Semi-Supervised Zero-Shot Classification with Label Representation Learning

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

Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
daes@ualberta.ca

Abstract

Given the challenge of gathering labeled training data, zero-shot classification, which transfers information from observed classes to recognize unseen classes, has become increasingly popular in the computer vision community. Most existing zero-shot learning methods require a user to first provide a set of semantic visual attributes for each class as side information before applying a two-step prediction procedure that introduces an intermediate attribute prediction problem. In this paper, we propose a novel zero-shot classification approach that automatically learns label embeddings from the input data in a semi-supervised large-margin learning framework. The proposed framework jointly considers multi-class classification over all classes (observed and unseen) and tackles the target prediction problem directly without introducing intermediate prediction problems. It also has the capacity to incorporate semantic label information from different sources when available. To evaluate the proposed approach, we conduct experiments on standard zero-shot data sets. The empirical results show the proposed approach outperforms existing state-of-the-art zero-shot learning methods.

1. Introduction

Visual recognition has made tremendous progress over the last decade. Many reliable and efficient recognition approaches have been developed based on the combination of powerful low-level features such as SIFT [19] and HoG [8], and robust machine learning techniques such as SVMs and Boosting. These recognition systems however typically require a sufficient amount of labeled training images for each class to achieve good classification performance. However, it is very expensive to collect many annotated training instances for every single class given the dramatic increase of the image categories. It is therefore important and desirable to develop classification systems that can significantly reduce the need for labeled training instances from each class.

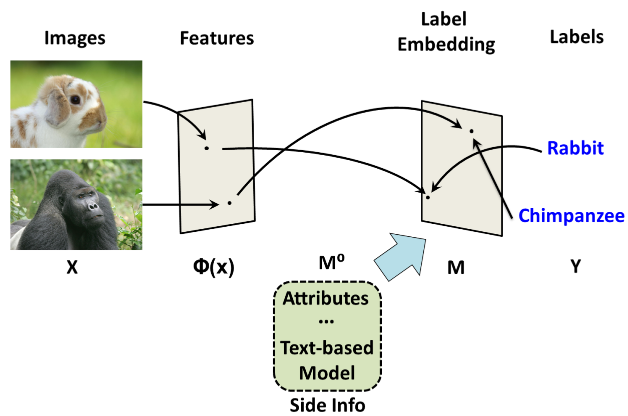


Figure 1: Illustration of the proposed framework. The model can automatically learn the label embeddings M, while side label information, denoted as M^0, is optional.

Zero-shot learning, introduced in [17] and [10] in parallel, offers a compelling solution where unseen classes that do not have any labeled instances are recognized based on knowledge transferred from observed classes with ample labeled instances. The general methodology pursued in the zero-shot classification literature requires the learner to have access to some mid-level semantic label representation that has been defined by human experts or extracted from auxiliary text sources to establish inter-class connections. By exploiting different types of mid-level label representation, current zero-shot learning methods can be categorized into two main groups. One group, attribute-based methods, exploit *attributes* shared between class categories [10, 16, 17, 22], which provides an intermediate label representation. For example, attributes such as "*black*", "*four-leg*" and "*has ears*" can be shared between animal class categories to provide a visually meaningful representation vector for each class label. The main drawback of attribute-based approaches is that they require labor-intensive manual annotation for the class-attribute associations. The other group, text-based methods [9, 13, 24, 25], extract mid-level label representations from large textual corpora such as *WordNet* and *Wikipedia*. By applying natural language processing (NLP) techniques to mine attributes from linguistic resources, these approaches greatly reduce the need for human effort. However, these textual label representations are induced independently from classifier training, hence they are not optimized for the ultimate goal of accurate classification. Overall, existing zero-shot methods still suffer from a number of major drawbacks. First, most approaches apply two independent steps for classification, mapping from the input to the mid-level representation and then from the mid-level representation to the class labels, which violates Vapnik's principle of solving the target problem directly rather than indirectly through intermediate problems [30]. Second, existing methods assume a pre-fixed label representation (or embedding) has been provided by human experts or extracted from linguistic data, regardless of their suitability for the target prediction problem. Third, as training classes and testing classes are disjoint in a two-step classification process, the trained mapping from low-level features to mid-level representations is subject to a projection domain shift problem in the testing phase [14].

In this paper, we propose a novel zero-shot classification approach that can automatically learn the label embeddings from the input data and perform multi-class classification across all classes within a semi-supervised max-margin classification framework. The proposed framework, illustrated in Figure 1, addresses the aforementioned drawbacks of existing zero-shot learning methods in a principled manner. First, the proposed approach does not solve any intermediate problems but rather directly learns model parameters of the target classification function. Second, instead of using fixed label representations, the proposed approach performs label representation learning while training the classification model, which is expected to produce adaptive label embeddings that are more informative for the target classification task. Moreover, the proposed framework can incorporate available label information, such as attribute based label representations and text-induced label representations, as prior knowledge for classification model training. Third, unlike standard zero-shot settings, the proposed semi-supervised framework takes both labeled data from the observed classes and unlabeled data from the unseen classes as input, and jointly learns a multi-class classification model over all classes. In this way, both the label representations and the model parameters are learned consistently across the labeled classes and unlabeled classes, which overcomes possible model shifting problems between the training data and testing data. Furthermore, given the fact that unlabeled data are abundant and easy to collect, this approach provides a mechanism to effectively exploit this readily available resource. To evaluate the performance of the proposed approach, we conduct experiments on standard zero-shot classification data sets. The empirical results demonstrate the superiority of the proposed approach compared to current zero-shot learning methods.

The remainder of the paper is organized as follows. Section 2 first provides a brief review of the related work. The proposed approach is then presented in Section 3. Section 4 provides an experimental evaluation, and finally the paper is concluded in Section 5.

2. Related Work

In this section, we briefly review the related work on zero-shot learning and label embedding learning.

Zero-Shot Learning. Learning classifiers in the absence of labeled data is a challenging problem, and achieving better-than-chance performance requires prior knowledge. Attributes [11] are the most well-known characteristics shared among different objects, which provide an intermediate representation layer between the low-level image features and the semantic labels. Most existing zero-shot classification models exploit attributes in a two-stage classification procedure: given an image, its attributes will be first predicted, then its class label will be predicted as a function of the attributes. In [10, 22, 32], the unseen object classes of images have been described as binary indicator vectors of the attributes to provide intermediate prediction problems. The *Direct Attribute Prediction* (DAP) method developed in [17] takes a similar form but with priors for the classes and attributes, and it uses a MAP prediction for unseen class labels. A topic model variant has been further explored in [33]. As attribute predictions are difficult in practice due to wide image variations, [16] presents a random forest model to account for the unreliability of attribute predictions. In addition to attributes, other external knowledge sources have also been explored for zero-shot classification. For example, [9] uses *Wikipedia* articles to produce the descriptions of labels; [25] utilizes the semantic hierarchy of *WordNet* to mine the parts (attributes) of object categories. Moreover, a zero-shot strategy of directly adapting the classifiers for observed classes to unseen classes has been explored based on the class relationships [20, 21, 24]. In particular, the methods in [21, 24] first compute the class relationships based on the *ImageNet* hierarchy and then estimate the classifier for an unseen label by combining nearest existing classifiers for observed labels; the work in [20] combines classifiers according to the label co-occurrences. Beyond object recognition, zero-shot learning has also been used for other computer vision applications, including *action recognition* [2] and *event detection* [31].

Label Embedding. Different from feature embedding, which provides ways to represent the input images, label embedding, which provides label representations, can be an effective way to share prediction model parameters across classes. [22] applies label embedding in zero-shot classification, but the embedding codes are provided through manual human effort rather than learned from data. To remedy this drawback, *DeViSE* [13] leverages textual data to learn semantic relationships between labels with a neural network model. Similarly, [27] produces continuous semantic word embeddings as label representations using an unsupervised language model. Nevertheless, these works produce label embeddings on textual corpora independent of the target classification task. In consequence, their label embeddings can be uncorrelated with the low-level image features and less informative for the ultimate classification task. [1] presents an approach that can jointly learn class/label embeddings and classifiers in the few-shot setting, but it still relies on side information to provide fixed label embeddings in the zero-shot setting. Unlike these label embedding methods, the approach we propose can learn label embeddings in zero-shot setting with or without side information. More importantly, the approach integrates label embedding learning with classifier training, which guarantees the predictability of the label embeddings from the low-level features and their informativeness for predicting the output class labels.

There are also several zero-shot learning works in the literature that incorporate unlabeled data into the training phase. [23] extends semantic knowledge transfer to the transductive setting by exploiting similarities in the unlabeled data distribution. [14] proposes a transductive multi-view zero-shot learning method, which explores unlabeled data from the unseen classes for projection adaptation, and embeds both low-level feature and multiple semantic representations to rectify the projection shift. [18] integrates semi-supervised classification over the observed classes with unsupervised clustering over the unseen classes in a unified max-margin multi-class classification formulation. However, none of these methods pursue automatic label representation learning from the input data.

3. Proposed Approach

In this section, we present a max-margin semi-supervised approach that trains a multi-class zero-shot classification model defined over both observed and unseen classes. The approach uses a new smooth surrogate loss function which allows the classification model to be efficiently trained over all classes while simultaneously learning adaptive label representations.

We use the following *notation*. For a matrix X, X_i denotes its i-th row vector, and $\|X\|_F$ denotes its Frobenius norm. We will use $\mathbf{1}$ to indicate a column vector with all 1 entries, assuming its length can be determined from the context. $\mathbf{1}_k$ denotes a column vector with all zeros except a single 1 at its k-th entry. I_t denotes an identity matrix with size t, and $\mathbf{0}_{r,c}$ refers to a $r \times c$ matrix with all zero values.

3.1. Semi-Supervised Learning Framework

We consider zero-shot learning in the following multi-class classification setting. Assume one is given a set of t training instances $\mathcal{D} = (X, Y)$ over K classes $\mathcal{Y} = \{1, \cdots, K\}$, where each row of $X \in \mathbb{R}^{t\times d}$ contains a feature vector for an image instance, and each row of $Y \in \{0,1\}^{t\times K}$, when observed, contains an indicator vector that indicates the class membership of the corresponding instance. Without loss of generality, we assume the first t_ℓ instances are labeled instances with class labels in the first K^ℓ *observed classes*, and the remaining t_u instances are unlabeled instances whose labels will belong to the remaining K^u *unseen classes*. Let Y^ℓ be the first t_ℓ rows of Y, which are observed, and Y^u be the last t_u rows of Y, which are latent. Then each row of Y^ℓ contains a single 1 in the first K^ℓ entries, while each row Y^u, once observed, will contain a single 1 in the last K^u entries.

We aim to perform zero-shot classification over the unseen classes by learning a multi-class classification model over all K classes in a semi-supervised manner. In particular, we proposed to perform learning on both the classification model parameters and the latent class labels, discriminatively, by minimizing a regularized classification loss:

$$\min_{Y^u\in\mathcal{Q},W} \sum_{i=1}^{t} \mathcal{L}(f(X_i, W), Y_i) + \frac{\alpha}{2}\|W\|_F^2 + \frac{\rho}{2}\text{tr}(Y^{u\top} L^u Y^u) \quad (1)$$

where $f(\cdot,\cdot)$ is the prediction function with model parameter matrix W, $\mathcal{L}(\cdot,\cdot)$ is a convex loss function, $\mathcal{Q}$ denotes the feasible set for Y^u, and $L^u \in \mathbb{R}^{t_u\times t_u}$ is a Laplacian matrix built over the t_u unlabeled instances such that $L^u = \text{diag}(A\mathbf{1}) - A$ for a similarity matrix $A \in \mathbb{R}^{t_u\times t_u}$. In our experiments, we compute the entries of A as the inverse Euclidean distance between the corresponding unlabeled instance pairs. Laplacian regularization has been typically used in semi-supervised learning scenarios to enforce the smoothness of the prediction values on unlabeled instances with respect to the intrinsic affinity structure of the input data [4]. Here we exploit the Laplacian regularizer to promote the smoothness of our prediction labels from the unseen classes. This framework treats all the K classes equally, and hence avoids the potential model shifting problem between the training data and testing data [14].

Unlike standard semi-supervised learning where labeled instances exist for all the classes, here we do not have any labeled instances for the K^u unseen classes. To facilitate information transfer from observed classes in the labeled data to the unseen classes, we further adopt a label embedding idea into the proposed framework. The intuition is similar to the idea of attribute-based label representations explored in the literature: since image class labels normally provide semantic descriptions of the image content, they can be described with a set of mid-level semantic visual features shared across classes. However, instead of using a pre-fixed label embedding, as in many previous works, we propose to learn label embeddings adaptively from the input data using a discriminative semi-supervised learning approach. In particular, by representing the K classes with a label embedding matrix $M \in \mathbb{R}^{K\times v}$, a semi-supervised co-embedding framework can be obtained:

$$\begin{aligned}\min_{Y^u\in\mathcal{Q},M,W}\quad & \sum_{i=1}^{t}\mathcal{L}(f(X_i,W),Y_iM)+\frac{\alpha}{2}\|W\|_F^2\\ &+\frac{\beta}{2}\|M-M^0\|_F^2+\frac{\rho}{2}\mathrm{tr}(Y^{u\top}L^uY^u)\end{aligned}\qquad (2)$$

where Y_iM maps the label indicator vector of the i-th instance into the embedding vector of its assigned class. Here M^0 is a pre-given prior for the label embedding matrix; when prior knowledge is not available one can simply set $M^0 = \mathbf{0}_{K,v}$. We consider a simple linear prediction function $f(X_i, W) = X_iW$ with model parameter $W \in \mathbb{R}^{d\times v}$, which maps an instance vector into the v-dimensional label embedding space. The parameter matrix W is shared across all K classes, since the classes are represented as v-dimensional embedding vectors in the same space.

By simultaneously learning the label embeddings and the prediction function, the proposed framework in (2) enforces both the predictability of the label embeddings from low-level input features and the informativeness of the embeddings for predicting the output class labels.

3.2. Max-Margin Training Loss

In principle, the proposed framework (2) can accommodate different training losses, such as least-squares loss or max-margin hinge loss. However, the label-embedding-based approach to zero-shot classification faces the same challenge of mid-level prediction unreliability as previous fixed-attribute-based zero-shot methods [16]: since images within a semantic category exhibit significant variation, e.g., different images can contain different subsets of the semantic properties of the given class, the prediction scores of the same label embedding vector can vary a lot within the same category of data. It is therefore more reasonable to use a multi-class large-margin classification model to determine an instance's label by comparing its prediction scores across all classes. In particular, we propose to use a bilinear co-embedding score model that determines the prediction score of the k-the class over an instance $\mathbf{x} \in \mathbb{R}^d$ as

$$s(\mathbf{x}^\top, \mathbf{1}_k^\top) = \mathbf{x}^\top WM^\top\mathbf{1}_k. \qquad (3)$$

The training loss function in the framework (2) above can then be expressed as a multi-class hinge loss [7]:

$$\begin{aligned}&\mathcal{L}(f(X_i,W),Y_iM)\\ &=\max_{k\in\mathcal{Y}}\left(1-Y_k\mathbf{1}_k+s(X_i,\mathbf{1}_k^\top)-s(X_i,Y_i)\right)_+\\ &=\max_{k\in\mathcal{Y}}\left(1-Y_k\mathbf{1}_k+X_iWM^\top\mathbf{1}_k-X_iWM^\top Y_i^\top\right)_+\end{aligned}\qquad (4)$$

where the capped-operator $(\cdot)_+ = \max(\cdot, 0)$. Note that this training loss compares the prediction scores of an instance across all classes, which diminishes the influence of within-category image variation. For example, for an image $\mathbf{x}$ with class label y, even if the image only weakly exhibits the properties of class y and its prediction score $s(\mathbf{x}^\top, \mathbf{1}_y^\top)$ is small, its prediction loss $\mathcal{L}(f(\mathbf{x}, W), M_y)$ can still be small as long as the prediction score $s(\mathbf{x}^\top, \mathbf{1}_y^\top)$ over the correct class is comparatively larger than the prediction scores $s(\mathbf{x}^\top, \mathbf{1}_k^\top)$ over all the other classes $k \in \mathcal{Y}\setminus y$.

Using the large-margin classification model, in the test phase, one can simply predict the label of an given instance $\mathbf{x}$ as the class k^* that maximizes the prediction score:

$$k^* = \arg\max_{k\in\mathcal{Y}}\ s(\mathbf{x}^\top, \mathbf{1}_k^\top) = \arg\max_{k\in\mathcal{Y}}\ (\mathbf{x}^\top WM^\top\mathbf{1}_k). \qquad (5)$$

Since the label matrix $Y^u \in \mathcal{Q}$ over the unlabeled instances is not known and must be learned in the framework (2), we need to specify the feasible set $\mathcal{Q}$ that enforces constraints on the unknown labels. Since it is assumed that Y^u contains labels from the K^u unseen classes, the first K^ℓ columns of Y^u should be zero, while each row of Y^u should contains a single 1 within the last K^u columns. Let $S = [I_{K^\ell}; \mathbf{0}_{K^u,K^\ell}]$ and $\bar{S} = [\mathbf{0}_{K^\ell,K^u}; I_{K^u}]$ be two column selection matrices for Y^u, such that Y^uS contains the first K^ℓ columns of Y^u and $Y^u\bar{S}$ contains the last K^u columns of Y^u. We then impose the following constraints:

$$\mathcal{Q} = \{Y^u \in \{0,1\}^{t_u\times K},\ Y^uS = \mathbf{0}_{t_u,K^\ell},\ Y^u\mathbf{1} = \mathbf{1}\}. \qquad (6)$$

Moreover, since there are no labeled instances for the last K^u classes, the process of recovering Y^u is a clustering process. To avoid degenerate clustering results where most instances are put into a few large clusters while other clusters contain few instances, we further consider a class balance constraint over Y^u: $a\mathbf{1}^\top \le \mathbf{1}^\top(Y^u\bar{S}) \le b\mathbf{1}^\top$, where a and b are user specified constants, $a < b$. This additional constraint enforces that each of the K^u classes obtains at least a and at most b instances from the overall t_u instances, leading to the final constraint set:

$$\mathcal{Q} = \left\{\begin{array}{l}Y^u \in \{0,1\}^{t_u\times K},\ Y^uS = \mathbf{0}_{t_u,K^\ell},\ Y^u\mathbf{1} = \mathbf{1},\\ a\mathbf{1}^\top \le \mathbf{1}^\top(Y^u\bar{S}) \le b\mathbf{1}^\top\end{array}\right\}. \qquad (7)$$

3.3. Smooth Surrogate of Max-Margin Hinge Loss

Multi-class large-margin losses, such as the one introduced in (4), have been popular for discriminatively training multi-class classification models in the literature. However, the non-smoothness of the hinge loss prevents convenient optimization. Although working with the dual training problem [7] can alleviate some of the difficulties with non-smoothness, the dual problem requires a much larger number of optimization variables and sacrifices the convenience of working in the original primal form. Instead, for zero-shot classification framework developed here, we prefer to work in the primal form since it allows the semantic label embeddings to be learned explicitly while allowing side information to be conveniently incorporated (discussed later)—convenient details that are lost in the dual formulation. Therefore, in this section we develop a smooth surrogate loss function that approximates the original max-margin hinge loss (4).

Note that the non-smoothness of the max-margin hinge loss in (4) arises from two maximization operations: the outer maximization over all $k \in \mathcal{Y}$, and the inner capped-operator $(\cdot)_+$. We therefore introduce smooth approximations for each operator.

Proposition 1 *For any vector $\mathbf{z} \in \mathbb{R}^n$, we have that*

$$\max_i \mathbf{z}_i \;\leq\; \tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) \;\leq\; \max_i \mathbf{z}_i + \tau \log n \quad (8)$$

for $\tau > 0$. The middle expression therefore provides a smooth approximation of the maximum function that becomes arbitrarily tight as $\tau \to 0$.

The proof is provided in the supplementary material file.

Proposition 2 *For any scalar $x \in \mathbb{R}$, we have the bounds $(x)_+ \leq \varphi_\tau(x) \leq (x)_+ + \frac{\tau}{4}$ for any $\tau > 0$, where*

$$\varphi_\tau(x) = \begin{cases} 0 & \text{if } \; -\tau \geq x \\ \frac{(x+\tau)^2}{4\tau} & \text{if } \; -\tau < x < \tau \\ x & \text{if } \; x \geq \tau \end{cases} \quad (9)$$

Therefore $\varphi_\tau(\cdot)$ provides a smooth approximation of the capped-operator $(\cdot)_+$ that becomes tight as as $\tau \to 0$.

The proof is provided in the supplementary material file.

By using these upper bound approximations of the two non-smooth operators, one can obtain a principled smooth form of surrogate max-margin loss that approximates the max-margin hinge loss in (4):

$$\begin{aligned} &\widehat{\mathcal{L}}(f(X_i, W), Y_i M) \\ =&\tau \log\left(\sum_{k\in\mathcal{Y}} e^{\varphi_\tau(1 - Y_k \mathbf{1}_k + X_i W M^\top \mathbf{1}_k - X_i W M^\top Y_i^\top)/\tau}\right) \end{aligned} \quad (10)$$

where the $\varphi_\tau(\cdot)$ function is defined in (9). In our experiments, we simply used $\tau = 1$ to obtain a reasonable trade-off between smoothness and approximation tightness; choosing τ too close to zero increases curvature and slows the convergence of any optimization method. With this surrogate loss, the semi-supervised learning problem becomes:

$$\begin{aligned} \min_{Y^u \in \mathcal{Q}, M, W} \quad & \sum_{i=1}^{t} \widehat{\mathcal{L}}(f(X_i, W), Y_i M) + \frac{\alpha}{2}\|W\|_F^2 \\ & + \frac{\beta}{2}\|M - M^0\|_F^2 + \frac{\rho}{2}\text{tr}(Y^{u\top} L^u Y^u). \end{aligned} \quad (11)$$

3.4. Side Information

As noted previously, auxiliary side information about suitable class label representations can also be made available in different forms. For example, intermediate class label representations based on shared lower level attributes have already been obtained via manual human effort for some data sets [10, 22, 32]. Alternatively, label representations can also be extracted from large textual corpora such as *WordNet* and *Wikipedia* using NLP techniques based on the class label phrases [9, 13, 24, 25]. An important aspect of the proposed approach is that these forms of side information can be used as prior knowledge to improve label embedding learning. In particular, one can encode the label representation matrix into the framework, whether obtained via human effort or NLP techniques, by using it as the prior label embedding matrix $M^0 \in \mathbb{R}^{K \times v}$. Even given this prior knowledge, however, the approach still learns an adaptive M from the input data instead of merely fixing it to M^0.

3.5. Training Algorithm

The training problem formulated in (11) is a joint minimization over three variable matrices: the latent label matrix Y^u, the label representation matrix M and the prediction model parameter W. Although the training objective is smooth it is not jointly convex in all three matrices. However, it is convex in each individual variable matrix given the others fixed, therefore we develop an alternating minimization approach to solve the joint training problem (11).

The matrices M and W are first initialized randomly. For Y^u, without side information, we perform k-means clustering over the unlabeled instances with $k = K^u$, then initialize Y^u with the clustering result. With side information, we randomly initialize Y^u. Then we perform alternating updates over the three matrices in the following three steps: First, given the current values for M and Y^u, we solve the convex minimization over W using the LBFGS algorithm. Second, given the current values for W and Y^u, we solve the convex minimization over M using the LBFGS algorithm. Finally, given the current values for W and M, we solve the constrained minimization problem over Y^u. However, with the indicator-based integer constraints over Y^u, the optimization problem remains difficult even with a convex objective, hence we relax the integer constraints $Y^u \in \{0,1\}^{t_u \times K}$ to continuous ones where $0 \leq Y^u \leq 1$. This leads to the following relaxed constraint set $\mathcal{Q}^*$:

$$\mathcal{Q}^* = \left\{ \begin{array}{l} Y^u \geq 0,\; Y^u S = \mathbf{0}_{t_u, K^\ell},\; Y^u \mathbf{1} = \mathbf{1}, \\ a\mathbf{1}^\top \leq \mathbf{1}^\top(Y^u \bar{S}) \leq b\mathbf{1}^\top \end{array} \right\}. \quad (12)$$

Then, for Y^u, we solve the convex minimization subject to the convex constraints $\mathcal{Q}^*$ using a conditional gradient descent algorithm (a.k.a. Frank-Wolfe algorithm) [12]. After obtaining the continuous optimal solution for Y^u, we can round it back into an indicator matrix by setting the largest entry in each row as 1 and other entries as zeros. The overall iterative alternation converges quickly in our experiments.

4. Experiments

In this section, we report our experiments on standard zero-shot classification data sets, comparing the proposed approach to a number of state-of-the-art methods.

4.1. Experimental Setup

Datasets. We conducted experiments on two standard data sets for zero-shot learning. *Animal with Attribute (AwA)* [17] consists of $30,475$ images of 50 animals classes, each containing at least 92 images, paired with a human provided 85-attribute inventory and corresponding class-attribute associations. We follow the commonly agreed experimental protocol in the literature, using the provided split of 40 training and 10 test classes ($24,295$ training, $6,180$ test images). *aPascal-aYahoo (aPaY)* [10] consists of a $12,695$ image subset of the PASCAL VOC 2008 data set across 20 object classes and $2,644$ images collected using the Yahoo image search engine across 12 object classes. Same as in previous work, we perform training on images from PASCAL VOC 2008 and test on images from Yahoo image engine. Additionally, 64 binary attributes that characterize shape, material and the presence of important parts of the visible objects are provided as part of the data set.

Comparison Methods. We compared our approach with three recent zero-shot classification methods: DAP, ALE and MM-ZSL. *Directed Attribute Prediction (DAP)* [17] is a well-known zero-shot learning work that first predicts the value of each attribute for a testing example and then infers the class label according to these predicted attributes. *Attribute Label Embedding (ALE)* [1] treats attribute-based image classification as a label-embedding problem, and maximizes the compatibility between the feature and label embeddings. *Max-Margin Zero-shot Learning (MM-ZSL)* [18] proposes a unified max-margin zero-shot classification formulation in a semi-supervised scheme by involving unlabeled data into the training phase.

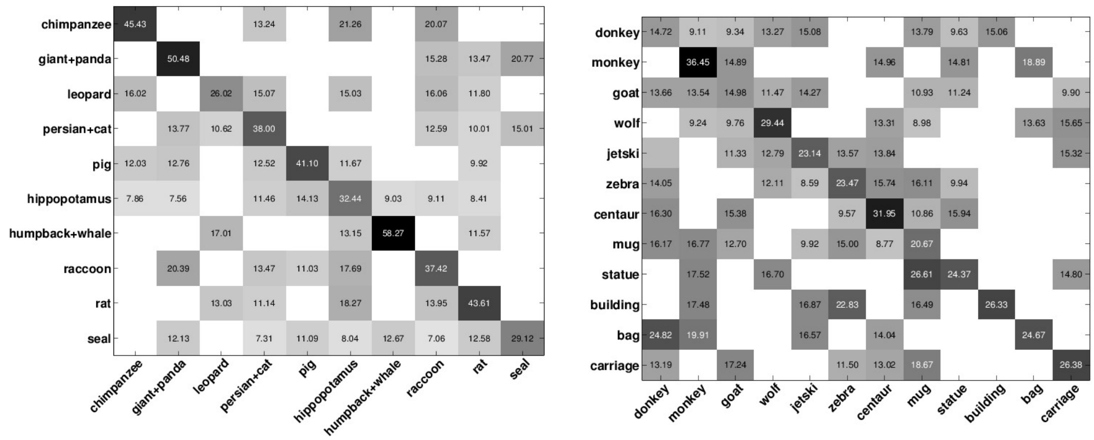
Implementation. For both *AwA* and *aPaY*, we used the pre-computed features within the datasets to represent each image. Specifically, the *AwA* data set provides features such as color histogram, SIFT [19], rgSIFT [29], PHOG [5], SURF [3] and local self-similarity histogram [26]. We first concatenated all features of each image into a vector with length $10,940$, and then performed dimensionality reduction with PCA to reduce the feature vector dimension to $2,000$. The *aPaY* data set provides bag-of-words style features for color, texture, HoG, and Edge. These features are stacked together to form a $9,751$-dimensional feature vector (see [10] for details). We reduced the feature dimension to $1,500$ using PCA. For the *DAP* and *MM-ZSL* methods, we directly used the code provided by the authors. We implemented the *ALE* method with the standard multi-class Structured SVM (SSVM) [28] code.

On each data set, we used all the data from observed classes for training and randomly select 20% of the images from unseen classes for training and used the rest images as testing data. For all the comparison methods, we performed parameter selection on the training data using 3-fold cross validation which separates the training data into a training set and a validation set. The trade-off parameters are selected based on the test performance on the labeled instances from the observed classes in the validation set. This process is repeated three times and the reported test accuracies in this section are averages of three runs. For the proposed method, we set the constants $a = 5$ and set $b = ceil(\frac{t_u}{2})$. The trade-off parameters α, β, and ρ are selected from $\{0.01, 0.1, 1, 10, 100\}$.

4.2. Zero-Shot Classification Results

We evaluated the proposed method and the comparison methods on the two data sets. Since all the comparison methods (DAP, ALE and MM-ZSL) require prior knowledge, we denote our proposed approach with input prior knowledge M^0 as the "*Proposed*" method. In order to demonstrate the usefulness of our automatically learned label embeddings, we have also evaluated our proposed approach without any prior knowledge, *i.e.* $M^0 = 0$, and we denote it as "*Proposed w/o* M^0". The average zero-shot classification results and standard deviations on the unseen classes are reported in Table 1. We can see the proposed method outperforms the other comparisons methods on both data sets with remarkable margins. On *AwA*, the proposed method improves the test accuracy of *MM-ZSL* by 0.5%, of *ALE* and *DAP* by 2.5% and 3.8% respectively. On *aPaY*, the proposed method improves the test accuracy of *MM-ZSL* by almost 5%, and beats *ALE* and *DAP* by 5.5% and 6.3% respectively. More interestingly, our proposed approach can produce compelling results even without any prior knowledge. For example, *Proposed w/o* M^0 produces better results than both *DAP* and *ALE* on *AwA*, and produces competitive results on *aPaY* as well. The label embedding learning can be one factor that leads to this advantage in performance, and the unified semi-supervised learning without separate intermediate problems can be another factor.

From Table 1, we can also see that all methods perform better on *AwA* than on *aPaY*. This is because the connectivity between classes in *AwA* is stronger than in *aPaY*. *AwA* has only animal classes whereas *aPaY* has random object classes. It is easier to learn the common properties of the classes in *AwA* than in *aPaY*. Moreover, the given attributes of *AwA* provide special descriptions tailored for animals, whereas the given semantic attributes of *aPaY* on shape, part, and material are not enough to describe an ob-

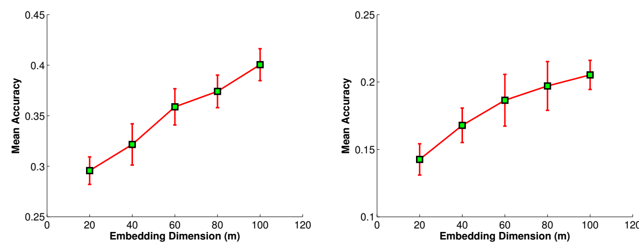


(a) *AwA* (b) *aPaY*

Figure 2: Confusion matrices of the test results on unseen classes for the proposed method on *AwA* and *aPaY*. Diagonal numbers indicate the correct prediction accuracy. Rows corresponds to the ground truth and columns to the predictions.

Table 1: Test accuracy (%) results on the *AwA* and *aPaY* data sets. Each row corresponds to a method. *Proposed* is the the proposed approach with attribute knowledge as side information and *Proposed w/o* M^0 denotes the proposed approach without any prior knowledge ($M^0 = 0$). MM-ZSL, ALE and DAP all use the attribute knowledge.

Method	AwA	aPaY
Proposed	**40.05 ± 2.25**	**24.71 ± 3.19**
Proposed w/o M^0	37.57 ± 2.16	19.14 ± 2.24
MM-ZSL	39.43 ± 2.27	19.77 ± 1.36
ALE	37.49 ± 2.62	19.28 ± 2.27
DAP	36.25 ± 2.57	18.43 ± 2.53



(a) *AwA* (b) *aPaY*

Figure 3: The mean accuracy averaged over all unseen test classes on *AwA* and *aPaY* with different embedding dimensions, *i.e.* v values, using *Proposed w/o* M^0.

ject comprehensively. Hence, based on the attributes, richer and more useful knowledge can be transferred from the seen classes to the unseen classes on *AwA* than on *aPaY*.

The zero-shot classification results of the proposed approach on the two data sets are also visualized in the confusion matrices in Figure 2. In each confusion matrix, the rows correspond to the ground truth and the columns correspond to the predictions. From the confusion matrix for *AwA*, we can observe that for some animal categories our classifier can achieve above 50% accuracy, *e.g. giant panda* (50.48%) and *humpback whale* (58.27%). Given the fact that the classifier is trained without any labeled data from these classes at all, these are quite exciting results. The confusion matrix for *aPaY* also shows impressive results on some categories, for example, *monkey* (36.45%) and *centaur* (31.95%). All these results demonstrate the effectiveness of the proposed approach for zero-shot classification.

4.3. Impact of Label Embedding Dimension

In order to study how the learned label embeddings affect the zero-shot classification performance, we conducted experiments on both data sets with varying embedding dimension, *i.e.* v value in our approach, from $\{20, 40, 60, 80, 100\}$. We investigated our approach without prior knowledge – otherwise v will be determined from the side information M^0. The mean accuracy results averaged over all test unseen classes with different v values are reported in Figure 3. We can see that the same trend is demonstrated on both data sets, that is, the mean test accuracy is increasing with the growing of v value. By comparing the figures here with Table 1, we can see that the performance of the proposed approach without M^0 is competitive to the other methods that use side information when v reaches 100. This suggests that with our proposed approach, automatically learned label embeddings have similar power as the attributes provided by human experts, if not more.

4.4. Impact of Laplacian Regularization

In our proposed approach, we used the Laplacian regularizer to enforce the smoothness of the latent labels on the training instances from the unseen classes. As we do not have any labeled instances from the unseen classes, we expect the Laplacian regularizer can assist the clustering task for the unseen classes and enhance the overall semi-supervised learning. To study the impact of this regularizer on zero-shot classification, we conducted experiments by dropping the smoothness regularization term in our formulation, *i.e.* $tr(Y^{u\top} L^u Y^u)$, by setting $\rho = 0$. We tested both variants of our proposed framework, *Proposed* and *Proposed w/o* M^0, with $\rho = 0$. The zero-shot classification results are reported in Table 2.

Table 2: Test accuracy(%) without Laplacian regularization.

Method	AwA	aPaY
Proposed	**37.54 ± 2.02**	**19.69 ± 1.98**
Proposed w/o M^0	33.75 ± 1.61	17.40 ± 1.66

By comparing the results in Table 1 and Table 2, we can see that by dropping the Laplacian regularization term, the performance of *Proposed* and *Proposed w/o* M^0 degrades on both *AwA* and *aPaY*. For the proposed approach with side information, i.e., *Proposed*, its performance drops about 2.5% and 5% respectively on the two data sets. This indicates that the Laplacian regularizer over the unlabeled instances is a very effective component in our proposed semi-supervised learning framework.

4.5. Exploration of Different Side Information

Finally, since our proposed framework can incorporate any kind of side information encoded as prior label embedding matrix M^0, we investigated the performance of the proposed approach with side information produced from different sources, including the human defined attributes and the label representation vectors produced from large textual corpora using NLP techniques. In particular, we considered *Explicit Semantic Analysis (ESA)* [6], which represents an input word by its appearance record vector over a set of concepts in Wikipedia, and *Word Embedding (WE)* [15], which learns word embeddings with neural networks using an earlier dump of Wikipedia. With each of the semantic tools (ESA and WE), we can transfer a class name into a representation vector which is seen as a row of M^0.

Intuitively, the *attributes* provide richer information than the other two types of external knowledge, since attribute-based label representations are directly provided by human experts based on their interpretations of the label concepts, while *ESA* and *WE* based label representations are extracted automatically from free textual documents. Table 3 presents the zero-shot classification performance achieved by the proposed method with different side information including null information. These results validated our intuition above, as *Proposed+Att* outperforms both *Proposed+WE* and *Proposed+ESA*. Nevertheless, all variants that use side information outperform the variant without side information. This suggests that all these information sources are useful. Moreover, by comparing the results in Table 1 and Table 3, we can see both *Proposed+ESA* and *Proposed+WE* outperform the *ALE* and *DAP* methods which use the attribute information. These results suggest that our proposed semi-supervised framework is effective in exploring different auxiliary sources. Moreover, to verify that learning label representations is useful than fixing them to the prior knowledge M^0, we checked the trade-off parameter β value selected in the experiments. We found $\beta = 1$ instead of any larger values (large β value will push M closer to M^0) were selected from $\{0.01, 0.1, 1, 10, 100\}$ in the parameter selection process of almost all three runs for all the three variants, *Proposed+Att*, *Proposed+WE* and *Proposed+ESA*. This suggests that our framework is really learning useful label embeddings.

Table 3: Experimental results with different side information, including attributes, Word Embedding (WE), Explicit Semantic Analysis (ESA) and null information.

Method	AwA	aPaY
Proposed+Att.	**40.05 ± 2.25**	**24.71 ± 3.19**
Proposed+WE	38.76 ± 1.56	22.29 ± 2.24
Proposed+ESA	38.29 ± 2.04	22.37 ± 2.62
Proposed w/o M^0	37.57 ± 2.16	19.14 ± 2.24

5. Conclusion

In this paper, we proposed a novel semi-supervised approach to address zero-shot learning, which overcomes the limitations of existing zero-shot methods in a principled manner. The proposed approach automatically learns useful label embeddings from the input data and trains a multi-class classification model over all the classes based on a new smooth surrogate training loss. Moreover, it has the capacity to encode prior label representation knowledge from different sources. We conducted extensive experiments to evaluate the proposed approach on two standard zero-shot classification data sets, *Animal with Attributes* and *aPascal-aYahoo (aPaY)*. The results showed that the proposed approach produces superior performance than the existing zero-shot learning methods recently developed in the literature. We have also investigated side information from different resources and showed that the proposed approach can effectively exploit these auxiliary knowledge.

Acknowledgments

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

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *Proceedings of CVPR*, 2013. 3, 6

[2] S. Antol, C. Zitnick, and D. Parikh. Zero-shot learning via visual abstraction. In *Proceedings of ECCV*, 2014. 2

[3] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proceedings of ECCV*, 2006. 6

[4] M. Belkin, P. Niyogi, and V. Sindhwani. On manifold regularization. In *Proceedings of AISTATS*, 2005. 3

[5] A. Bosch, A. Zisserman, and X. Munoz. Representing shape with a spatial pyramid kernel. In *Proceedings of CIVR*, 2007. 6

[6] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, Nov. 2011. 8

[7] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *JMLR*, 2:265–292, 2001. 4

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of CVPR*, 2005. 1

[9] M. Elhoseiny, B. Saleh, and A. Elgammal. Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of ICCV*, 2013. 2, 5

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *Proceedings of CVPR*, 2009. 1, 2, 5, 6

[11] V. Ferrari and A. Zisserman. Learning visual attributes. In *Proceedings of NIPS*, 2007. 2

[12] M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956. 5

[13] A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. DeViSE: A deep visual-semantic embedding model. In *Proceedings of NIPS*, 2013. 2, 3, 5

[14] Y. Fu, T. Hospedales, . Xiang, Z. Fu, and S. Gong. Transductive multi-view embedding for zero-shot recognition and annotation. In *Proceedings of ECCV*, 2014. 2, 3

[15] E. Gabrilovich and S. Markovitch. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI*, 2007. 8

[16] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Proceedings of NIPS*, 2014. 1, 2, 4

[17] H. Lampert, C.and Nickisch and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of CVPR*, 2009. 1, 2, 6

[18] X. Li and Y. Guo. Max-margin zero-shot learning for multi-class classification. In *Proceedings of AISTATS*, 2015. 3, 6

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 1, 6

[20] T. Mensink, E. Gavves, and C. Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of CVPR*, 2014. 2

[21] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka. Metric learning for large scale image classification: Generalizing to new classes at near-zero cost. In *Proceedings of ECCV*, 2012. 2

[22] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In *Proceedings of NIPS*, 2009. 1, 2, 3, 5

[23] M. Rohrbach, S. Ebert, and B. Schiele. Transfer learning in a transductive setting. In *Proceedings of NIPS*, 2013. 3

[24] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of CVPR*, 2011. 2, 5

[25] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where – and why? semantic relatedness for knowledge transfer. In *Proceedings of CVPR*, 2010. 2, 5

[26] E. Shechtman and M. Irani. Matching local self-similarities across images and videos. In *Proceedings of CVPR*, 2007. 6

[27] R. Socher, M. Ganjoo, H. Sridhar, O. Bastani, C. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *Proceedings of NIPS*, 2013. 3

[28] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, Dec. 2005. 6

[29] K. van de Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 32(9):1582–1596, Sept 2010. 6

[30] V. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998. 2

[31] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Proceedings of CVPR*, 2014. 2

[32] F. Yu, L. Cao, R. Feris, J. Smith, and S. Chang. Designing category-level attributes for discriminative visual recognition. In *Proceedings of CVPR*, 2013. 2, 5

[33] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *Proceedings of ECCV*, 2010. 2

Supplentary Material for "Semi-Supervised Zero-Shot Classification with Label Representation Learning"

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

Dale Schuurmans
Department of Computing Science
University of Alberta
Edmonton, AB T6G 2E8, Canada
daes@ualberta.ca

1. Proofs of Propositions

Proposition 1 *For any vector $\mathbf{z} \in \mathbb{R}^n$, we have that*

$$\max_i \mathbf{z}_i \;\le\; \tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) \;\le\; \max_i \mathbf{z}_i + \tau \log n \quad (1)$$

for $\tau > 0$. The middle expression therefore provides a smooth approximation of the maximum function that becomes arbitrarily tight as $\tau \to 0$.

Proof: First, to prove the left inequality in (1), note that it is easy to verify the following

$$\max_i \mathbf{z}_i = \max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} \le \max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} - \tau F^*(\mathbf{p}) \quad (2)$$

where $\Delta = \{\mathbf{p} \in \mathbb{R}^n : \mathbf{p} \ge 0, \mathbf{p}^\top \mathbf{1} = 1\}$ is the probability simplex and $F^*(\mathbf{p}) = \mathbf{p}^\top \log(\mathbf{p})$ is the negative entropy function; the inequality in (2) follows simply because $\tau > 0$ and $F^*(\cdot)$ takes only non-positive values over its domain Δ. Next observe that the maximization problem on the right hand side of (2) corresponds to the definition of the Fenchel conjugate of $\tau F^*(\mathbf{p})$, which is given by $\tau F(\mathbf{z}/\tau)$ for the log-sum-exp function $F(\mathbf{z}/\tau)$ [**?**]; therefore we have

$$\max_{\mathbf{p}\in\Delta} \mathbf{p}^\top \mathbf{z} - \tau F^*(\mathbf{p}) \;= \tau F(\mathbf{z}/\tau) \;= \tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}), \quad (3)$$

establishing the left inequality in (1).

To prove the right inequality in (1), first note that

$$\tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) = c + \tau \log(\sum_{i=1}^{n} e^{(\mathbf{z}_i - c)/\tau}) \quad (4)$$

for any $c \in \mathbb{R}$, via simple algebra, hence

$$\tau \log(\sum_{i=1}^{n} e^{\mathbf{z}_i/\tau}) = \max_i \mathbf{z}_i + \tau \log(\sum_{i=1}^{n} e^{(\mathbf{z}_i - \max_i \mathbf{z}_i)/\tau}).$$

The inequality then follows because $\mathbf{z}_i - \max_i \mathbf{z}_i \le 0$ for all i, hence $e^{(\mathbf{z}_i - \max_i \mathbf{z}_i)/\tau} \le 1$ for all i as long as $\tau > 0$. □

Proposition 2 *For any scalar $x \in \mathbb{R}$, we have the approximation: $(x)_+ \le \varphi_\tau(x) \le (x)_+ + \frac{\tau}{4}$ for any $\tau > 0$, where*

$$\varphi_\tau(x) = \begin{cases} 0 & \textit{if} \quad -\tau \ge x \\ \frac{(x+\tau)^2}{4\tau} & \textit{if} \quad -\tau < x < \tau \\ x & \textit{if} \quad x \ge \tau \end{cases} \quad (5)$$

Therefore $\varphi_\tau(\cdot)$ provides a smooth approximation of the capped-operator $(\cdot)_+$ that becomes tight as as $\tau \to 0$.

Proof: Recall that the capped-operator $(x)_+ = \max(0, x)$ by definition, and note that the following holds

$$\max(0, x) \;=\; \max_{0\le p\le 1} px \;\le\; \max_{0\le p\le 1} px - \tau F^*(p) \quad (6)$$

for a convex regularizer

$$F^*(p) = p^2 - p = -p(1-p);$$

in particular, the inequality in (6) follows because $\tau > 0$ and $F^*(\cdot)$ only takes non-positive values on the domain $0 \le p \le 1$. Next observe that $px - \tau F^*(p) = px + \tau p(1-p)$ is a quadratic concave function of p, hence the maximizer, $\arg\max_{0\le p\le 1} px - \tau F^*(p)$, can be easily recovered as

$$p = \begin{cases} 0 & \text{if} \quad -\tau \ge x \\ \frac{(x+\tau)}{2\tau} & \text{if} \quad -\tau < x < \tau \\ 1 & \text{if} \quad x \ge \tau \end{cases} \quad (7)$$

Plugging this solution back to the maximization objective yields the $\varphi_\tau(\cdot)$ function defined in (5):

$$\max_{0\le p\le 1} px - \tau F^*(p) = \begin{cases} 0 & \text{if} \quad -\tau \ge x \\ \frac{(x+\tau)^2}{4\tau} & \text{if} \quad -\tau < x < \tau \\ x & \text{if} \quad x \ge \tau \end{cases} \quad (8)$$

Equations (6) and (8) and the definition (5) establish that $(x)_+ \leq \varphi_\tau(x)$ for all $x \in \mathbb{R}$ and $\tau > 0$.

To show that $(x)_+ \leq \varphi_\tau(x) + \frac{\tau}{4}$ for all x, note that $(x)_+ = \varphi_\tau(x)$ for $x \leq -\tau$ and $x \geq \tau$ so it remains only to show that the inequality holds for $-\tau < x < \tau$. In this interval, we have that $\varphi_\tau(x) = \frac{(x+\tau)^2}{4\tau}$, so we need to upper bound the gap $g(x) = \varphi_\tau(x) - (x)_+ = \frac{(x+\tau)^2}{4\tau} - (x)_+$. First consider the subinterval $-\tau < x \leq 0$, where the gap is given by $g(x) = \varphi_\tau(x) = \frac{(x+\tau)^2}{4\tau}$. On this subinterval we have $g'(x) > 0$ so the gap value is strictly increasing in x, hence its maximum value is obtained at the rightmost point $x = 0$, yielding $\max_{-\tau<x\leq 0} g(x) = g(0) = \frac{\tau}{4}$. Similarly, on the subinterval $0 \leq x < \tau$ the gap is given by $g(x) = \varphi_\tau(x) - x = \frac{(x-\tau)^2}{4\tau}$. On this subinterval we have $g'(x) < 0$ so the gap value is strictly decreasing in x, hence its maximum value is obtained at the leftmost point $x = 0$, which again yields $\max_{0\leq x<\tau} g(x) = g(0) = \frac{\tau}{4}$. Thus, $\varphi_\tau(x) - (x)_+ \leq \frac{\tau}{4}$ for all x. $\square$