta/N = 0.4$ for *AWA*, and $\nu/N = 0.025, \beta/N = 0.2$ for *CUB200*. BZ-SCR has a smaller gap between training and test curves. (c) The error rates of BZ-SCR w.r.t. β/N, with $\nu/N = 10^{-4}$ for *AWA* and $\nu/N = 0.025$ for *CUB200*.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grants U1509206, 61472353, and 61672456, in part by the Alibaba-Zhejiang University Joint Institute of Frontier Technologies, and the Fundamental Research Funds for Central Universities in China.

A Dual Derivation of Eqs. (10) and (11)

For Eq. (9), we write its Lagrangian function:

$$\mathcal{L} = \sum_{i,r\in\mathcal{Y}_S} s_i L(\rho_{ir}) + \sum_i \beta s_i R(\rho_i^t) + \nu \mathbf{1}_K^T w - \theta^T w \quad (16)$$
$$- \sum_{i,r} Q_{ir} \{\rho_{ir} + F(x_i, y_i) - F(x_i, r) - \Delta(y_i, r)\},$$

where $Q \in \mathbb{R}^{N\times C}$, $\theta \in \mathbb{R}^K$ ($\theta \geqslant 0$) are the Lagrangian multipliers for the equality constraints of ρ_{ir} and the inequality constraint of $w \geqslant 0$, respectively. Then we have:

$$\frac{\partial \mathcal{L}}{\partial \rho_{ir}} = \begin{cases} s_i L'(\rho_{ir}) - Q_{ir}, & r \in \mathcal{Y}_S \\ \beta s_i L'(cov_i) \frac{\partial cov_i}{\partial \rho_{ir}} - Q_{ir}, & r \in \mathcal{Y}_T \end{cases}.$$

Since $L(\rho) = \ln(1 + e^{\rho})$, $cov_i = \mathrm{cov}(\Delta_i^t, \rho_i^t) - D[\Delta_i^t]$, let $\partial\mathcal{L}/\partial\rho_{ir} = 0$ (KKT conditions), we have:

$$Q_{ir} = \begin{cases} \frac{s_i}{1+\exp(-\rho_{ir})}, & r \in \mathcal{Y}_S \\ \frac{\Delta(y_i,r) - \mathrm{E}[\Delta_i^t]}{|\mathcal{Y}_T|} \frac{\beta s_i}{1+\exp(-cov_i)}, & r \in \mathcal{Y}_T \end{cases}, \quad (17)$$

which is exactly Eq. (11).

Furthermore, the derivative of $\mathcal{L}$ to w is:

$$\frac{\partial \mathcal{L}}{\partial w} = \nu \mathbf{1}_K - \theta - \sum_{ir} Q_{ir} \left(h^{(iy_i)} - h^{(ir)}\right),$$

where $h^{(ir)} \in \mathbb{R}^K$ is the stacked vector of $h_j(x_i, r)$ for $j \in [1, K]$. Let $\partial\mathcal{L}/\partial w = 0$, and based on $\theta \geqslant 0$, we have:

$$\sum_{ir} Q_{ir} \left(h^{(iy_i)} - h^{(ir)}\right) = \nu \mathbf{1}_K - \theta \leqslant \nu \mathbf{1}_K, \quad (18)$$

which is the dual constraint of Q for the dual problem.

By substituting ρ_{ir} ($r \in \mathcal{Y}_S$) from Eq. (17) and θ from Eq. (18) into Eq. (16), we have:

$$\min_{w,\rho} \mathcal{L} = \sum_{i,r\in\mathcal{Y}_S} \{s_i \ln s_i - H(Q_{ir})\} + J(Q, \Delta)$$
$$+ \sum_i \min_{\rho_i^t} \left\{\beta s_i R(\rho_i^t) - \langle Q_i^t, \rho_i^t \rangle\right\},$$

where $H(Q_{ir}) = Q_{ir} \ln Q_{ir} + (s_i - Q_{ir}) \ln (s_i - Q_{ir})$; $J(Q, \Delta) = \sum_{i,r} Q_{ir} \Delta(y_i, r)$; $Q_i^t, \rho_i^t \in \mathbb{R}^{|\mathcal{Y}_T|}$ are the stacked vectors of Q_{ir} and ρ_{ir} for $r \in \mathcal{Y}_T$, respectively. Based on the definition of Fenchel dual function, we have:

$$\min_{\rho_i^t} \left\{\beta s_i R(\rho_i^t) - \langle Q_i^t, \rho_i^t \rangle\right\}$$
$$= -\beta \max_{\rho_i^t} \left\{\langle Q_i^t/\beta, \rho_i^t \rangle - s_i R(\rho_i^t)\right\} = -\beta R_{s_i}^{\star}\left(Q_i^t/\beta\right),$$

where $R_{s_i}^{\star}(\cdot)$ is the Fenchel dual function of $s_i R(\cdot)$.

Finally, based on the Lagrangian duality, the dual problem should be $\max_Q \min_{w,\rho} \mathcal{L}$, which is given by:

$$\max_Q - \sum_{i,r\in\mathcal{Y}_S} H(Q_{ir}) - \sum_i \beta R_{s_i}^{\star}\left(Q_i^t/\beta\right) + J(Q, \Delta),$$

where we omit the constant term. Together with the constraint Eq. (18), the above optimization is equivalent to Eq. (10). Moreover, since $L(\cdot)$ is convex and cov_i is linear to ρ_i^t, the primal objective function is convex and the duality gap is 0.

References

[Akata *et al.*, 2013] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for attribute-based classification. In *Computer Vision and Pattern Recognition*, pages 819–826, 2013.

[Akata *et al.*, 2015] Zeynep Akata, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[Akata *et al.*, 2016] Zeynep Akata, Florent Perronnin, Zaïd Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(7):1425–1438, 2016.

[Demiriz *et al.*, 2002] Ayhan Demiriz, Kristin P Bennett, and John Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1-3):225–254, 2002.

[Freund and Schapire, 1997] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[Frome *et al.*, 2013] Andrea Frome, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[Guo *et al.*, 2016] Yuchen Guo, Guiguang Ding, Xiaoming Jin, and Jianmin Wang. Transductive zero-shot recognition via shared model space learning. In *AAAI Conference on Artificial Intelligence*, pages 3434–3500, 2016.

[Hastie *et al.*, 2001] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. The elements of statistical learning. 2001. *NY Springer*, 2001.

[Jiang *et al.*, 2014a] Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proceedings of the ACM International Conference on Multimedia*, pages 547–556. ACM, 2014.

[Jiang *et al.*, 2014b] Lu Jiang, Deyu Meng, Shoou-I Yu, Zhenzhong Lan, Shiguang Shan, and Alexander Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems*, pages 2078–2086, 2014.

[Kumar *et al.*, 2010] M Pawan Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[Lampert *et al.*, 2009] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition*, 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 Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[Meng and Zhao, 2015] Deyu Meng and Qian Zhao. What objective does self-paced learning indeed optimize? *arXiv preprint arXiv:1511.06049*, 2015.

[Mikolov *et al.*, 2013] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

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

[Palatucci *et al.*, 2009] Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. Zero-shot learning with semantic output codes. In *Advances in Neural Information Processing Systems*, pages 1410–1418, 2009.

[Pi *et al.*, 2016] Te Pi, Xi Li, Zhongfei Zhang, Deyu Meng, Fei Wu, Jun Xiao, and Yueting Zhuang. Self-paced boost learning for classification. In *International Joint Conference on Artificial Intelligence*, pages 1932–1938, 2016.

[Rätsch *et al.*, 2007] Gunnar Rätsch, Manfred K Warmuth, and Karen A Glocer. Boosting algorithms for maximizing the soft margin. In *Advances in Neural Information Processing Systems*, pages 1585–1592, 2007.

[Schapire and Freund, 2012] Robert E Schapire and Yoav Freund. *Boosting: Foundations and algorithms*. MIT press, 2012.

[Socher *et al.*, 2013] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[Wang *et al.*, 2016] Donghui Wang, Yanan Li, Yuetan Lin, and Yueting Zhuang. Relational knowledge transfer for zero-shot learning. In *AAAI Conference on Artificial Intelligence*, pages 2145–2151, 2016.

[Zhang *et al.*, 2015] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. A self-paced multiple-instance learning framework for co-saliency detection. In *International Conference on Computer Vision*, pages 594–602, 2015.

[Zhao *et al.*, 2015] Qian Zhao, Deyu Meng, Lu Jiang, Qi Xie, Zongben Xu, and Alexander G Hauptmann. Self-paced learning for matrix factorization. In *AAAI Conference on Artificial Intelligence*, 2015.

[Zhou, 2012] Zhi-Hua Zhou. *Ensemble methods: foundations and algorithms*. CRC Press, 2012.