Joint Dictionaries for Zero-Shot Learning

Article · September 2017

CITATIONS
0

READS
55

4 authors, including:

Soheil Kolouri
HRL Laboratories, LLC
30 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Yuri Owechko
HRL Laboratories, LLC
128 PUBLICATIONS **1,033** CITATIONS
SEE PROFILE

Kyungnam Kim
HRL Laboratories, LLC
54 PUBLICATIONS **1,803** CITATIONS
SEE PROFILE

Joint Dictionaries for Zero-Shot Learning

Soheil Kolouri *
HRL Laboratories, LLC
skolouri@hrl.com

Mohammad Rostami *
University of Pennsylvania
mrostami@seas.upenn.edu

Yuri Owechko
HRL Laboratories, LLC
yowechko@hrl.com

Kyungnam Kim
HRL Laboratories, LLC
kkim@hrl.com

Abstract

A classic approach toward zero-shot learning (ZSL) is to map the input domain to a set of semantically meaningful attributes that could be used later on to classify unseen classes of data (e.g. visual data). In this paper, we propose to learn a visual feature dictionary that has semantically meaningful atoms. Such dictionary is learned via joint dictionary learning for the visual domain and the attribute domain, while enforcing the same sparse coding for both dictionaries. Our novel attribute aware formulation provides an algorithmic solution to the domain shift/hubness problem in ZSL. Upon learning the joint dictionaries, images from unseen classes can be mapped into the attribute space by finding the attribute aware joint sparse representation using solely the visual data. We demonstrate that our approach provides superior or comparable performance to that of the state of the art on benchmark datasets.

Introduction

Most classification algorithms require a large pool of manually labeled data to learn the optimal parameters of a classifier. The recent exponential growth of visual data, the growing need for fine-grained multi-label annotations, and consistent emergence of new classes (e.g. new products), however, has rendered manual labeling of data practically infeasible. Transfer learning has been proposed as a remedy to deal with this issue (Lampert, Nickisch, and Harmeling 2009). The idea is to learn on a limited number of classes and then through knowledge transfer, learn how to classify images from the new classes either using only few labeled data points, i.e. few- and one-shot learning (Fei-Fei, Fergus, and Perona 2006), or in the extreme case without any labeled data, i.e. zero-shot learning (ZSL) (Lampert, Nickisch, and Harmeling 2009). These transfer learning approaches address the challenge of annotated data unavailability and open the door towards lifelong learning machines.

To learn target classes with no labeled data, one needs to be able to generalize the relationship between the source data and its labels to the target classes. To address this challenge in ZSL, an intermediate shared space (i.e. the space of semantic attributes) is exploited, which allows for knowledge transfer from labeled classes to the unlabeled classes. The overarching idea in ZSL is that the source and the target classes share common attributes. The semantic attributes (e.g., can fly, is green) are often provided as accessible side information (e.g. verbal description of a class), which uniquely describe classes of data. To achieve ZSL the relationship between seen data and its corresponding attributes are first learned in the training phase. In testing stage, this allows for parsing a target image from an unseen class into its semantic attributes to predict corresponding label.

To clarify the ZSL core idea and the required steps to perform ZSL, consider the following sentence: 'Tardigrades (also known as water bears or moss piglets) are water-dwelling, eight-legged, segmented micro animals'[1]. Given this textual description, one can easily identify the creature in Figure 1, Left as a Tardigrade even though she may have never seen one before. Performing this task requires three capabilities: 1) parsing the textual information into semantic features, so we can describe the class *Tardigrade* as 'bear-like', 'piglet-like', 'water-dwelling', 'eight-legged', 'segmented', and 'microscopic animal', 2) parsing the image into its visual attributes (See Figure 1), and 3) matching the parsed visual features to the parsed textual information which often requires extensive prior knowledge. Recent textual features extracted from large unlabeled text corpora; including *word2vec* (Mikolov et al. 2013) and *glove* (Pennington, Socher, and Manning 2014) enable a learner to efficiently parse textual information. Deep convolutional neural networks (CNNs) (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016; Huang et al. 2017) have revolutionized the field of computer vision and they enable a learner to extract rich visual features from images. An extensive body of work in the field of ZSL is concentrated on modeling the relationship between visual features and semantic attributes (Palatucci et al. 2009; Akata et al. 2013; Socher et al. 2013; Norouzi et al. 2014; Lampert, Nickisch, and Harmeling 2009; Zhang and Saligrama 2015; Ding, Shao, and Fu 2017).

In this paper, we provide a novel approach to model the relationship between the visual features and the textual information. Our specific contributions are:

1. New formulation of ZSL via joint dictionary learning
2. Extending the classic joint dictionary learning formula-

*Equal contribution

[1]Source: Wikipedia

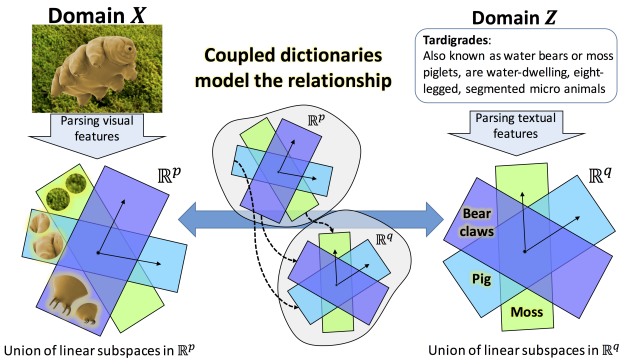

Figure 1: High-level overview of our approach. Left & right: visual and attribute feature extraction and representation using union of subspaces. Middle: constraining the dictionary atoms to be coupled.

tion to an attribute aware formulation that addresses the domain shift/adaptation problem (Kodirov et al. 2015)

3. Demonstrating the benefit of a transductive learning scheme to reduce the hubness phenomenon (Dinu, Lazaridou, and Baroni 2014; Shigeto et al. 2015)

Related Work

ZSL methods often focus on learning the relationship between the visual space and the semantic attribute space. Palatucci et al. (Palatucci et al. 2009) proposed to learn a linear compatibility between the visual space and the semantic attribute space. Lampert et al. (Lampert, Nickisch, and Harmeling 2009) posed the problem as an attribute classification problem and learned individual binary attribute classifiers in the training stage and used the ensemble of classifiers to map visual features to their semantic attributes. Yu and Aloimonos (Yu and Aloimonos 2010) approached the problem from a probabilistic point of view and proposed to use generative models to learn prior distributions for image features with respect to each attribute. More recently, various authors have proposed to embed image features and semantic attributes in a shared metric space (i.e. a latent embedding) (Akata et al. 2013; Romera-Paredes and Torr 2015; Zhang and Saligrama 2015) while forcing the embedded representations for image features and their corresponding semantic attributes to be similar. Akata et al. (Akata et al. 2013), for instance, proposed a model that embeds the image features and the semantic attributes in a common space (i.e. a latent embedding) where the compatibility between them is measured via a bilinear function. Similarly, Romera-Paredes and Torr (Romera-Paredes and Torr 2015) utilized a principled choice of regularizers that enable the authors to derive a simple closed form solution to learn a linear mapping that embeds the image features and the semantic attributes in a low dimensional shared linear subspace. Others have identified the major problems and challenges in ZSL to be the domain shift problem (Kodirov et al. 2015) and the hubness phenomena (Dinu, Lazaridou, and Baroni 2014; Shigeto et al. 2015). In short, the domain shift problem raises from the fact that the distribution of features corresponding to the same attribute for seen and unseen images could be very different (e.g. stripes of tigers versus zebras). The hubness problem, on the other hand, states that there will often be attributes that are similar (have small distance) to vastly different visual features in the embedding space. Various transductive approaches are presented to overcome the hubness problem (Fu et al. 2015; Yu et al. 2017).

The use of sparse dictionaries to model the space of visual features and semantic attributes as union of linear subspaces has been shown to be an effective modeling scheme in recent ZSL papers (Yu et al. 2017; Isele, Rostami, and Eaton 2016; Kodirov et al. 2015; Zhang and Saligrama 2015). Zhang et al. (Zhang and Saligrama 2015) showed that modeling the test image features as sparse linear combination of train image features is beneficial and formulated a ZSL method based on this principal. Using similar ideas, Isele et. al. (Isele, Rostami, and Eaton 2016) used joint dictionary learning to learn a dynamical control system using high level task descriptors in an online lifelong zero-shot reinforcement learning setting. Our JD-ZSL build on similar ideas as in (Yu et al. 2017; Isele, Rostami, and Eaton 2016; Kodirov et al. 2015) and introduce a novel ZSL method based on learning joint sparse dictionaries for the image features and the semantic attributes. At its core, JD-ZSL is equipped with a novel entropy minimization regularizer (Grandvalet and Bengio 2004), which facilitates the solution to the ZSL problem by reducing the domain shift effect. We further show that a transductive approach applied to our attribute aware JD-ZSL formulation provide state-of-the-art or close to state-of-the-art performance on various benchmark datasets. Finally it should be noted that the idea of using joint dictionaries to map data from a given metric space to a second related space was pioneered by Yang et al. (Yang et al. 2010) in super-resolution applications.

Figure 1 captures the gist of our idea. Visual features are extracted via CNNs, left sub-figure, and the semantic attributes are provided via textual feature extractors like word2vec or via human annotations, right sub-figure. Both the visual features and the semantic attributes are assumed to be representable sparsely in a shared union of linear subspaces, left and right sub-figures. The idea here is then to enforce the sparse representation vectors for both domains be equal and thus effectively couple the learned dictionaries for the the visual and the attribute spaces. The intuition from a co-view perspective (Yu et al. 2014) is that both the visual and the attribute features provide information about the same class, and so each can augment the learning of the other. Each underlying class is common to both views, and one can find task embeddings that are consistent for both the visual features and their corresponding attributes. Having learned the coupled dictionaries, zero-shot classification can be performed by mapping images of unseen classes into the attribute space, where classification can be simply done via nearest neighbor or via a more elaborate scheme like label propagation. Given the coupled nature of the learned dictionaries, an image could be mapped to its semantic attributes by first finding the sparse representation with respect

to the visual dictionary, and next the semantic attribute dictionary can be used to recover the attribute vector from the joint sparse representation which could then be used for classification.

Problem Statement and Technical Rational

Consider a visual feature metric space $\mathcal{X}$ of dimension p, an attribute metric space $\mathcal{Z}$ with dimension q, and a class label set $\mathcal{Y}$ with dimension K which ranges over a finite alphabet of size K (images can potentially have multiple memberships to the classes). As an example $\mathcal{X} = \mathbb{R}^p$ for the visual features extracted from a deep CNN and $\mathcal{Z} = \{0,1\}^q$ when a binary code of length q is used to identify the presence/absence of various characteristics in an object (Lampert, Nickisch, and Harmeling 2009). We are given a labeled dataset $\mathcal{D} = \{((\mathbf{x_i}; \mathbf{z_i}), \mathbf{y_i})\}_{\mathbf{i=1}}^{\mathbf{N}}$ of features of seen images and their corresponding semantic attributes, where $\forall i : \mathbf{x_i} \in \mathcal{X}, \mathbf{z_i} \in \mathcal{Z}$, and $\mathbf{y_i} \in \mathcal{Y}$. We are also given the unlabeled attributes of unseen classes $\mathcal{D}' = \{\mathbf{z'_j}\}_{\mathbf{j=1}}^{\mathbf{M}}$ (i.e. we have access to textual information for a wide variety of objects but do not have access to the corresponding visual information). In ZSL the set of seen and unseen classes are disjoint and it is assumed that the semantic attributes are class specific. The goal is then to use $\mathcal{D}$ and $\mathcal{D}'$ to learn the relationship between $\mathcal{X}$ and $\mathcal{Z}$ so when an unseen image (image from an unseen class) is fed to the system, its corresponding attributes and consequently its label could be predicted. Finally, we assume that $\psi : \mathcal{Z} \rightarrow \mathcal{Y}$ is the mapping between the attribute space and the label space and ψ is a known linear mapping, $\mathbf{y} = \psi(\mathbf{z}) = \mathbf{V}\mathbf{z}$.

To further clarify the problem, consider an instance of ZSL in which features extracted from images of horses and tigers are included in seen visual features $X = [\mathbf{x_1}, ..., \mathbf{x_N}]$, where $\mathbf{x_i} \in \mathcal{X}$, but X does not contain features from images containing zebras. On the other hand, the semantic attributes contain information of all seen $Z = [\mathbf{z_1}, ..., \mathbf{z_N}]$ for $\mathbf{z_i} \in \mathcal{Z}$ and unseen $Z' = [\mathbf{z'_1}, ..., \mathbf{z'_M}]$ for $\mathbf{z'_j} \in \mathcal{Z}$ classes including the zebras. Intuitively, by learning the relationship between the image features and the attributes "has hooves", "has mane", and "has stripes" from the seen images, we must be able to assign an image of a zebra to its corresponding attribute, while we have never seen a zebra before. More formally, we want to learn the mapping $\phi : \mathcal{X} \rightarrow \mathcal{Z}$ which relates the visual space and the attribute space. Having learned this mapping, for an unseen image one can recover the corresponding attribute vector using the image features and then classify the image using the mapping $\mathbf{y} = (\psi \circ \phi)(\mathbf{x})$, where '$\circ$' represents function composition.

Technical Rational

For the rest of our discussion we assume that $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Z} = \mathbb{R}^q$, and $\mathcal{Y} = \mathbb{R}^K$. The simplest ZSL approach is to assume that the mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ is linear, $\phi(\mathbf{x}) = W^T\mathbf{x}$ where $W \in \mathbb{R}^{p \times q}$, and then minimize the regression error $\frac{1}{N}\sum_i \|W^T\mathbf{x}_i - \mathbf{z}_i\|_2^2$ to learn W. Despite existence of a closed form solution for W, the solution contains the inverse of the covariance matrix of X, $(\frac{1}{N}\sum_i (x_i x_i^T))^{-1}$, which requires a large number of data points for accurate estimation. To overcome this problem, various regularizations are considered for W. Decomposition of W as $W = P\Lambda Q$, where $P \in \mathbb{R}^{p \times l}$, $\Lambda \in \mathbb{R}^{l \times l}$, $Q \in \mathbb{R}^{l \times q}$, and $l < min(p, q)$ can also be helpful. Intuitively, P is a right linear operator that projects $\mathbf{x}$'s into a shared low dimensional subspace, Q is a left linear operator that projects $\mathbf{z}$ into the same shared subspace, and Λ provides a bi-linear similarity measure in the shared subspace. The regression problem then can be transformed into maximizing $\frac{1}{N}\sum_i \mathbf{x}_i^T P\Lambda Q\mathbf{z}_i$, which is a weighted correlation between the embedded $\mathbf{x}$'s and $\mathbf{z}$'s. This is the essence of many ZSL techniques including Akata et al. (Akata et al. 2013) and Romera-Paredes et al.(Romera-Paredes and Torr 2015). This technique can be extended to nonlinear mappings using kernel methods. However, the choice of kernels remains a challenge.

On the other side of the spectrum, the mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ can be chosen to be highly nonlinear, as in deep neural networks (DNN). Let a DNN be denoted as $\phi(.;\theta)$, where θ represents the parameters of the network (i.e. synaptic weights and biases). ZSL can then be addressed by minimizing $\frac{1}{N}\sum_i \|\phi(\mathbf{x}_i;\theta) - \mathbf{z}_i\|_2^2$ with respect to θ. Alternatively, one can nonlinearly embed $\mathbf{x}$'s and $\mathbf{z}$'s in a shared metric space via deep nets, $f(\mathbf{x};\theta_x) : \mathbb{R}^p \rightarrow \mathbb{R}^l$ and $g(\mathbf{z};\theta_z) : \mathbb{R}^q \rightarrow \mathbb{R}^l$, and maximize their similarity measure in the embedded space, $\frac{1}{N}\sum_i f(\mathbf{x}_i;\theta_x)^T g(\mathbf{z}_i;\theta_z)$, as in (Lei Ba et al. 2015). Nonlinear methods are computationally expensive, require a large training dataset, and can easily overfit to the training data. On the other hand, linear ZSL algorithms are efficient, easy to train, and generalizable but they are often outperformed by nonlinear methods. As a compromise, we model nonlinearities in data distributions as union of linear subspaces with coupled dictionaries. By jointly learning the visual and attribute dictionaries, we effectively model the relationship between the metric spaces. This allows a nonlinear scheme with a computational complexity comparable to linear techniques.

Zero Shot Learning using Joint Dictionaries

Joint dictionary learning has been proposed to couple related features from two metric spaces (Yang et al. 2010; Shekhar et al. 2014). Yang et al. (Yang et al. 2010) proposed the approach to tackle the problem of image super-resolution and Shekhar et al. (Shekhar et al. 2014) used joint dictionary learning for multimodal biometrics recognition. Following a similar framework, the gist of our approach is to learn the mapping $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ through two dictionaries, $D_x \in \mathbb{R}^{p \times r}$ and $D_z \in \mathbb{R}^{q \times r}$ that model X and $[Z, Z']$, respectively, where $r > max(p, q)$. The goal is to find a shared sparse representation (i.e. sparse code) $\mathbf{a}_i$ for $\mathbf{x}_i$ and $\mathbf{z}_i$, such that $\mathbf{x}_i = D_x\mathbf{a}_i$ and $\mathbf{z}_i = D_z\mathbf{a}_i$. Below we describe the training and testing phases of our proposed method.

Training phase

The standard dictionary learning is based on minimizing the empirical average estimation error $\frac{1}{N}\|X - D_x A\|_F^2$ on a given training set X, where ℓ_1 regularization on A enforces

sparsity:

$$D_x^*, A^* = \underset{D_x, A}{\operatorname{argmin}} \frac{1}{N}\|X - D_x A\|_F^2 + \lambda\|A\|_1 \quad \text{s.t. } \|D_x^{[i]}\|_2^2 \leq 1. \tag{1}$$

Here λ is the regularization parameter, which controls the sparsity of A, and $D_x^{[i]}$ is the i'th column of D_x. Alternatively, following the Lagrange multiplier technique, the Frobenius norm of D_x could be used as a regularizer in place of the costraint.

In our joint dictionary learning framework, we aim to learn D_x and D_z such that they share the sparse coefficients A to represent the seen visual features X and their corresponding attributes Z, respectively. An important twist here is that the attribute dictionary, D_z, is also required to sparsify the semantic attributes of other (unseen) classes, Z'. To obtain such coupled dictionaries we propose the following optimization,

$$\begin{aligned}\underset{D_x, A, D_z, B}{\operatorname{argmin}} \{&\frac{1}{Np}(\|X - D_x A\|_F^2 + \frac{p\lambda}{r}\|A\|_1) + \\ &\frac{1}{Nq}\|Z - D_z A\|_F^2 + \frac{1}{Mq}(\|Z' - D_z B\|_F^2 + \\ &\frac{q\lambda}{r}\|B\|_1)\} \quad \text{s.t.:} \|D_x^{[i]}\|_2^2 \leq 1,\ \|D_z^{[i]}\|_2^2 \leq 1\end{aligned} \tag{2}$$

The above formulation combines the dictionary learning problem for X and Z by coupling them via A, and also enforces D_z to be a sparsifying dictionary (i.e. a good model) for Z'. The optimization in Eq (2), while convex in each individual term, is highly nonconvex in all variables. Following the approach proposed in (Yang et al. 2012) we use an Expectation Maximization (EM) like alternation to update dictionaries D_x and D_z. To do so, we rewrite the optimization problem into the following two steps:

1. For a fixed D_x update D_z via the following optimization:

$$\begin{aligned}&\min_{D_z, B} \frac{1}{Mq}(\|Z' - D_z B\|_F^2 + \frac{q\lambda}{r}\|B\|_1) + \\ &\quad \frac{1}{Nq}\|Z - D_z A^*\|_F^2 \\ &\text{s.t. } A^* = \underset{A}{\operatorname{argmin}} \frac{1}{p}\|X - D_x A\|_F^2 + \frac{\lambda}{r}\|A\|_1, \\ &\|D_z^{[i]}\|_2^2 \leq 1\end{aligned} \tag{3}$$

   A is found using a Lasso optimization problem, and FISTA (Beck and Teboulle 2009) is used to update D_z and B.

2. For a fixed D_z update D_x via:

$$\begin{aligned}&\min_{D_x} \ \|X - D_x A^*\|_F^2 \\ &\text{s.t. } A^* = \underset{A}{\operatorname{argmin}} \frac{1}{q}\|Z - D_z A\|_F^2 + \frac{\lambda}{r}\|A\|_1, \\ &\quad \|D_x^{[i]}\|_2^2 \leq 1,\end{aligned} \tag{4}$$

   which involves a Lasso optimization together with a simple regression with a close form solution.

Zero-Shot Prediction of Unseen Attributes

In the testing phase we are only given the extracted features from unseen images, $X' = [\mathbf{x}'_1, ..., \mathbf{x}'_l] \in \mathbb{R}^{p \times l}$ and the goal is to predict their corresponding semantic attributes. Here we introduce a progression of methods, which clarifies the logic behind our method, and enables us to efficiently predict the semantic attributes of the unseen images based on the learned dictionaries in the training phase.

Attribute Agnostic Prediction The attribute agnostic (AAg) formulation, is the naive way of predicting semantic attributes from an unseen image $\mathbf{x}'_i$. In the AAg formulation, we first find the sparse representation $\boldsymbol{\alpha}_i$ of the unseen image $\mathbf{x}'_i$ with respect to the learned dictionary D_x by solving the following Lasso problem,

$$\boldsymbol{\alpha}_i = \operatorname{argmin}_{\mathbf{a}} \frac{1}{p}\|\mathbf{x}_i - D_x \mathbf{a}\|_2^2 + \frac{\lambda}{r}\|\mathbf{a}\|_1. \tag{5}$$

Here, one can simply use $\boldsymbol{\alpha}_i$ and compare it to the sparse codes of the unseen attributes, $\mathbf{b}_j$. In our experiments, however, we found that this approach is not suitable in our JD-ZSL setting as the dictionaries could have redundant atoms that cause two similar image features or attributes to have different sparse codes. Instead, we do the comparison in the attribute space and predict the corresponding attribute via $\hat{z}_i = D_z \boldsymbol{\alpha}_i$. In the attribute-agnostic formulation, the sparse coefficients are calculated without any information from the attribute space. Not using the information from the attribute space would lead to the domain shift problem, in the sense that there is no guarantee that $\boldsymbol{\alpha}_i$ would reconstruct a meaningful attribute in $\mathcal{Z}$. In other words, $\hat{z}_i = D_z \boldsymbol{\alpha}_i$ could be far from the unseen attributes, $\mathbf{z}'_m$, and therefore could not be assigned to any known attribute with high confidence. To alleviate this problem we progress to an extended solution, which we denote as the Attribute Aware (AAw) prediction.

Attribute Aware Prediction In the attribute-aware (AAw) formulation we would like to find the sparse representation $\boldsymbol{\alpha}_i$ to not only approximate the input visual feature, $\mathbf{x}'_i \approx D_x \boldsymbol{\alpha}_i$, but also provide an attribute prediction, $\hat{z}_i = D_z \boldsymbol{\alpha}_i$, that is well resolved in the attribute space and does not suffer from the domain shift problem. Note that, ideally $\hat{\mathbf{z}}_i = \mathbf{z}'_m$ for some $m \in \{1, ..., M\}$. To achieve this we define the soft assignment of $\hat{\mathbf{z}}_i$ to $\mathbf{z}'_m$, denoted by p_m, using the Student's t-distribution as a kernel to measure similarity between $\hat{\mathbf{z}}_i = D_z \boldsymbol{\alpha}_i$ and $\mathbf{z}'_m$,

$$p_m(\boldsymbol{\alpha}_i) = \frac{(1 + \frac{\|D_z \boldsymbol{\alpha}_i - \mathbf{z}'_m\|_2^2}{\rho})^{-\frac{\rho+1}{2}}}{\sum_k (1 + \frac{\|D_z \boldsymbol{\alpha}_i - \mathbf{z}'_k\|_2^2}{\rho})^{-\frac{\rho+1}{2}}} \tag{6}$$

where ρ is the kernel parameter. The choice of t-distribution is due to its long tail and low sensitivity to the choice of kernel parameter, ρ. Ideally, $p_m(\boldsymbol{\alpha}_i) = 1$ for some $m \in \{1, ..., M\}$ and $p_j(\boldsymbol{\alpha}_i) = 0$ for $j \neq m$. The ideal soft-assignment $\mathbf{p} = [p_1, p_2, ..., p_M]$ then would be one-sparse and therefore would have minimum entropy. This motivates our attribute-aware formulation, which regularizes the AAg

formulation in Equation 5 with the entropy of $\mathbf{p}$.

$$\boldsymbol{\alpha}_i = \text{argmin}_{\mathbf{a}} \underbrace{\frac{1}{p}\|\mathbf{x}'_i - D_x\mathbf{a}\|_2^2 - \gamma \sum_m p_m(\mathbf{a}) log(p_m(\mathbf{a}))}_{g(\mathbf{a})} + \frac{\lambda}{r}\|\mathbf{a}\|_1 \tag{7}$$

where γ is the regularization parameter for entropy of the soft-assignment probability vector $\mathbf{p}$. Such entropy minimization scheme has been successfully used in several work (Grandvalet and Bengio 2004; Huang, Tran, and Tran 2016) whether as a sparsifying regularization or to boost the confidence of classifiers. We note that the entropy regularization enforces the prediction to be close to one of the unseen attributes, but it can potentially backfire in that a low-entropy solution (aligned to a prototype) doesn't necessarily have to be the correct solution. In our experiments, we consistently observed higher performance for the AAw formulation.

The entropy regularization turns the optimization in Eq. (7) into a nonconvex problem. In (Huang, Tran, and Tran 2016), the authors use a generalized gradient descent approach similar to FISTA to optimize this non-convex problem. We use a similar scheme to optimize the objective function in Eq. (7). In short, we relax $g(\mathbf{a})$ using its quadratic approximation around the previous estimation of $\mathbf{a}$, $\mathbf{a}_{k-1}$, and update $\mathbf{a}$ as the solution of the following problem

$$\mathbf{a_k} = \text{argmin}_{\mathbf{a}} \frac{1}{2t}\|\mathbf{a} - (\mathbf{a_{k-1}} - \mathbf{t}\nabla\mathbf{g}(\mathbf{a_{k-1}}))\|_{\mathbf{2}}^{\mathbf{2}} + \frac{\lambda}{r}\|a\|_1 \tag{8}$$

Equation (8) is a LASSO problem and can be solved efficiently using FISTA. It only remains to compute ∇g:

$$\nabla g(\mathbf{a}) = \frac{1}{p} D_x^T(D_x\mathbf{a} - \mathbf{x}') - \frac{\gamma}{(\sum_k l_k(\mathbf{a}))^2} \sum_m \{(1 + log(p_m(\mathbf{a}))) \times (\frac{\partial l_m(\mathbf{a})}{\partial \mathbf{a}} \sum_k l_k(\mathbf{a}) - l_m(\mathbf{a}) \sum_k \frac{\partial l_k(\mathbf{a})}{\partial \mathbf{a}})\} \tag{9}$$

where:

$$l_m(\mathbf{a}) = (\mathbf{1} + \frac{\|\mathbf{D_z a} - \mathbf{z'_m}\|_{\mathbf{2}}^{\mathbf{2}}}{\rho})^{-\frac{\rho+\mathbf{1}}{\mathbf{2}}},$$

$$\frac{\partial l_m(\mathbf{a})}{\partial \mathbf{a}} = -\frac{\rho+1}{\rho}(D_z^T(D_z\mathbf{a} - \mathbf{z}))(\mathbf{1} + \frac{\|\mathbf{D_z a} - \mathbf{z'_m}\|_{\mathbf{2}}^{\mathbf{2}}}{\rho})^{-\frac{\rho+\mathbf{3}}{\mathbf{2}}}.$$

Due to the non-convex nature of the objective function, a good initialization is needed to achieve a sensible solution. Therefore we initialize $\boldsymbol{\alpha}$ from the solution of the AAg formulation. Finally the corresponding attributes are estimated by $\hat{z}_i = D_z\boldsymbol{\alpha}_i$, for $i = 1, ..., l$.

From Predicted Attributes to Labels

In order to predict the image labels, one needs to assign the predicted attributes, $\hat{Z} = [\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_l]$, to the M attributes of the unseen classes Z' (i.e. prototypes). This task can be performed in two ways, namely the inductive approach and the transductive approach. In the inductive scheme the inference could be performed using a nearest neighbor (NN) approach in which label of each individual $\hat{\mathbf{z}}_i$ is assigned to be the label of its nearest neighbor $\mathbf{z}'_m$. In such approach the structure of $\hat{\mathbf{z}}_i$'s is not taken into account and the hubness problem could easily degrade the performance of the ZSL algorithm. Looking at the t-SNE embedding visualization (Maaten and Hinton 2008) of $\hat{\mathbf{z}}_i$'s and $\mathbf{z}'_m$'s in Figure 2, details are explained later, it can be seen that NN does not provide an optimal label assignment.

In the transductive setting, on the other hand, the attributes for all test images (i.e. unseen) are first predicted to form $\hat{Z} = [\hat{\mathbf{z}}_1, ..., \hat{\mathbf{z}}_l]$. Next, a graph is formed on $[Z', \hat{Z}]$ where the labels for Z' are known and the task is to infer the labels of $\hat{Z}$. This problem can be formulated as a graph-based semi-supervised label propagation (Belkin, Matveeva, and Niyogi 2004; Zhou et al. 2003). We follow the work of Zhou et al. (Zhou et al. 2003) and spread the labels of Z' to $\hat{Z}$. More precisely, we first reduce the dimension of $[Z', \hat{Z}]$ via t-SNE (Maaten and Hinton 2008) and then form a graph in the lower dimension and perform label propagation on this graph. Figure 2 reconfirms that label propagation in a transductive setting could significantly improve the performance of ZSL and resolve the hubness and domain shift issues as also demonstrated in (Fu et al. 2015; Yu et al. 2017).

Theoretical Discussion

The core step for ZSL in our scheme is to compute the joint sparse representation for an unseen image. Note that in the testing phase, the sparse representation $\mathbf{a}$ is estimated using (5), while the dictionaries are learned for optimal sparse representations as in (2). More specifically, we need to demonstrate that the following two problems lead to close approximations:

$$\begin{aligned} \boldsymbol{\alpha}^* &= \text{argmin}_a \|\mathbf{x} - D_x\mathbf{a}\|_2^2 + \|\mathbf{z} - D_z\mathbf{a}\|_2^2 + \lambda\|a\|_1 \\ &= \text{argmin}_a \| \begin{bmatrix}\mathbf{x}\\ \mathbf{z}\end{bmatrix} - \begin{bmatrix}D_x\\ D_z\end{bmatrix} \mathbf{a}\|_{\mathbf{2}}^{\mathbf{2}} + \lambda\|\mathbf{a}\|_{\mathbf{1}} \\ \boldsymbol{\alpha}^+ &= \text{argmin}_a \|\mathbf{x} - D_x\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_1, \end{aligned} \tag{10}$$

in order to conclude that we can solve for $\boldsymbol{\alpha}^+$ in ZSL regime (i.e. prediction attributes for unseen images) to estimate $\boldsymbol{\alpha}^*$ with good accuracy. Note that the major challenge in the testing phase is that we are using the dictionary $D_x \in \mathbb{R}^{p\times r}$ to find the shared sparse parameters, $\boldsymbol{\alpha}$, instead of $\tilde{D} = [D_x, D_z]^T \in \mathbb{R}^{(p+q)\times r}$. To study the effect of this change, we first point out that Eq. 1 can be interpreted as result of a maximum a posteriori (MAP) inference from a Bayesian perspective. This means that from a probabilistic perspective, $\boldsymbol{\alpha}$'s are drawn from a Laplacian distribution and the dictionary D is a Gaussian matrix with elements drawn i.i.d: $d_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma^2)$. This means that given a drawn dataset, we learn MAP estimate of a Gaussian matrix. In order to analyze the effect, we rely on the following

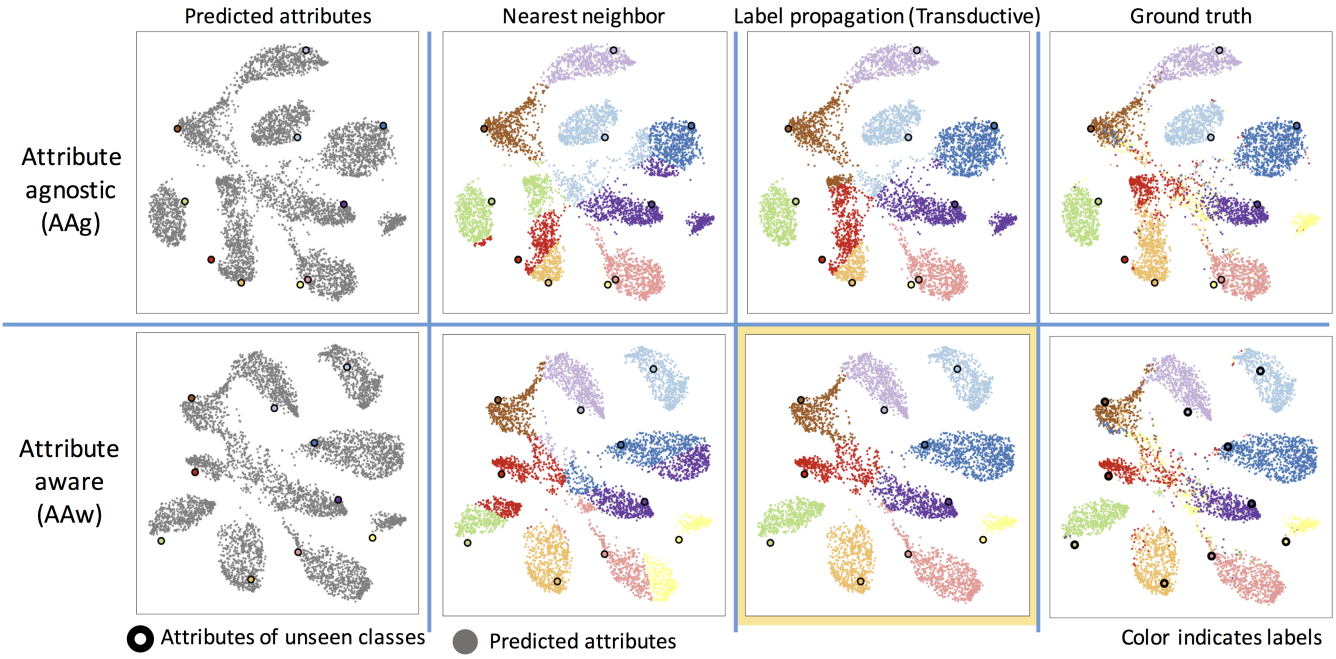

Figure 2: Attributes predicted from the input visual features for the unseen classes of images for AWA dataset using our attribute agnostic and attribute aware formulations respectively in top and bottom rows. The nearest neighbor and label propagation assignment of the labels together with the ground truth labels are visualized. It can be seen that the attribute aware formulation together with the label propagation scheme overcomes the hubness and domain shift problems. Best seen in color.

theorem about LASSO with Gausian matrices (Negahban et al. 2009):

Theorem 1 (Negahban et al. 2009): Let $\boldsymbol{\alpha}_s$ be the unique sparse solution of the linear system $\mathbf{x} = D\mathbf{a}$ with $\|\mathbf{a}\|_0 = k$ and $D \in \mathbb{R}^{p \times n}$. If $\boldsymbol{\alpha}^\ddagger$ is the LASSO solution for the system from noisy observations, then with high probability: $\|\boldsymbol{\alpha}_s - \boldsymbol{\alpha}^\ddagger\|_2 \leq c' \sqrt{k \frac{\log r}{p}}$, where $c' \in \mathbb{R}^+$ is a constant which depends on the loss function which measures the data fidelity, here the Euclidean distance.

Lemma 1: Attribute prediction error in ZSL setting is upper-bounded proportional to $(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{q+p}})$.

Proof: note that if $\boldsymbol{\alpha}^*$ is a solution of $[\mathbf{x}^\mathbf{T}, \mathbf{z}^\mathbf{T}]^\mathbf{T} = \tilde{\mathbf{D}}\mathbf{a}$, trivially it is also a solution for $\mathbf{x} = D_x\mathbf{a}$ as well. Now using Theorem 1:

$$\|\mathbf{z}^* - \mathbf{z}^+\| \leq \|D_x(\alpha^* - \alpha^+)\|$$

$$\|D_x(\alpha^* - \alpha^+)\| \leq c'\|D_z\|_2 \sqrt{k \log r}(\frac{1}{\sqrt{p}} + \frac{1}{\sqrt{q+p}}) \tag{11}$$

Note we have used the triangular inequality first and then the theorem in the above deduction and $\|\cdot\|_2$ denotes spectral norm for a matrix. This result accords with intuition. First, it advises sparseness of $\mathbf{z}$, i.e. smaller k, decreases the error which means that a good sparsifying dictionary would lead to less ZSL error. Second, the error is proportional to inverse of both $\sqrt{p}$ and $\sqrt{p+q}$, meaning that rich visual and attribute descriptions lead to minimal ZSL error. This suggests that for our approach to work, existence of a good sparsifying dictionary as well as rich visual and attribute data is essential. Finally, although increasing the number of dictionary columns r intuitively can improve sparsity, i.e. decrease k, this result shows that it can potentially increase the ZSL error, and should be tuned for an optimal performance.

Experiments

We carried out experiments on three benchmark ZSL datasets and empirically evaluated the resulting performance against nascent ZSL algorithms.

Datasets: We conducted our experiments on three benchmark datasets namely: the Animals with Attributes (AwA1) (Lampert, Nickisch, and Harmeling 2014), the SUN attribute (Patterson and Hays 2012), and the Caltech-UCSD-Birds 200-2011 (CUB) bird (Wah et al. 2011) datasets. The AwA1 dataset is a coarse-grained dataset containing 30475 images of 50 types of animals with 85 corresponding attributes for these classes. Semantic attributes for this dataset are obtained via human annotations. The images for the AWA1 dataset are not publicly available; therefore we use the publicly available features of dimension 4096 extracted from a VGG19 convolutional neural network, which was pretrained on the ImageNet dataset. Following the conventional usage of this dataset, 40 classes are used as source classes to learn the model and the remaining 10 classes are used as target (unseen) classes to test the performance of zero-shot classification. The SUN dataset is a fine-grained dataset and contains 717 classes of different scene categories with 20 images per category (14340 images total). Each im-

Method		SUN	CUB	AwA
(Romera-Paredes and Torr 2015)‡		82.10	-	75.32
(Zhang and Saligrama 2015)†		82.5	30.41	76.33
(Zhang and Saligrama 2016)†		82.83	42.11	80.46
(Bucher, Herbin, and Jurie 2016)†		84.41	43.29	77.32
(Xu et al. 2017)†		83.5	53.6	84.5
(Li et al. 2017) †		-	61.79	87.22
(Ye and Guo 2017)†		85.40	57.14	85.66
(Ding, Shao, and Fu 2017)†		86.0	45.2	82.8
(Wang and Chen 2017)†		-	42.7	79.8
(Kodirov, Xiang, and Gong 2017)†		91.0	61.4	84.7
Ours	AAg (5)	82.05	35.81	77.73
Ours	AAw (6)	83.22	38.36	83.33
Ours	Transductive AAw (TAAw)	85.90	47.12	88.23
Ours	TAAw hit@3	94.52	58.19	91.73
Ours	TAAw hit@5	98.15	69.67	97.13

Table 1: Zero-shot classification results for three benchmark datasets. All methods use VGG19 features trained on the ImageNet dataset and the original continuous (or binned) attributes provided by the datasets. Here, † indicates that the results are extracted directly from the corresponding paper, ‡ indicates that the results are reimplemented with VGG19 features, and − indicates that the results are not reported.

age is annotated with 102 attributes that describe the corresponding scene. Following (Lampert, Nickisch, and Harmeling 2014), 707 classes are used to learn the dictionaries and the remaining 10 classes are used for testing. The CUB200 dataset is a fine-grained dataset containing 200 classes of different types of birds with 11788 images with 312 attributes and boundary segmentation for each image. The attributes are obtained via human annotation. The dataset is divided into four almost equal folds, where three folds are used to learn the model and the fourth fold is used for testing. For both SUN and CUB200-2011 datasets we used features from VGG19 trained on the ImageNet dataset, which have 4096 dimensions. We note that our results using ResNet50 and DenseNet (Huang et al. 2017) features will be published in an extended version of this paper.

Tuning parameters: The optimization regularization parameters λ, ρ, γ as well as the number of dictionary atoms r need to be tuned for maximal performance. We used standard k-fold cross validation to search for the optimal parameters for each dataset. After splitting the datasets accordingly into training, validation, and testing sets, we used performance on the validation set for tuning the parameters in a brute-force search. we used the common evaluation metrics in ZSL, flat hit@K classification accuracy, to measure the performance. This means that a test image is said to be classified correctly if it is classified among the top K predicted labels. We report hit@1 rate to measure ZSL image classification performance and hit@3 and hit@5 for image retrieval performance. Each experiment is performed ten times and the mean is reported in Tabel 1.

Results: Figure 2 demonstrates the 2D t-SNE embedding for predicted attributes and actual class attributes of the AWA dataset. The actual attributes are depicted by colored circles with black edges. The first column of Figure 2 demonstrates the attribute prediction for AAg and AAw formulations. It can be clearly seen that the entropy regularization in AAw formulation improves the clustering quality, decreases data overlap, and reduces the domain shift problem. The nearest neighbor label assignment is shown in the second column, which demonstrates the domain shift and hubness problems with NN label assignment in the attribute space. The third column of Figure 2 shows the transductive approach in which a label propagation is performed on the graph of the predicted attributes. Note that the label propagation addresses the domain shift and hubness problem and when used with the AAw formulation provides significantly better zero-shot classification accuracy.

Performance comparison results are summarized in Table 1. As pointed out by Xian et al. (Xian et al. 2017) the variety of used image features (e.g. various DNNs and various combinations of these features) as well as the variation of used attributes (e.g. word2vec, human annotation), and different data splits make direct comparison with the ZSL methods in the literature very challenging. In Table 1 we provide a fair comparison of our JDZSL performance to the recent methods in the literature. All compared methods use the same visual features (i.e. VGG19) and the same attributes (i.e. the continuous or binned) provided in the dataset. Table 1 provides a comprehensive explanation of the shown results. Note that our method achieves state-of-the-art or close to state-of-the-art performance.

We report the hit@1 accuracy on unseen classes in the first nine rows of the table to measure image classification performance. For the sake of transparency and to provide the complete picture to the reader, we included results for the AAg formulation using nearest neighbor, the AAw using nearest neighbor, and AAw using the transductive approach, denoted as transductive attribute aware (TAA) formulation. As it can be seen, while the AAw formulation significantly improves the AAg formulation and adding the transductive approach (i.e. label propagation on predicted attributes) to the AAw formulation further boosts the classification accuracy, as also shown in Figure 2. In addition, our approach leads to better and comparable performance in all three datasets which include zero-shot scene and object recgonition tasks. More importantly, while the other methods can perform well on a specific dataset, our algorithm leads to competitive performance on all the three datasets.

Conclusions

A ZSL formulation is developed that models the relationship between visual features and semantic attributes via joint sparse dictionaries. We showed that while a classic joint dictionary learning approach suffers from the domain shift problem, an entropy regularization scheme can help with this phenomenon and provide superior performance. In addition, we demonstrated that a transductive approach towards assigning labels to the predicted attributes can boost the performance considerably and lead to state-of-the-art zero-shot classification. Finally, we compared our method to the nascent approaches in the literature and demonstrated its competitiveness on benchmark datasets.

References

[Akata et al. 2013] Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 819–826.

[Beck and Teboulle 2009] Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* 2(1):183–202.

[Belkin, Matveeva, and Niyogi 2004] Belkin, M.; Matveeva, I.; and Niyogi, P. 2004. Regularization and semi-supervised learning on large graphs. In *Conference on Learning Theory*, 624–638. Springer.

[Bucher, Herbin, and Jurie 2016] Bucher, M.; Herbin, S.; and Jurie, F. 2016. Improving semantic embedding consistency by metric learning for zero-shot classiffication. In *European Conference on Computer Vision*, 730–746. Springer.

[Ding, Shao, and Fu 2017] Ding, Z.; Shao, M.; and Fu, Y. 2017. Low-rank embedded ensemble semantic dictionary for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2050–2058.

[Dinu, Lazaridou, and Baroni 2014] Dinu, G.; Lazaridou, A.; and Baroni, M. 2014. Improving zero-shot learning by mitigating the hubness problem. *arXiv preprint arXiv:1412.6568*.

[Fei-Fei, Fergus, and Perona 2006] Fei-Fei, L.; Fergus, R.; and Perona, P. 2006. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence* 28(4):594–611.

[Fu et al. 2015] Fu, Y.; Hospedales, T. M.; Xiang, T.; and Gong, S. 2015. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence* 37(11):2332–2345.

[Grandvalet and Bengio 2004] Grandvalet, Y., and Bengio, Y. 2004. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, volume 17, 529–536.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *International Conference on Computer Vision*, 770–778.

[Huang et al. 2017] Huang, G.; Liu, Z.; Weinberger, K. Q.; and van der Maaten, L. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, 4700–4708.

[Huang, Tran, and Tran 2016] Huang, S.; Tran, D. N.; and Tran, T. D. 2016. Sparse signal recovery based on non-convex entropy minimization. In *IEEE International Conference on Image Processing*, 3867–3871. IEEE.

[Isele, Rostami, and Eaton 2016] Isele, D.; Rostami, M.; and Eaton, E. 2016. Using task features for zero-shot knowledge transfer in lifelong learning. In *Proc. of International Joint Conference on Artificial Intelligence*, 1620–1626.

[Kodirov et al. 2015] Kodirov, E.; Xiang, T.; Fu, Z.; and Gong, S. 2015. Unsupervised domain adaptation for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 2452–2460.

[Kodirov, Xiang, and Gong 2017] Kodirov, E.; Xiang, T.; and Gong, S. 2017. Semantic autoencoder for zero-shot learning. 3174–3183.

[Krizhevsky, Sutskever, and Hinton 2012] Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

[Lampert, Nickisch, and Harmeling 2009] Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 951–958.

[Lampert, Nickisch, and Harmeling 2014] Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(3):453–465.

[Lei Ba et al. 2015] Lei Ba, J.; Swersky, K.; Fidler, S.; et al. 2015. Predicting deep zero-shot convolutional neural networks using textual descriptions. In *International Conference on Computer Vision*, 4247–4255.

[Li et al. 2017] Li, Y.; Wang, D.; Hu, H.; Lin, Y.; and Zhuang, Y. 2017. Zero-shot recognition using dual visual-semantic mapping paths. 3279–3287.

[Maaten and Hinton 2008] Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research* 9(Nov):2579–2605.

[Mikolov et al. 2013] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.

[Negahban et al. 2009] Negahban, S.; Yu, B.; Wainwright, M.; and Ravikumar, P. 2009. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in neural information processing systems*, 1348–1356.

[Norouzi et al. 2014] Norouzi, M.; Mikolov, T.; Bengio, S.; Singer, Y.; Shlens, J.; Frome, A.; Corrado, G. S.; and Dean, J. 2014. Zero-shot learning by convex combination of semantic embeddings. *International Conference on Learning Representations*.

[Palatucci et al. 2009] Palatucci, M.; Pomerleau, D.; Hinton, G. E.; and Mitchell, T. M. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, 1410–1418.

[Patterson and Hays 2012] Patterson, G., and Hays, J. 2012. Sun attribute database: Discovering, annotating, and recognizing scene attributes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2751–2758. IEEE.

[Pennington, Socher, and Manning 2014] Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Empirical Methods on Natural Language Processing*, volume 14, 1532–43.

[Romera-Paredes and Torr 2015] Romera-Paredes, B., and Torr, P. 2015. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*, 2152–2161.

[Shekhar et al. 2014] Shekhar, S.; Patel, V. M.; Nasrabadi, N. M.; and Chellappa, R. 2014. Joint sparse representation for robust multimodal biometrics recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(1):113–126.

[Shigeto et al. 2015] Shigeto, Y.; Suzuki, I.; Hara, K.; Shimbo, M.; and Matsumoto, Y. 2015. Ridge regression, hubness, and zero-shot learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 135–151. Springer.

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

[Socher et al. 2013] Socher, R.; Ganjoo, M.; Manning, C. D.; and Ng, A. 2013. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, 935–943.

[Wah et al. 2011] Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

[Wang and Chen 2017] Wang, Q., and Chen, K. 2017. Zero-shot visual recognition via bidirectional latent embedding. *International Journal of Computer Vision* 124(3):356–383.

[Xian et al. 2017] Xian, Y.; Lampert, C. H.; Schiele, B.; and Akata, Z. 2017. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *arXiv preprint arXiv:1707.00600*.

[Xu et al. 2017] Xu, X.; Shen, F.; Yang, Y.; Zhang, D.; Shen, H. T.; and Song, J. 2017. Matrix tri-factorization with manifold regularizations for zero-shot learning. In *Proceedings of the IEEE International Conference on Computer Vision*, 3798–3807.

[Yang et al. 2010] Yang, J.; Wright, J.; Huang, T. S.; and Ma, Y. 2010. Image super-resolution via sparse representation. *IEEE transactions on image processing* 19(11):2861–2873.

[Yang et al. 2012] Yang, J.; Wang, Z.; Lin, Z.; Cohen, S.; and Huang, T. 2012. Coupled dictionary training for image super-resolution. *IEEE Transactions on Image Processing* 21(8):3467–3478.

[Ye and Guo 2017] Ye, M., and Guo, Y. 2017. Zero-shot classification with discriminative semantic representation learning. 17140–17148.

[Yu and Aloimonos 2010] Yu, X., and Aloimonos, Y. 2010. Attribute-based transfer learning for object categorization with zero/one training example. *European Conference on Computer Vision* 127–140.

[Yu et al. 2014] Yu, Z.; Wu, F.; Yang, Y.; Tian, Q.; Luo, J.; and Zhuang, Y. 2014. Discriminative coupled dictionary hashing for fast cross-media retrieval. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, 395–404. ACM.

[Yu et al. 2017] Yu, Y.; Ji, Z.; Li, X.; Guo, J.; Zhang, Z.; Ling, H.; and Wu, F. 2017. Transductive zero-shot learning with a self-training dictionary approach. *arXiv preprint arXiv:1703.08893*.

[Zhang and Saligrama 2015] Zhang, Z., and Saligrama, V. 2015. Zero-shot learning via semantic similarity embedding. In *International Conference on Computer Vision*, 4166–4174.

[Zhang and Saligrama 2016] Zhang, Z., and Saligrama, V. 2016. Zero-shot learning via joint latent similarity embedding. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 6034–6042.

[Zhou et al. 2003] Zhou, D.; Bousquet, O.; Lal, T. N.; Weston, J.; and Schölkopf, B. 2003. Learning with local and global consistency. In *Advances in neural information processing systems*, volume 16, 321–328.