Zero-Shot Classification with Discriminative Semantic Representation Learning

Meng Ye[†]   Yuhong Guo

Computer and Information Sciences, Temple University[†]

School of Computer Science, Carleton University

meng.ye@temple.edu, yuhong.guo@carleton.ca

Abstract

Zero-shot learning, a special case of unsupervised domain adaptation where the source and target domains have disjoint label spaces, has become increasingly popular in the computer vision community. In this paper, we propose a novel zero-shot learning method based on discriminative sparse non-negative matrix factorization. The proposed approach aims to identify a set of common high-level semantic components across the two domains via non-negative sparse matrix factorization, while enforcing the representation vectors of the images in this common component-based space to be discriminatively aligned with the attribute-based label representation vectors. To fully exploit the aligned semantic information contained in the learned representation vectors of the instances, we develop a label propagation based testing procedure to classify the unlabeled instances from the unseen classes in the target domain. We conduct experiments on four standard zero-shot learning image datasets, by comparing the proposed approach to the state-of-the-art zero-shot learning methods. The empirical results demonstrate the efficacy of the proposed approach.

1. Introduction

With the rapid increase of image collections, the class categories involved also expand quickly. The popular standard image classification models such as deep neural networks [15, 28] however require a massive amount of labeled training data from all classes to function properly. To cope with the expensive and sometimes impractical annotation needs required for building prediction systems over the newly appeared classes, zero-shot learning (ZSL), which transfers information from the seen classes with labeled instances to recognize the new classes that have not been seen in the labeled training data, has recently received increasing attention in the research community. ZSL has already been investigated on various computer vision tasks, including image categorization [17, 7, 2, 18, 14], event detection

[32] and action recognition [10, 19, 5].

ZSL can be viewed as a special case of unsupervised domain adaptation, where the source domain (labeled data from the seen classes) and the target domain (unlabeled data from the unseen classes) have completely disjoint label spaces. Hence additional side information has typically been required to build inter-class connections to facilitate the information adaptation across class categories. *Attributes*, which denote high-level visual entities or visual characteristics, have been one most common type of side information exploited in zero shot learning. High-level visual attributes have been exploited in the literature to improve image classification performance [6, 31, 27]. In ZSL, such attributes have been primarily used to provide high-level semantic representations for the class labels. For example, in the *Animal with Attributes (AwA)* dataset [16], each animal class category has been described as a prototype vector of attributes such as *'black', 'stripes', 'four legs'*, etc. By mapping both seen classes and unseen classes into the semantic space based on the same set of attributes, information adaptation can be achieved for zero-shot categorization [23, 16, 17]. Besides attributes, word embeddings have also been used to produce semantic label representations, i.e., prototype label vectors, and build the inter-class connections in some ZSL works [7, 22, 29, 2]. Some other works have also exploited the class taxonomy structures to infer label relationships for ZSL [25, 12, 3].

From the methodology perspective, many ZSL methods have been developed in the literature. They can be roughly grouped into the following three types: (1) visual feature projection methods; (2) semantic similarity matching methods; and (3) sparse coding methods. The *visual feature projection* methods first project an instance (visual feature vector of the image) into the same semantic space as the prototype label vectors, and then assign a prediction label to it by comparing its similarity with all the prototypes of unseen classes [17, 1, 2, 7, 26]. The *semantic similarity matching* methods work in a different way [17, 22, 34]. Their training part remain the same as conventional image classification problem, i.e., a classifier is trained on the seen classes.

When a test instance comes, they use the trained classifier to acquire prediction scores of the instance belonging to each of the seen classes. Finally, the scores are combined with the semantic similarities between each pair of seen and unseen classes to derive the prediction scores on the target unseen classes. The *sparse coding* methods on the other hand are kind of a combination of the previous two types of methods [12, 14]. They transform instances into the label embedding space by exploiting the sparse coding techniques, while assigning each test instance into the unseen class that has the closest semantic embedding. However, most of these ZSL methods fail to exploit the unlabeled instances from the unseen classes in the same way as the labeled instances from the seen classes during the training phase, which makes them prone to domain shift and leads to overfitting to the seen classes.

In this paper we propose a novel zero-shot learning method based on sparse non-negative matrix factorization, which can be categorized into the *sparse coding* group. We treat the labeled data from the seen classes as the source domain and the unlabeled data from the unseen classes as the target domain, and consider ZSL as a special case of unsupervised domain adaptation. To bridge the divergence across domains and enable information transfer across label categories, we perform sparse non-negative matrix factorization on the data in both domains to induce a common dictionary across the two domains that contains components for an *expanded* set of high-level semantic visual attributes, while representing the instances across the two domains in this common semantic dictionary space. Moreover, we simultaneously align the relevant part of the semantic representation vectors of the labeled instances to its attribute-based class label vectors under a discriminative hinge loss. We formulate the overall learning process as a semi-supervised sparse non-negative matrix factorization problem and propose an iterative optimization algorithm based on projected gradient descent to solve it. In addition, we propose to further exploit the learned semantic representations and use a label propagation technique to perform test instance classification in the target domain. Comparing to previous works, the proposed approach simultaneously learns discriminative semantic representations of the instances from both the seen and unseen classes, which can avoid the potential domain shift problem and naturally enable cross-domain information transfer. We conduct experiments on four standard ZSL datasets and the empirical results show that the proposed approach can outperform the state-of-the-art ZSL methods.

2. Related Work

Visual Feature Projection. Many ZSL methods explore semantic relations between seen classes and unseen classes to achieve the goal of automatically categorizing instances into unseen classes. The *visual feature projection* methods first train a projection model based on the training instances and the attribute vectors (or semantic embeddings) of the training classes. Then given a test instance, they project the instance onto the semantic space and assign it into one of the unseen classes by comparing the semantic output with the prototypes of unseen classes. Many different projection strategies have been adopted in the literature, including attribute direct prediction [16, 17], linear mapping [1, 2], convolutional neural networks [7], and simple two layer linear networks [26]. These methods however fail to take the unlabeled instances from the unseen classes into account during the projection function learning process.

Semantic Similarity Matching. Instead of projecting visual feature into the semantic space like the *visual feature projection* approaches, the *semantic similarity matching* methods train a classic classifier on the training data over the seen classes. In the test phase, they first apply the learned classifier to categorize the test instance in terms of the seen classes, and then use the semantic similarity matching between the seen classes and unseen classes to further assign it into the unseen classes. For example, the Indirect Attribute Prediction (IAP) method [16] trains a probabilistic classifier on seen classes. In the test phase, the prediction scores on seen classes are used to predict an attribute distribution, which is further used to predict unseen class distribution. In [22] the authors used a convolutional neural network to directly predict the seen class label of an image, and then use the convex combination of the seen class word embeddings [20] to match with unseen class embeddings. The Semantic Similarity Embedding (SSE) method [34] proposes to represent each unseen class as a distribution/histogram of seen classes. Non-linear embedding functions are learned to map instances into this space to compare with those unseen classes. More recently, the work in [4] introduces phantom classes and proposes to train phantom classifiers as bases for synthesizing classifiers for real classes. These methods fail to exploit the unlabeled data from the target domain in the training process as well.

Sparse Coding. There is a large amount of work in sparse coding and dictionary learning, but very few have tackled ZSL. The work in [12] proposes to represent each category as its *supercategory* plus a combination of *attributes*. Since each category only contains several attributes, they learn a sparse projection matrix to embed seen/unseen categories. They also require the hierarchical category information from WordNet. Another work in [14] proposes to first learn a dictionary on the source data with sparse coding, and then learn the target domain dictionary and target data semantic labels by minimizing the reconstruction error. They also exploited *adaptation regularisation constraint* and *visual-semantic similarity constraint* (VSS). This work is different from ours in that it separately learns two dic-

tionaries for the source and target domains, while we learn a unified semantic dictionary for both domains. Recently, [35] proposes a novel Joint Latent Similarity Embedding (JLSE) method for ZSL. They proposed to learn a joint latent space which is insensitive to noises and can fit source and target instances very well. They have reported great improvements over previous state-of-the-art. We hence will compare our proposed approach to this work.

3. Approach

3.1. Problem Formulation

We consider ZSL in the following unsupervised domain adaptation setting. We have n_s labeled images (X^s, Y^s) from K^s seen classes $\mathcal{Y}^s = \{1, 2, ..., K^s\}$ in the source domain, and a set of n_u unlabeled images X^u from K^u unseen classes $\mathcal{Y}^u = \{K^s+1, \cdots, K\}$ with $K = K^s + K^u$ in the target domain. Here $X^s \in \mathbb{R}^{n_s \times d}$ and $X^u \in \mathbb{R}^{n_u \times d}$ are the input feature matrices, and $Y^s \in \{0, 1\}^{n_s \times K}$ is a class membership indicator matrix, which contains a single 1 in the first K^s columns of each row. We use $X = [X^s; X^u] \in \mathbb{R}^{n \times d}$ to denote all the input images represented as d-dimensional row vectors. We also assume the attribute-based prototype vectors for all the K classes are available in the form of a label representation matrix $M \in \mathbb{R}^{K \times m}$, where the k-th row of M represents the semantic prototype vector for the k-th class and is typically sparse. The problem of ZSL is to transfer information from the source domain to accurately categorize the unlabeled images X^u into the right unseen class from $\mathcal{Y}^u$ in the target domain.

Notation. The following notations are used in the remaining of the paper. We use I_n to denote an identity matrix of size $n \times n$, and use $\mathbf{0}_{r,c}$ to denote a $r \times c$ matrix with all 0s. We use $\mathbf{1}$ to denote a column vector with all 1s, and use $\mathbf{1}_k$ to denote a column zero vector with a single 1 at its k-th entry, assuming the vector length can be determined from the context. We use X_i to denote the i-th row of the matrix X, use X_{ij} to denote the entry of X at its i-th row and j-th column. We use $\|X_i\|$ to denote the Euclidean norm of the vector X_i, and use $\|X\|_F$ and $\|X\|_1$ to denote the Frobenius norm and the entrywise ℓ_1 norm of matrix X respectively. We use $\lambda_{max}(X)$ to denote the largest eigenvalue of X in terms of magnitude.

3.2. Sparse Non-Negative Matrix Factorization

Previous ZSL works typically identify a projection function that maps the input instances into the semantic label prototype space based on the labeled data in the seen classes, while either ignoring the unlabeled data in the unseen classes or handling them in separate steps. A potential problem is that such projection functions identified may overfit the seen classes and do not work well on the unla-

beled data in the unseen classes, which can eventually hurt the ZSL performance. This inspires us to jointly identify the high level latent representations of both the labeled and unlabeled images from the seen and unseen classes in the same semantic space. From the unsupervised domain adaptation perspective, we are also motivated to learn transferable latent representations of the data to address the domain divergence problem [9, 21]. We hence propose to learn latent intermediate representations of the images from both the seen and unseen classes by performing the following unified sparse non-negative matrix factorization (NMF) on X with a common set of non-negative basis components:

$$\min_{Z \geq 0, \Phi \geq 0} \quad \frac{1}{2}\|X - Z\Phi\|_F^2 + \mu\|\Phi\|_1 + \rho\|Z\|_1 \quad (1)$$

where $\Phi \in \mathbb{R}^{a \times d}$ is the component matrix (i.e., dictionary) that contains a basis vectors as its rows; $Z \in \mathbb{R}^{n \times a}$ is the latent representation matrix that contains the coefficient vectors as its rows; and ℓ_1 norm regularizers are used to induce entrywise sparsity in Φ and Z. It has been shown in the literature, sparse NMF can allow one to discover qualitatively better parts-based representations than the regular NMF on images [11]. Here by using sparse NMF, we aim to discover latent representations that can help adapt prediction information across the class boundaries.

3.3. Max-Margin Semantic Alignment with Label Representations

To enable the label information transfer from the seen classes to the unseen classes and achieve effective ZSL, it is desirable that the sparse NMF above can map the images into latent representation vectors in the same semantic space as the label prototype vectors; i.e., the components in the dictionary Φ should be corresponding to the attributes that describe the class labels. Moreover, we will also need to ensure the latent image representations obtained in Z can be discriminative for their class labels. Towards this goal, we propose to align the latent representation vectors of the labeled images in Z with their corresponding label prototype vectors by enforcing each image to have the smallest distance to the prototype vector of its corresponding class label, such that $\|Z_i - Y_i^s M\|^2 \leq \|Z_i - M_k\|^2, \forall k \in \mathcal{Y}^s$.

Restricting the dictionary components to the attributes involved in the class label vectors however requires the attribute set to be broad enough to cover all the contents in the image data X, which is typically not true. Hence we further propose to introduce an additional set of b latent components into the sparse NMF model, such that $a = m + b$. These additional components can capture the background content in the images to help the accurate discovery of the m attribute-based components by minimizing the reconstruction error in Eq.(1). Without loss of generality, we assume the first m components in Φ correspond to the la-

---
Algorithm 1 Projected Subgradient Descent Algorithm

Input: $X_i, Z_i, \Phi, Y^s, M, B$
Initialization: $\tau = \frac{2}{\lambda_{\max}(\Phi\Phi^\top)}$.
Repeat
1. subgradient descent $Z_i = Z_i - \tau \partial g(Z_i)$
2. projection: $Z_i = \max(Z_i, 0)$
Until Converge

---

bel attributes, while the last b components are the additional latent ones. Let $B = [I_m; \mathbf{0}_{d,m}]$ be a column selection matrix for Z, such that ZB contains the first m columns of Z. We then need to align ZB with the corresponding label prototype representation vectors in M. By incorporating this semantic alignment into the sparse NMF model using a max-margin hinge loss, we formulate the following discriminative learning model in the form of semi-supervised sparse NMF:

$$\min_{Z \geq 0, \Phi \geq 0} \frac{1}{2}\|X - Z\Phi\|_F^2 + \gamma \sum_{i=1}^{n_s} \boldsymbol{\xi}_i + \mu\|\Phi\|_1 + \rho\|Z\|_1 \quad (2)$$

where the alignment hinge loss is defined as $\boldsymbol{\xi}_i = \max_{k \in \mathcal{y}^s} \left(\Delta(Y_i\mathbf{1}_k = 0) - \|Z_iB - M_k\|^2 + \|Z_iB - Y_i^sM\|^2\right)_+$; the indicator function $\Delta(c)$ has value 1 if and only if the condition c is true, and the capped operator is defined as $(\cdot)_+ = \max(\cdot, 0)$.

Note Z can be separated into two parts, $Z = [Z^s; Z^u]$, where $Z^s \in \mathbb{R}^{n_s \times a}$ and $Z^u \in \mathbb{R}^{n_u \times a}$ contain the latent representation vectors for the images from the seen classes and unseen classes respectively. The hinge loss in the discriminative semi-supervised NMF model above can push the latent representation vectors Z^sB to align with their corresponding class label representation vectors, and hence Z^uB, expressed in the same space as Z^sB, can be used to determine the class labels of each unlabeled image from X^u in the following way:

$$Y_i^u = \mathbf{1}_{k^*}^\top, \quad \text{with} \ \ k^* = \arg\min_{k \in \mathcal{y}^u} \|Z_i^uB - M_k\|^2. \quad (3)$$

3.4. Optimization Algorithm

To solve the learning problem in Eq.(2), we propose to use an iterative alternating optimization algorithm. In each iteration, it alternatively learns the latent representation matrix Z and the component matrix Φ in two steps.
(i) Learning Z by fixing Φ: The learning problem over Z can be decomposed into a set of n independent subproblems, one for each row of Z. For the i-th row of Z, the subproblem can be written as:

$$\min_{Z_i \geq 0} g(Z_i) = \frac{1}{2}\|X_i - Z_i\Phi\|^2 + \gamma\Delta(i \leq n_s)\boldsymbol{\xi}_i + \rho Z_i\mathbf{1} \quad (4)$$

Since the hinge loss function $\boldsymbol{\xi}_i$ is non-smooth, we use a projected subgradient descent algorithm, presented in Algorithm 1, to perform minimization, where the subgradient can be computed as:

$$\partial g(Z_i) = \begin{cases} Z_i\Phi\Phi^\top - X_i\Phi^\top & \text{if } \Delta(i \leq n_s)\boldsymbol{\xi}_i = 0; \\ \begin{pmatrix} Z_i\Phi\Phi^\top - X_i\Phi^\top + \\ 2\gamma(M_{k^*} - Y_i^sM)B^\top \end{pmatrix} & \text{otherwise}; \end{cases} \quad (5)$$

with $k^* = \arg\max_{k \in \mathcal{y}^s} (\Delta(Y_i\mathbf{1}_k = 0) + \mathcal{D}(i,k))_+$ and $\mathcal{D}(i,k) = \|Z_iB - Y_i^sM\|^2 - \|Z_iB - M_k\|^2$. Let $h(\cdot) = I(\cdot) - \tau\partial g(\cdot)$ be the gradient descent operator in the step 1 of Algorithm 1. We choose the step-size parameter τ to ensure $h(\cdot)$ is non-expansive, i.e., $\|h(Z_i) - h(Z_i')\| \leq \|Z_i - Z_i'\|$ for any feasible Z_i and Z_i', which guarantees the convergence of the algorithm [33]. This leads to $0 < \tau \leq \frac{2}{\lambda_{max}(\Phi\Phi^\top)}$.
(ii) Learning Φ by fixing Z: The minimization over Φ can be written as:

$$\min_{\Phi \geq 0} \ g(\Phi) = \frac{1}{2}\|X - Z\Phi\|_F^2 + \mu\text{tr}(E^\top\Phi) \quad (6)$$

where E is a $a \times d$ matrix with all 1s. We use a projected gradient descent algorithm to solve this linear constrained quadratic programming problem. The projected gradient descent algorithm has the same procedure as the Algorithm 1, except we work on Φ and use the following gradient instead of subgradient:

$$\nabla g(\Phi) = Z^\top Z\Phi - Z^\top X + \mu E \quad (7)$$

The step size τ is chosen in the same principle as stated above to ensure the convergence of the algorithm. In this case, it will be $\tau = \frac{2}{\lambda_{\max}(Z^\top Z)}$.

4. Prediction with Label Propagation

The semantic representations, Z^uB, obtained by our proposed model can be viewed as signatures of the unlabeled instances in the attribute-based label representation space. They contain rich information that can be used beyond the ZSL prediction procedure in Eq.(3), including computing the matching degree scores between the instances and the class labels, and calculating the affinities between instances in the discriminative semantic space. We hence propose to use a label propagation methodology to classify the unlabeled instances into the unseen classes by exploiting such rich information.

We first compute the matching scores of each unlabeled instance with the unseen classes and use these scores as the prediction confidence values to initialize a predicted label matrix $Y \in \mathbb{R}^{n_u \times K^u}$; i.e., we set $Y_{ij} = \kappa(Z_i^uB, M_j^u)$, where M^u denotes a submatrix that contains the last K^u rows of M, and $\kappa(\cdot, \cdot)$ denotes a cosine similarity function.

Next we construct a k-nearest neighbor (k-NN) graph on the n_u unlabeled test instances. We propose to use

the learned representation matrix $Z^u B$ to complement the original feature matrix X^u to represent the n_u instances. Moreover, in order to give equal weights to the two types of features, we first perform PCA dimensionality reduction on X^u to reduce its dimension to the same size as the $Z^u B$ part, and then normalize each row of the dimension reduced X^u and $Z^u B$ separately using their Euclidean norms to obtain the normalized $\hat{X}^u$ and $Z^{\hat{u}} B$ respectively. Finally we use $S = [\hat{X}^u, Z^{\hat{u}} B]$ as the feature matrix for the n_u instances. After computing the squared Euclidean distance between each pair of instances, such that $d(S_i, S_j) = \|S_i - S_j\|^2$, we can construct the k-NN graph by computing the RBF kernel based affinity matrix W in the following way:

$$W_{ij} = \begin{cases} \exp\left(\frac{-d(S_i, S_j)}{2\sigma^2}\right), & \text{if } i \in \text{KNN}(j) \text{ or } j \in \text{KNN}(i) \\ 0, & \text{otherwise} \end{cases}$$

where $\text{KNN}(i)$ denotes the k-nearest neighbors of the i-th instance. Given this affinity matrix W, a normalized Laplacian matrix L can be computed as $L = Q^{-1/2} W Q^{-1/2}$, where Q is a diagonal matrix with $Q_{ii} = \sum_j W_{ij}$. Finally we can perform standard regularized label propagation [8], which provides the following prediction score matrix:

$$Y^* = (I_{n^u} - \alpha L)^{-1} \times Y \tag{8}$$

where $\alpha \in [0, 1]$ is a regularization trade-off parameter. Then the label matrix Y^u can be produced by setting

$$Y_i^u = \mathbf{1}_{k^*}^\top, \quad \text{with } k^* = \underset{k \in \mathcal{Y}^u : j = k - K^s}{\arg\max} Y_{ij}^* \tag{9}$$

5. Experiments

5.1. Experimental Setting

Datasets. We conducted experiments on four standard ZSL datasets: (1) attribute-Pascal-Yahoo (aPY) [6]; (2) Animal with Attribute (AwA) [16]; (3) Caltech-UCSD Bird 200-2011 (CUB) [30]; and (4) SUN-Attribute (SUN) [24]. The aPY dataset contains 12,695 images over 20 classes from the Pascal dataset and 2,644 images over 12 classes collected from Yahoo. Each image in this dataset is labeled with a 64-dim binary vector to denote the attributes. The AwA dataset contains 30,475 images from 50 classes of animals. Each class is associated with a 85-dim attribute vector. CUB is a dataset for fine-grained classification. It contains 11,788 images and 200 class categories. Each image is labeled with a 312-dim vector with continuous values. The SUN-Attribute dataset contains 717 categories with 20 images in each category, which ends up 14,140 images in total, each annotated with a class label and a 102-dim attribute vector. In our experiments we only use class-level attribute vectors if provided, otherwise all the attribute vectors of the images belong to the same class are averaged to

Table 1: Statistics of the four datasets used in the experiments, represented in source/target format. The class splits are provided in the first two datasets. We follow [1] to use the same 50 test classes for the CUB dataset. For the SUN dataset, we use the same 10 test classes as in [13].

Dataset	# instances	# classes	# attributes
aPY	12695/2644	20/12	64
AwA	24295/6180	40/10	85
CUB	8855/2933	150/50	312
SUN	14140/200	707/10	102

serve as the class-level attribute vector. The overall statistic information of the four datasets are summarized in Table 1.

Image Features. Features extracted from Convolutional Neural Networks (CNN) have been well generalized for different kinds of tasks. To take advantage of deep networks for better ZSL performance, we used the same CNN features, 4096-dim vectors extracted from the verydeep-19 network [28], as adopted in previous ZSL works [34].

Parameter Selection. In our model there are three hyper-parameters γ, μ and ρ. We conducted parameter selection using the data in the seen classes for each dataset. Given a dataset with K^s seen classes and K^u target unseen classes, we further split the seen classes into $K_{train} = \lfloor K^s \times \frac{K^s}{K^s + K^u} \rfloor$ training classes and $K_{val} = K^s - K_{train}$ validation classes. We performed ZSL to conduct parameter selection by using the K_{train} classes as the seen classes and the K_{val} classes as the unseen test classes. All the three parameters are selected from the range $\{10^a | a = -3, -2, ..., 2, 3\}$. After parameter selection, we used the selected parameters to perform ZSL with the original seen and unseen classes.

Model Initialization. The iterative training of the proposed model needs to start at a good initialization of the two model parameter matrices, the representation matrix $Z = [Z^s; Z^u]$ and the dictionary matrix Φ. Simple random initialization can lead to very poor solutions. In this work, we adopted an informative initialization procedure. First, we can directly initialize the latent representation of the labeled instances Z^s as the corresponding class prototype vectors; i.e., $Z^s = Y^s M B^\top$. Then, we solve the following matrix factorization problem on the labeled data based on the initial Z^s: $\min_{\Phi^m} \|X^s - Y^s M \Phi^m\|_F^2$, which has a closed-form solution and yields $\Phi^m = (M^\top Y^{s\top} Y^s M + \epsilon I)^{-1} M^\top Y^{s\top} X^s$, where the small constant ϵ is added to avoid numerical problem. This solution Φ^m can be used as the initialization for the first m rows of dictionary Φ, which are corresponding to the m attributes. The rest b rows of

Φ can be randomly initialized. Finally, given the initialized Φ, we can pursue an initialization of Z^u by solving a matrix factorization problem on the unlabeled test data: $\min_{Z^u} \|X^u - Z^u\Phi\|_F^2$, which provides the initialization $Z^u = X^u\Phi^\top(\Phi\Phi^\top + \epsilon I)^{-1}$. Although these initialization values are out of the feasible region, i.e., they do not satisfy the non-negative constraints, they can greatly incorporate information retrieved from the data and labels. Moreover, in the iterative training procedure, Z and Φ will be immediately pushed into the feasible region after one iteration. In our experiments, we found such an initialization procedure can lead to better results than either random initialization or feasible initialization.

5.2. Zero-Shot Classification

We compared the proposed approach with a baseline IAP [16] with CNN features and a few state-of-the-art ZSL methods recently developed in the literature and reported the results in Table 2. We tested two different versions of the proposed discriminative semantic representation learning (DSRL) approach. The first version directly performs prediction by comparing the learned semantic vectors Z^uB with the class label prototype vectors as shown in Eq.(3). We denote this version as *DSRL*. The second version uses the label propagation technique to classify the test instances in the unseen classes. We denote this version as *DSRL-LP*. The number of additional latent semantic components in our model was set to $b = 10$ on all the datasets. For label propagation, we used $k = 10$ to build the k-NN graph, while σ was computed as the mean of distances in the k-NN graph, and α was set to 0.5 for equal preference over the initial prediction and the propagation factor. We repeat each experiment five times with different model initializations and reported the average multi-class classification accuracy results and the standard deviations. Among the comparison methods, UDA-ZSL [14] is most relevant to our proposed approach as it also adopted a sparse coding framework and treated ZSL as unsupervised domain adaptation to bridge the domain shift. However in their model, the source dictionary and target dictionary are learnt separately. They conducted experiments in several different settings. We compared to their results produced in the same experimental setting as other methods, i.e., using CNN features and label attributes. SSE-INT and SSE-ReLU are two variants of the semantic similarity matching method proposed in [34]. JLSE is a more recently developed state-of-the-art method [35], which uses dictionary learning for joint latent similarity embedding.

From Table 2, we can see that the proposed DSRL-LP method consistently outperforms all the comparison methods on all the four datasets, with substantial margins on some datasets. In particular, on the CUB dataset, the proposed DSRL-LP outperforms the best comparison method

JLSE by 15.36%. The CUB dataset is for fine-grained classification which is quite challenging for general ZSL methods. Most attributes in this dataset are designed as 'color' and 'shape' of birds, e.g. 'wing_color', 'back_color', 'eye_color', 'wing_shape', etc. There is a clear correspondence between the attribute vector and the visual appearance of an image. Our discriminative NMF framework can nicely catch such attributes as visual components and achieve a good alignment between the latent representation vectors of the images and the attribute-based class prototype vectors, while effectively transferring prediction information based on the visual knowledge from the source to the target domain. Nevertheless, even if the datasets, e.g., AwA and SUN, contain some attributes that are not designed specifically for visual component identification, the proposed approach still can achieve a consistent mapping from the visual components to the attribute concepts and produce useful semantic instance representations for ZSL; on AwA and SUN, DSRL-LP outperforms the best comparison results by 8.10% and 1.57% respectively.

Between the two variants of the proposed approach, DSRL-LP and DSRL, we can see with label propagation, DSRL-LP can boost the performance substantially on three out of the four datasets, and outperforms DSRL by 9.84%, 6.88% and 3.40% on AwA, CUB and SUN respectively. This suggests that the rich semantic information from Z^u is useful for ZSL. However, we do observe a performance drop on the aPY dataset. To investigate the reason, we produced the confusion matrices for the DSRL prediction results without label propagation on the four datasets, which are presented in Figure 1. We can see that the confusion matrix on the aPY dataset contains more noise than on the other datasets, which suggests large prediction uncertainties. In such case, label propagation can lead to a propagation of the noise and degrade the prediction performance. But it is worth to notice that without label propagation, the proposed DSLR outperforms the best comparison method, JLSE, by 5.94% on aPY.

5.3. Study of the Semantic Representations

As mentioned before, the matrix Z in our model serves as a high-level semantic representation of the instances, whose submatrix ZB should be well aligned with the class prototype vectors. To catch a glimpse of of the quality of the learned Z representation for class separation, we computed the inter-class cosine similarity matrix for the 12 unseen classes of the aPY dataset by using the average of the in-class instance representations in Z^u or Z^uB as the class representation vectors. We also compared to the results obtained by using the average of the original input instances within each class as the class representation vector, and using the attribute-based class prototype vector directly. We visualized the inter-class similarity scores based on these

Table 2: Zero-shot classification results in terms of multi-class classification accuracy on the four ZSL datasets. "-" indicates results not reported.

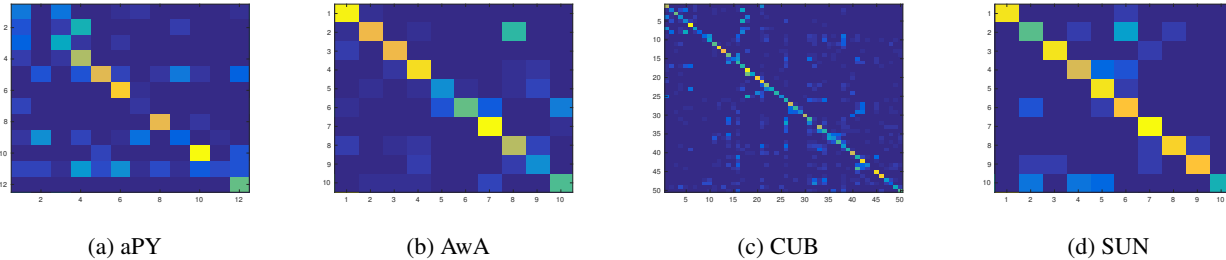| Method	aPY	AwA	CUB	SUN	Avg.
IAP [16]	21.14	49.16	25.43	48.50	36.06
UDA-ZSL [14]	-	73.2	39.5	-	-
SSE-INT [34]	44.15±0.34	71.52±0.79	30.19±0.59	82.17±0.76	57.01±0.62
SSE-ReLU [34]	46.23±0.53	76.33±0.83	30.41±0.20	82.50±1.32	58.87±0.72
JLSE [35]	50.35±2.97	79.12±0.53	41.78±0.52	83.83±0.29	63.77±1.08
DSRL	**56.29±0.44**	77.38±0.06	50.26±0.04	82.00±0.00	66.48±0.14
DSRL-LP	51.29±1.42	**87.22±0.27**	**57.14±0.07**	**85.40±0.22**	**70.26±0.50**

(a) aPY (b) AwA (c) CUB (d) SUN

Figure 1: Visualization of confusion matrix from the DSRL prediction results on the four datasets. Brighter color stands for higher value.
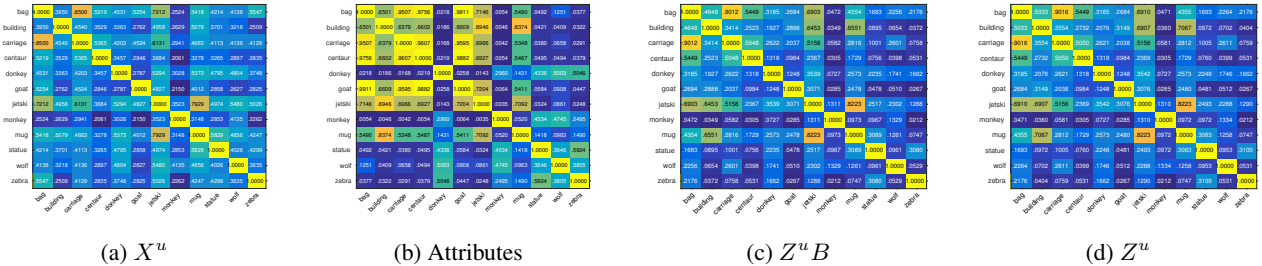

(a) X^u (b) Attributes (c) $Z^u B$ (d) Z^u

Figure 2: Visualization of cosine similarity values between 12 unseen classes in the aPY dataset. Each class is represented as average of the following types of data within the class: (a) Verydeep-19 features X^u, (b) attribute vectors in M, (c) the learned semantic representation $Z^u B$ that are aligned with the label attribute vectors, and (d) the learned semantic representation Z^u.

four different class representations and presented the results in Figure 2. We can see that the two figures produced by using the learned semantic representations, Z^u or $Z^u B$, have less off-diagonal noise than the figure produced by using the original feature X^u, which suggests better class separation ability. We further computed the simple average of the off-diagonal values in each similarity matrix to evaluate its quality, and smaller off-diagonal values indicate better representation and stronger discriminative power. We obtained an average off-diagonal value of 0.4131 from the similarity matrix computed with original features X^u, and obtained much smaller average values of 0.2300 and 0.2265 from the similarity matrices produced by using our learned representations Z^u and $Z^u B$ respectively. Even with the expert-

provided attribute-based class prototypes, the average off-diagonal value on the inter-class similarity matrix is 0.3301, which is larger than our values. These results show that the latent semantic representations of the instances learned by our proposed model have great discriminative power for class separation.

Moreover, one important function of our model in supporting ZSL lies in aligning the semantic representations ZB of the instances with the corresponding attribute-based class prototype vectors. For effective alignment, we expect that each feature column in ZB corresponds to one semantic attribute concept, while the corresponding component in Φ becomes the visual description of the attribute. To verify whether an effective discriminative representation has

Table 3: Similar attribute pairs computed from visual component features: ⟨ a1, a2 ⟩ denote a pair of attributes.

aPY	⟨face, nose⟩, ⟨face, hair⟩, ⟨pot, leaf⟩, ⟨Jet Engine, Propeller⟩
AwA	⟨ocean, plankton⟩, ⟨black, white⟩
CUB	⟨wing_color_blue, upperpart_color_blue⟩ ⟨wing_color_rufous, upperpart_color_rufous⟩ ⟨underpart_color_rufous, breast_color_rufous⟩ ⟨underpart_color_green, breast_color_green⟩
SUN	⟨open area, natural light⟩, ⟨sailing, diving⟩, ⟨hiking, rugged scene⟩, ⟨diving, scary⟩

(a) AwA (b) CUB

Figure 3: Performance with different b values. The worst performance is gained when $b = 0$.

been produced by our model, we hence compare the semantic attributes with their visual description components in Φ. Our intuition is that if two attributes are visible and have similar appearance, we expect their corresponding visual description components in Φ can reflect this closeness, and vice versa. We thus computed the similarity value between each pair of attributes using their visual description vectors in Φ. In Table 3 we present some meaningful pairs we find from the ones with top similarity values. We can see that the visual component description vectors learned in the dictionary Φ can really reflect the semantic concepts of the attributes. For example in the aPY dataset, 'face' and 'nose' are visually similar and conceptually related, so are 'Jet Engine' and 'Propeller'. In SUN dataset, 'sailing' and 'diving' are very similar. It is even more interesting on the fine-grained CUB dataset where most of the attributes are related to color and shape. Here we can see that the learned visual components correctly relate different parts of a bird with the same color to each other; e.g. 'wing_color_rufous' vs. 'upperpart_color_rufous'. This suggests the latent feature representations produced in our model are well aligned with the class attributes.

5.4. Latent Dictionary Components

In the proposed model, in addition to the class attribute components, we have also considered b additional latent components. Our assumption is that the additional latent components can increase the capacity of the model on handling various background noise or content that is not covered by the existing attributes and hence help the accurate discovery of the attribute-based components under the minimization of the NMF reconstruction error. But do we really need these additional latent components? How does the b value affect the performance of the proposed approach? To answer these questions, we conducted experiments on two datasets, AwA and CUB, with a set of different b values from the range of $\{0, 5, 10, 15, 20\}$. For each b value, the experiments are conducted in the same way as before. The results with different b values are presented in Figure 3.
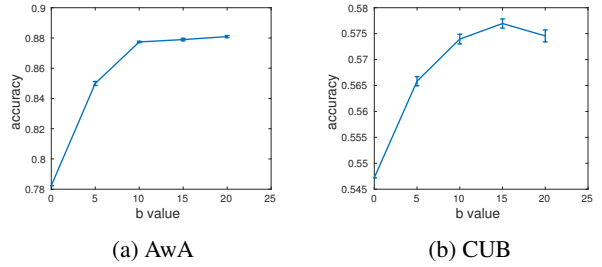
We can see that on both datasets, the worst performance is produced when $b = 0$, i.e., no additional latent components. Moreover, the gap between the test accuracy when $b = 0$ and the best accuracy is substantially large on both datasets. This result clearly suggests that the additional latent components are useful and play a critical role in the proposed ZSL model. The results also intuitively make sense since when $b = 0$, under the reconstruction error, all the background noise will be pushed into the components for the class attributes, which will negatively affect the learning of the attribute dictionary and hence the ZSL.

With the increase of the b value, the performance of the proposed approach improves substantially, especially when b value is small. However, when b value gets too large, e.g, $b = 20$ on CUB, it can harm the accurate learning of the existing attribute components and hence the ZSL performance. Nevertheless, within a reasonable value range, e.g., between 10 and 15, the performance change is very small and the proposed approach achieves great performance.

6. Conclusion

In this paper we proposed a novel zero-shot method to simultaneously learn latent representations for images from both the seen and unseen classes based on a common dictionary that contains basis components for an *expanded* set of semantic attributes. By aligning the relevant part of the semantic representation vectors of the labeled instances to its attribute-based class label vectors under a discriminative max-margin hinge loss, the learned instance representation vectors can naturally reveal their relevance to different class categories. We formulated the overall learning process as a semi-supervised sparse non-negative matrix factorization problem and proposed an iterative optimization algorithm based on projected gradient descent to solve it. We have also adopted the label propagation methodology to fully exploit the semantic instance representation vectors produced by our model and perform test instance classification over the unseen classes. We conducted experiments on four standard ZSL datasets and showed that the proposed approach can outperform the state-of-the-art ZSL methods.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, 2013. 1, 2, 5

[2] Z. Akata, S. Reed, D. Walter, H. Lee, and B. Schiele. Evaluation of output embeddings for fine-grained image classification. In *CVPR*, 2015. 1, 2

[3] Z. Al-Halah and R. Stiefelhagen. How to transfer? zero-shot object recognition via hierarchical transfer of semantic attributes. In *WACV*, 2015. 1

[4] S. Changpinyo, W.-L. Chao, B. Gong, and F. Sha. Synthesized classifiers for zero-shot learning. In *CVPR*, 2016. 2

[5] M. L. Chuang Gan, Y. Yang, Y. Zhuang, and A. G. Hauptmann. Exploring semantic interclass relationships (sir) for zero-shot action recognition. In *AAAI*, 2015. 1

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

[7] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *NIPS*, 2013. 1, 2

[8] Y. Fujiwara and G. Irie. Efficient label propagation. In *ICML*, 2014. 5

[9] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 3

[10] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *ICCV*, 2013. 1

[11] P. Hoyer. Non-negative matrix factorization with sparseness constraints. *JMLR*, 5:1457–1469, 2004. 3

[12] S. J. Hwang and L. Sigal. A unified semantic embedding: Relating taxonomies and attributes. In *NIPS*, 2014. 1, 2

[13] D. Jayaraman and K. Grauman. Zero-shot recognition with unreliable attributes. In *Advances in Neural Information Processing Systems*, pages 3464–3472, 2014. 5

[14] E. Kodirov, T. Xiang, Z. Fu, and S. Gong. Unsupervised domain adaptation for zero-shot learning. In *ICCV*, 2015. 1, 2, 6, 7

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[16] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 5, 6, 7

[17] C. H. Lampert, H. Nickisch, and S. Harmeling. Attribute-based classification for zero-shot visual object categorization. *PAMI*, 36(3):453–465, 2014. 1, 2

[18] J. Lei Ba, K. Swersky, S. Fidler, et al. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *ICCV*, 2015. 1

[19] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 1

[20] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 2

[21] J. Ni, Q. Qiu, and R. Chellappa. Subspace interpolation via dictionary learning for unsupervised domain adaptation. In *CVPR*, 2013. 3

[22] M. Norouzi, T. Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. S. Corrado, and J. Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 1, 2

[23] M. Palatucci, D. Pomerleau, G. E. Hinton, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 1

[24] G. Patterson and J. Hays. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *CVPR*, 2012. 5

[25] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, 2010. 1

[26] B. Romera-Paredes and P. Torr. An embarrassingly simple approach to zero-shot learning. In *ICML*, 2015. 1, 2

[27] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*, 2012. 1

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1, 5

[29] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013. 1

[30] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5

[31] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 1

[32] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014. 1

[33] M. Xiao and Y. Guo. A novel two-step method for cross language representation learning. In *NIPS*, 2013. 4

[34] Z. Zhang and V. Saligrama. Zero-shot learning via semantic similarity embedding. In *ICCV*, 2015. 1, 2, 5, 6, 7

[35] Z. Zhang and V. Saligrama. Zero-shot learning via joint latent similarity embedding. In *CVPR*, 2016. 3, 6, 7