Neurocomputing 139 (2014) 34–46

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Attribute relation learning for zero-shot classification

Mingxia Liu [a,b], Daoqiang Zhang [a,*], Songcan Chen [a]

[a] School of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China
[b] School of Information Science and Technology, Taishan University, Taian 271021, China

ARTICLE INFO

Article history:
Received 4 March 2013
Received in revised form
31 August 2013
Accepted 18 September 2013
Available online 3 April 2014

Keywords:
Attribute
Attribute relation
Zero-shot learning
Classification

ABSTRACT

In computer vision and pattern recognition communities, one often-encountered problem is that the limited labeled training data are not enough to cover all the classes, which is also called the zero-shot learning problem. For addressing that challenging problem, some visual and semantic attributes are usually used as mid-level representation to transfer knowledge from training classes to unseen test ones. Recently, several studies have investigated to exploit the relation between attributes to aid the attribute-based learning methods. However, such attribute relation is commonly *predefined* by means of external linguistic knowledge bases, preprocessed in advance of the learning of attribute classifiers. In this paper, we propose a unified framework that learns the attribute–attribute relation and the attribute classifiers *jointly* to boost the performances of attribute predictors. Specifically, we unify the attribute relation learning and the attribute classifier design into a common objective function, through which we can not only predict attributes, but also *automatically* discover the relation between attributes from data. Furthermore, based on the afore-learnt attribute relation and classifiers, we develop two types of learning schemes for zero-shot classification. Experimental results on a series of real benchmark data sets suggest that mining the relation between attributes do enhance the performances of attribute prediction and zero-shot classification, compared with state-of-the-art methods.

© 2014 Elsevier B.V. All rights reserved.

1. Introduction

For traditional object classification problems in computer vision and pattern recognition, an intent classifier is usually designed based on a training set with samples from specific classes, and then evaluated on a test set using samples from the same categories. However, we are currently faced with more challenging problem scenarios, such as (1) multi-class classification problems where there are not enough labeled training data to cover all object classes, and (2) multi-task ones where the available training data do not provide examples of desired outputs, which are called zero-shot learning [1,2]. As a meaningful and challenging problem setting, zero-shot learning has attracted increasing attention in recent years [1–9]. It is worth noting that many design schemes for traditional classifiers cannot be directly used for zero-shot learning, due to the fact that the training and the target classes are commonly disjoint or different.

On the other hand, attributes describing objects (e.g. “red”, “stripe” and “similar to dog”) can be obtained relatively more conveniently than traditional class labels, because humans can typically provide good prior knowledge about attributes for some objects though they may be unable to name them. For example, Fig. 1 illustrates images of pandas and tigers for which we might not know their precise names (class labels) if we have never seen them before but can partly describe these objects via some of their attributes such as *furry*, *black*, *big*, and *with ear* etc. Moreover, as attributes are assigned on a per-class basis other than a per-image basis, it can significantly reduce the manual effort on adding a new object category.

Here, the term “attribute”, as defined in Webter's dictionary, means “an inherent characteristic” of an object. For zero-shot learning problems, researchers have explored various kinds of attributes, e.g., similarity to known object categories [4,10], appearance adjectives (such as shape, texture, and color) [9,11], and the presence or absence of parts [3]. Accordingly, these attributes are classified as (1) similarity based attributes [4,10], (2) semantic attributes which can be described in language [9,11], and (3) discriminative attributes [3]. Due to the fact that attributes can be frequently shared by different objects, they can be deemed as one type of transferable knowledge [3,9]. In fact, it has also been shown that attributes behave effectively in leveraging knowledge between different object categories and compensating for the lack of labeled data [1,8,12,13], which is especially valuable for zero-shot learning problems. Consequently, a class of attribute-based methods in face verification [6], image annotation [14], image

* Corresponding author. Tel.: +86 2584896481x12217.
E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

http://dx.doi.org/10.1016/j.neucom.2013.09.056
0925-2312/© 2014 Elsevier B.V. All rights reserved.

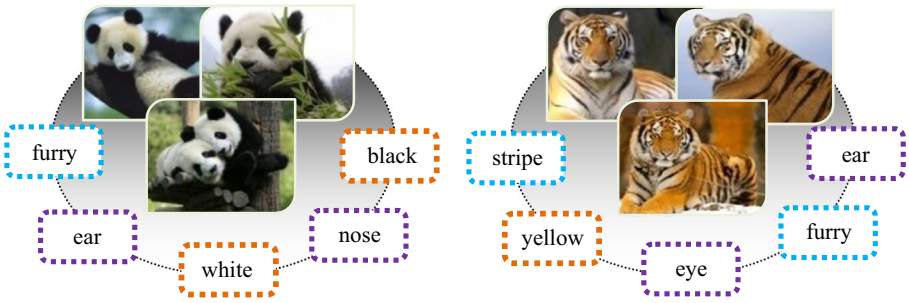


Fig. 1. Describing objects by attributes.

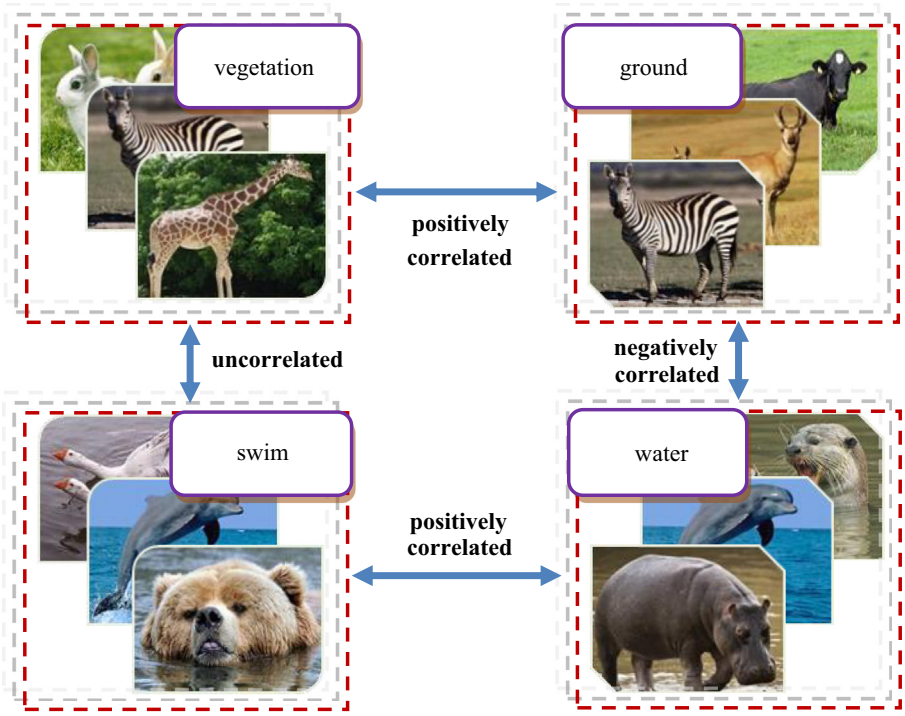


Fig. 2. Illustration of the relationship between attributes.

search and retrieval [15–18] has been developed with promising results.

Among various attribute-based approaches, the exploitation of the relationship between *attribute* and *object class*, which may convey prior information about object categories, has been proven beneficial to enhance the performance of object classification [8,19,20]. Likewise, intuitively the relationship between attributes of objects can also convey supplementary information from another aspect. As illustrated in Fig. 2, attributes corresponding to "vegetation" and "ground" are usually positively correlated, "ground" and "water" might be negatively correlated, while "swim" and "vegetation" tend to be uncorrelated. Unfortunately, in traditional attribute-based methods [3,9,14,21], the *relationship between attributes* is seldom considered.

As mentioned above, there exit some correlation relation (i.e., positive, negativeand uncorrelated ones, etc.) between attributes. In the literature, some works have also demonstrated that mining such relation is helpful to boost the performance of attribute classifiers. For example, researchers in [4,14,19,22] have considered the attribute–attribute relation to some degree. However, the relation they adopted is largely predefined in terms of some criteria such as correlation [4,22] and mutual information [14,19], or by means of external linguistic knowledge bases. Intuitively, it will be more reasonable and more prone to be optimal if we can learn the relationship between attributes automatically from data.

In this paper, we propose to unify the attribute relation learning and the attribute classifier design into a common objective function, through which we can not only predict attributes (e.g. "white" and "not furry"), but also *automatically* discover the (correlation) relation between attributes (e.g. "side mirror" is highly related to "head light"). The major contribution of this paper is two-fold: (1) we propose a model called attribute relation learning (ARL) to jointly learn the attribute–attribute relation and the attribute classifiers, through which the (correlation) relation between attributes can be discovered automatically and explicitly; (2) we develop two learning schemes for attribute-based zero-shot classification based on the afore-learnt attribute relation and classifiers. To the best of our knowledge, our work is the first attempt to formulate the learning of attribute–attribute relation and attribute classifiers into a unified objective function, and to reuse such relation to boost traditional attribute classifiers that are trained separately.

The rest of this paper is organized as follows. We introduce related works on zero-shot learning, multi-task relation learning and attribute relation learning in Section 2. In Section 3, we first present our proposed model for jointly learning attribute–attribute

relation and attribute classifiers, and then introduce an alternating optimization algorithm to solve the optimization problem. Section 4 introduces the attribute–class mapping method and the proposed two types of attribute-based zero-shot classification schemes with attribute relation incorporated. Experimental results are reported in Section 5. Finally, we conclude this paper and indicate several issues for future works in Section 6.

2. Related works

In the following, we will first briefly review the related works on zero-shot learning, multi-task relation learning, and then review the most relevant works on attribute relation learning.

2.1. Zero-shot learning

To the best of our knowledge, the concept of zero-shot learning can be traced back to one of the early works proposed by Larochelle et al. [1], which attempts to solve the problem of predicting novel samples that were unseen in the training categories. It is an important problem setting especially when the labeled training data are not enough to cover all the classes. Two key components are necessary in zero-shot learning [1]. One is how to generate an intermediate level representation to transfer knowledge from observed training classes to unseen target ones, and the other deals with mapping test data in the intermediate level representation to novel categories.

For the first problem, various representation of transferable knowledge have been proposed, such as distance metrics [23,24], class priors [25,26], discriminating aspects [27,28], and descriptive attributes [3,10,12,29]. Larochelle et al. [1] adopted a character font bitmap as the intermediate class description for zero-shot learning, and mapped novel classes of hand written digits into these bitmaps. However, such kind of intermediate class description cannot be generated to deal with non-trivial problems such as object recognition and scene recognition. Researchers in [2,30] proposed methods to obtain the intermediate class description such as semantic knowledge base to perform zero-shot learning. Meanwhile, in computer vision community, researchers have suggested that attributes are useful to describe familiar and unfamiliar things [3,7,12,29,31–34], to serve as mid-level features for object classification [9,10,19] and image retrieval [16,17,20,35], and to facilitate zero-shot learning paradigms [3,9,13]. Attributes have become an important way to perform zero-shot learning by transferring knowledge from training set to novel test set.

For the second problem, Lampert et al. [9] proposed an attribute-based object class model for zero-shot recognition, given the association between object categories and attributes. Informally, the attribute-based classification method models object classes through an inventory of descriptive attributes provided by human beings. They also develop two techniques to perform zero-shot learning with the help of attributes, i.e. Direct Attribute Prediction (DAP) and Indirect Attribute Prediction (IAP). It is also suggested in [9] that DAP is superior to IAP in zero-shot classification problems. Following the pioneering work of Lampert et al. [9], a number of attribute-based methods have been recently developed for addressing different kinds of applications [1,8,11,13,36].

2.2. Multi-task relation learning

Multi-task learning [37–39] learns multiple tasks jointly, seeking to improve the generalization performance of a learning task with the help of some other related tasks. A major assumption for this learning paradigm is that all tasks are related so that the joint learning is beneficial, and hence modeling the relation between tasks accurately is critical.

Recently, the so-called task relationship learning approach has been proposed to discover the relationship between tasks automatically, e.g., the multi-task Gaussian process (GP) model [40] and a regularized method [41] under the Bayesian framework. Using a task covariance matrix, this approach provides an effective way to characterize different types of pairwise task relation. To be specific, the task covariance matrix $\mathbf{C}$ is defined as a parameter matrix for the matrix–variate normal distribution [42] over the model parameters in least square regression or support vector machine (SVM) [43]:

$$\mathbf{W}=[\mathbf{w}_1,\mathbf{w}_2,\ldots,\mathbf{w}_M]\sim \mathcal{MN}_{D\times M}(\mathbf{W}|0_{D\times M},\mathbf{I}_D\otimes\mathbf{C}) \tag{1}$$

where $\mathbf{w}_i\in\mathbb{R}^D$is the model parameter vector for the ith task, $\mathcal{MN}_{D\times M}(\mathbf{M},\mathbf{A}\otimes\mathbf{C})$ denotes a matrix–variate normal distribution with mean $\mathbf{M}\in\mathbb{R}^{D\times M}$, row covariance matrix $\mathbf{A}\in\mathbb{R}^{D\times D}$ and column covariance matrix $\mathbf{C}\in\mathbb{R}^{M\times M}$. The probability density function of the matrix–variate normal distribution is as follows:

$$p(\mathbf{W}|\mathbf{M},\mathbf{A},\mathbf{C})=\frac{\exp\left(-(1/2)tr(\mathbf{A}^{-1}(\mathbf{X}-\mathbf{M})\mathbf{C}^{-1}(\mathbf{X}-\mathbf{M})^T)\right)}{(2\pi)^{MD/2}|\mathbf{A}|^{M/2}|\mathbf{C}|^{D/2}} \tag{2}$$

where $tr(\cdot)$ denotes the trace of a square matrix, $|\cdot|$ denotes the determinant of a square matrix, and $\mathbf{C}^{-1}$ denotes the inverse of a non-singular matrix $\mathbf{C}$ or the pseudo-inverse when it is singular.

Given the prior defined in Eq. (1) and the likelihood (i.e., Gaussian noise model for regression problem or logistic model for classification problem), the maximum a posteriori (MAP) solution is obtained by solving the following problem:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{C}}\ \sum_{m=1}^{M}\sum_{i=1}^{N} l(y_i^m,f_m(\mathbf{x}_i^m))+\frac{\lambda}{2}tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T)\\ &s.t.\quad \mathbf{C}\geq 0,\quad tr(\mathbf{C})=1\end{aligned} \tag{3}$$

where $l(\cdot,\cdot)$ denotes the empirical loss corresponding to the likelihood, λ is a regularization parameter to balance the tradeoff between the empirical loss and the regularization term, and the constraint $\mathbf{C}\geq 0$ in Eq. (3) is used to restrict $\mathbf{C}$ as positive semi-definite because it denotes a task covariance matrix. The second constraint in Eq. (3) is derived from the matrix–variate normal prior in Eq. (1) and is used to regularize the task relation. One can see that the matrix $\mathbf{C}^{-1}$ is used to model the covariance between the columns in the model parameter $\mathbf{W}$, and hence to model the task relationship since each column in $\mathbf{W}$ represents the model parameters of the corresponding task.

There are two advantages of the task relationship learning approach: (1) three types of relation can be modeled, i.e., positive, negative and uncorrelated relation; (2) the relationship between tasks can be learned from data automatically through the joint learning of multiple tasks. More recently, this approach has been widely applied to many domains demonstrating its effectiveness, such as multi-task learning [44–48], transfer learning [49], multi-label learning [50,51], and multi-view learning [52].

It is worth noting that the data used for tasks in multi-task learning vary from each other, while different attribute classifiers share the same data in attribute-based classification. However, to characterize the relationship between attributes, we can adopt the task relationship learning approach in multi-task learning to model the attribute–attribute relation in this work.

2.3. Attribute relation learning

As mentioned above, attributes can be used to serve as a kind of intermediate representation of transferable knowledge in zero-shot learning scenarios, which is called attribute-based classification. Among extensive study on attribute-based classification,

some researchers have demonstrated that considering the *attribute–class* relation which is a type of prior information can result in improved generalization performance [8,14,19,20]. In contrast, less attention is focused on *attribute–attribute* relation learning, where attributes are expected to benefit from each other.

Recently, some works have investigated incorporating attribute–attribute relation to attribute-based classification. For example, Wang et al. [19] and Kovashka et al. [14] proposed models that treat object attributes and class labels as latent variables, where the relation between attributes is captured through a tree-structured graph. In this graph, vertices denote attributes and edges are restricted to pairs of attributes that have the highest mutual information. However, this type of attribute–attribute relation is pre-computed. Rohrbach et al.[4] use external linguistic knowledge bases (e.g. *WordNet*, *Wikipedia*, and *Yahoo Web*) to mine the semantic relationship between object classes and corresponding attributes. It has been shown that web search for part–whole relationship is a better way to mine the association of attribute–attribute for object categories. Similarly, researchers of [12,22] mine attribute–attribute association from language to achieve unsupervised knowledge transfer. In these works, extra expert knowledge for natural language processing is required, and the object–attribute relation other than the attribute–attribute relation learning is their focus. Siddiquie et al. [20] employ structured learning to form a framework for images ranking and retrieval based on multi-attribute queries, demonstrating modeling pair-wise correlation between different attributes can lead to better performance.

Different from previous work, we use the task relationship learning approach used in multi-task learning [40,41] to model the relationship between attributes. In this way, we can learn the positive, negative and uncorrelated relation between attributes automatically from data. Furthermore, we propose to reuse the discovered relation between attributes to boost the classification performance of novel object categories in zero-shot learning scenarios, by incorporating such relation into traditional classifiers, e.g., Support Vector Machine and kernel ridge regression.

3. Our attribute relation learning (ARL) approach

In this section, we will describe our proposed approach of learning the relation between attributes automatically from data in detail. As multiple attributes are usually used in the attributes-based classification, we resort to the task relationship learning approach in multi-task learning [41] to model the correlation relation between attributes. We start by describing the notations used in this paper, and then propose the unified formulation to learn the attribute–attribute relation and attribute classifiers simultaneously from a multi-task learning respective. Finally, an alternating optimization method is introduced to solve the proposed objective function.

3.1. Notations

Suppose we are given M attributes for each object class. For the m-th attribute classifier, the training set $\mathbf{D}_m$consists of N_m data points$\{\mathbf{x}_i\}_{i=1}^{N_m} \in \mathbb{R}^d$, represented as$\mathbf{D}_m = [\mathbf{x}_1, \cdots, \mathbf{x}_{N_m}] \in \mathbb{R}^{d \times N_m}$. For each data point $\mathbf{x}_i \in \mathbb{R}^d$, its corresponding attributes label vector is denoted as $y_i^m \in \{1, 0\}$. Note that only binary attributes are considered in this work. For each attribute classifier, we want to learn a linear function $f_m(\mathbf{x}) = \mathbf{w}_m^T\mathbf{x} + b_m$, where $\mathbf{w}_m$ is the weight vector and $b_m$is the bias term. Denote $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]$ where $\mathbf{w}_m \in \mathbb{R}^d$, and$\mathbf{b} = [b_1, \cdots, b_M]^T$. Traditionally, the parameters of attribute classifiers are learned through training corresponding classifiers independently, without any attribute relation considered. In the following, we will describe our approach where the attribute classifier and the attribute relation can be learned jointly in a unified object function regularized by a task covariance matrix.

3.2. Problem formulation

Following the notations in the previous sub-section, we propose the following objective function to learn multiple attribute classifiers jointly from a multi-task learning perspective:

$$\min_{\mathbf{W}} \sum_{m=1}^{M} \sum_{i=1}^{N_m} Loss(y_i^m, f_m(\mathbf{x}_i^m)) + \lambda_1 \sum_{m=1}^{M} R(\mathbf{w}_m) \tag{4}$$

where the first term is a loss function which gives the empirical loss on the training data, and second term is a regularizer to control the complexity of weight vector $\mathbf{w}_m$, while λ_1 is a regularization parameter used to tune the tradeoff between the empirical loss and the regularization term. Actually, the model defined in Eq. (4) is a general framework for joint attribute classifiers' learning, as one can utilize various loss functions (e.g., least squares loss and hinge loss) and various regularizers (e.g., l_1-norm, l_2-norm and nuclear norm) to adapt to the problems at hand. For simplicity, we adopt the least squares loss function and l_2-norm regularizer in this paper. Accordingly, the problem defined in Eq. (4) can be rewritten as follows:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\{\rho_i^m\}_{m=1}^{M}} \sum_{m=1}^{M} \sum_{i=1}^{N_m} (\rho_i^m)^2 + \frac{\lambda_1}{2} tr(\mathbf{W}\mathbf{W}^T) \\ &s.t. \quad y_i^m - (\mathbf{w}_m^T\mathbf{x}_i^m + b_m) = \rho_i^m, \quad \forall m, m = 1, \ldots, M \end{aligned} \tag{5}$$

where ρ_i^m is a slack variable.

Following the work in [38], we resort to the column covariance matrix of $\mathbf{W}$ to model the relationship between $\mathbf{w}_m$'s. Accordingly, we get the following objective function:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\mathbf{C},\{\rho_i^m\}_{m=1}^{M}} \sum_{m=1}^{M} \sum_{i=1}^{N_m} (\rho_i^m)^2 + \frac{\lambda_1}{2} tr(\mathbf{W}\mathbf{W}^T) + \frac{\lambda_2}{2} tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T) \\ &s.t. \quad y_i^m - (\mathbf{w}_m^T\mathbf{x}_i^m + b_m) = \rho_i^m, \quad \forall m, m = 1, \ldots, M \\ &\mathbf{C} \geq 0, \quad tr(\mathbf{C}) = 1 \end{aligned} \tag{6}$$

where λ_1 and λ_2 are two regularization parameters, and $\mathbf{C}$ is the column covariance matrix of the weight matrix $\mathbf{W}$. The constraint $\mathbf{C} \geq 0$ in Eq. (6) is used to restrict $\mathbf{C}$ as positive semi-definite because it denotes a task covariance matrix. The second constraint in Eq. (6) is used to regularize the task relation. Here the inverse covariance matrix $\mathbf{C}^{-1}$ plays a role of coupling pairs of weight vectors, reflecting the attribute–attribute relationship.

To avoid the data imbalance problem partly where one attribute classifier with so many data points dominates the empirical loss, we reformulate the problem defined in Eq. (6) as follows:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\mathbf{C},\{\rho_i^m\}_{m=1}^{M}} \sum_{m=1}^{M} \frac{1}{N_m} \sum_{i=1}^{N_m} (\rho_i^m)^2 + \frac{\lambda_1}{2} tr(\mathbf{W}\mathbf{W}^T) + \frac{\lambda_2}{2} tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T) \\ &s.t. \quad y_i^m - (\mathbf{w}_m^T\mathbf{x}_i^m + b_m) = \rho_i^m, \quad \forall m, m = 1, \ldots, M \\ &\mathbf{C} \geq 0, \quad tr(\mathbf{C}) = 1 \end{aligned} \tag{7}$$

In the attribute-based zero-shot learning scenarios, all attributes usually share a same training set. In such cases, the problem defined in Eq. (7) can be rewritten as follows:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{b},\mathbf{C},\{\rho_i^m\}_{m=1}^{M}} \sum_{m=1}^{M} \frac{1}{N} \sum_{i=1}^{N} (\rho_i^m)^2 + \frac{\lambda_1}{2} tr(\mathbf{W}\mathbf{W}^T) + \frac{\lambda_2}{2} tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T) \\ &s.t. \quad y_i^m - (\mathbf{w}_m^T\mathbf{x}_i^m + b_m) = \rho_i^m, \quad \forall m, m = 1, \ldots, M \\ &\mathbf{C} \geq 0, \quad tr(\mathbf{C}) = 1 \end{aligned} \tag{8}$$

which is called attribute relation learning (ARL) model in this paper. Through this model, we can obtain not only the parameters

of all attribute classifiers but also the attribute–attribute relation explicitly and automatically.

It is easy to find that the problem of Eq. (8) is jointly convex with respect to **W**, **b** and **C**. For reasons of efficiency, an alternating optimization method is adopted to optimize this model, with more details described in the following.

3.3. Alternating optimization algorithm

Now we are in the position of considering the optimization of the model defined in Eq. (8) with three variables to be optimized. As the variables in Eq. (8) are jointly convex, the problem can be solved by an alternating optimization method. The first step is to optimize **W** and **b** given a fixed **C**, and the second step is to optimize **C** when **W** and **b** are fixed.

3.3.1. Optimizing *W* and *b* when *C* is fixed

Given a fixed **C**, the problem defined in Eq. (8) can be presented in the following form:

$$\min_{\mathbf{W},\mathbf{b},\mathbf{C},\{\rho_i^m\}} \sum_{m=1}^{M}\frac{1}{N_i}\sum_{i=1}^{N}(\rho_i^m)^2+\frac{\lambda_1}{2}tr(\mathbf{W}^T\mathbf{W})+\frac{\lambda_2}{2}tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T)$$
$$s.t.\quad y_i^m-(\mathbf{w}_m^T\mathbf{x}_i^m+b_m)=\rho_i^m,\qquad \forall m, m=1,\ldots,M \tag{9}$$

With the well-known kernel trick [53], it is easy to show that the dual form of Eq. (9) can be written as follows:

$$\min_{\alpha}\quad \tfrac{1}{2}\alpha^T\tilde{\mathbf{K}}\alpha-\sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_i^m y_i^m$$
$$s.t.\quad \sum_{i=1}^{N}\alpha_i^m=0,\quad \forall m,\ m=1,\ldots,M \tag{10}$$

where $\boldsymbol{\alpha}=(\alpha_1^1,\ldots\alpha_N^1,\ldots,\alpha_1^M,\ldots,\alpha_N^M)^T$, and Λ is diagonal with elements value N if the corresponding data point belongs to the m-th attribute classifier. Note that $\tilde{\mathbf{K}}=\mathbf{K}+(1/2)\Lambda$, and **K** is the kernel matrix on all data points for all attributes classifiers with element

$$k(\mathbf{x}_{i1}^{m1},\mathbf{x}_{i2}^{m2})=\mathbf{e}_{m1}^T\mathbf{C}(\lambda_1\mathbf{C}+\lambda_2\mathbf{I}_M)^{-1}\mathbf{e}_{m2}(\mathbf{x}_{i1}^{m1})^T\mathbf{x}_{i2}^{m2} \tag{11}$$

where $\mathbf{e}_m$ is the m-th column vector of $M\times M$ identity matrix $\mathbf{I}_M$.

Note that $\mathbf{W}=\sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_i^m\mathbf{x}_i^m\mathbf{e}_m^T\mathbf{C}(\lambda_1\mathbf{C}+\lambda_2\mathbf{I}_M)^{-1}$. For reasons of efficiency, we employ the sequential minimal optimization (SMO) algorithm [54] to solve the problem defined in Eq. (10).

3.3.2. Optimizing *C* when *W* and *b* are fixed

If **W** and **b** are fixed, the problem in Eq. (8) can be expressed in the following form:

$$\min_{\mathbf{C}} tr(\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^T) s.t. \mathbf{C}\succeq 0,\ \ tr(\mathbf{C})=1 \tag{12}$$

The optimal **C** that minimizes Eq. (11) has the following closed-form solution:

$$\mathbf{C}=\frac{(\mathbf{W}^T\mathbf{W})^{1/2}}{\boldsymbol{tr}\left((\mathbf{W}^T\mathbf{W})^{1/2}\right)} \tag{13}$$

The previous two steps are performed alternatively, until the optimization procedure converges or the maximal iteration number is reached. Algorithm 1 lists the main steps of our attribute relation learning (ARL) model.

Algorithm 1. Attribute relation learning (ARL)

Input: data matrix **X**; attribute labels $\mathbf{Y}=\{y_m\}_{m=1}^{M}$; the regularization parameters λ_1,λ_2; the maximum number for iteration *MaxIter*.

Output: **W**, **b**, **C**, **Cor**.

begin

1: Initialize **C** with an identity matrix $(1/N)\mathbf{I}$;

2: Compute the kernel matrix **K** according to Eq. (11) and $\tilde{\mathbf{K}}=\mathbf{K}+(1/2)\Lambda$;

3: **for** $k=1,\ldots,$ *MaxIter*

(1) Given fixed **C**, compute the optimal α and the bias term **b** using the SMO algorithm to solve the optimization problem in Eq. (10);

(2) Compute $\mathbf{W}=\sum_{m=1}^{M}\sum_{i=1}^{N}\alpha_i^m\mathbf{x}_i^m\mathbf{e}_m^T\mathbf{C}(\lambda_1\mathbf{C}+\lambda_2\mathbf{I}_M)^{-1}$

(3) Given fixed Wandb, compute Caccording to Eq. (13);

end for

4: Compute the attribute correlation matrix $\mathbf{Cor}_{ij}=\mathbf{C}_{ij}/(\mathbf{C}_{ii}\times\mathbf{C}_{jj})^{1/2},\quad \forall i,j,=1,\cdots,M.$

end

4. Incorporating attribute relation into zero-shot classification

As mentioned in [9], there are two major components in zero-shot learning. The first one concerns generating an intermediate level representation, e.g., attributes which are employed in this work. And the second one is to map test data points represented in the mid-level form to novel categories. In this paper, we adopt the probabilistic Direct Attribute Prediction (DAP) model introduced in [9] to perform such attribute–class mapping. In the following, we first briefly review the scheme of DAP, and then propose two schemes for attribute-based zero-shot classification by reusing the attribute relation learned from the proposed ARL method.

4.1. DAP Model

For zero-shot learning problems where the training and test classes are disjoint, we denote the K training classes as $\mathbf{S}=\{s_i\}_{i=1}^{K}$ and the L novel test classes as $\mathbf{U}=\{u_i\}_{i=1}^{L}$. If for all the training and test classes an attribute representation $a\in\mathbf{Y}$ is available, we can learn a non-trivial classifier $\alpha:\ \mathbf{X}\rightarrow\mathbf{S}$ by transferring information between **S** and **U** through **Y**.

As one general method to integrate attributes into multi-class classification, the scheme of DAP model is illustrated in Fig. 4. It uses a mid-level layer of attribute variables to decouple the inputs from the layer of class labels. In the training process, the output class label of each sample induces a deterministic labeling of the attribute layer, where the parameters of each attribute classifiers can be determined. In the test process, the attribute values are predicted by the trained attribute classifiers, from which the test class labels can be inferred. In the following, we explain the graphical structure in Fig. 3 in a probabilistic way [9].

Given a set of training data $\mathbf{D}=\{\mathbf{x}_i\}_{i=1}^{N}$ where $\mathbf{x}_i\in\mathbb{R}^d$, and an attribute representation $a=(a_1,a_2,\ldots,a_M)$ for any training class s, we can learn M probabilistic classifiers for each attribute a_m, and compute the posterior probability $p(a_m|\mathbf{x})$ of this attribute. Then the complete image–attribute layer can be expressed as

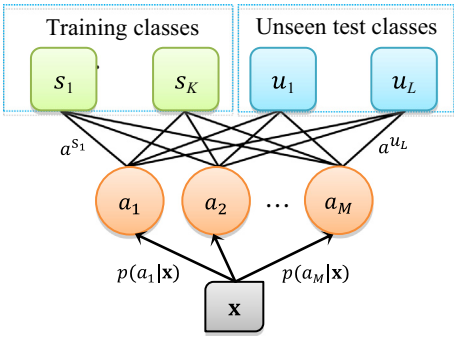


Fig. 3. Direct Attribute Prediction model for zero-shot classification.

$p(a|\mathbf{x}) = \prod_{m=1}^{M} p(a_m|\mathbf{x})$ which is provided by the learnt attribute classifiers. In the test process, we assume that every test class u induces its attribute vector a^u in a deterministic way, i.e. $p(a|u)$ is equal to 1 if $a = a^u$ and 0 otherwise.

Under Bayes' rule, the attribute–class layer can be expressed as $p(u|a)$ is equal to $p(u)/p(a^u)$ if $a = a^u$ and 0 otherwise. Then we can compute the posterior of a test class for a given image by combing the image–attribute layer and the attribute–class layer through the following formulation [9]:

$$p(u|\mathbf{x}) = \sum_{a \in \{0,1\}^M} p(u|\mathbf{x})p(a|\mathbf{x}) = \frac{p(u)}{p(a^u)} \prod_{m=1}^{M} p(a_m^u|\mathbf{x}) \tag{14}$$

If there is no extra prior information, we assume that all class priors $p(u)$ are equal, which allows us to ignore the factor $p(u)$ in the following. We assume the factor $p(a)$ follows a factorial distribution $p(a) = \prod_{m=1}^{M} p(a_m)$. The empirical mean $p(a_m) = (1/K) \sum_{i=1}^{K} a_m^{s_k}$ over all training classes can be used as attribute prior where s_k is the k-th class of the training set. Finally, a test image $\mathbf{z}$ is classified as the best output class from all test classes $u_1, \cdots, u_L$ using the maximum a posteriori (MAP) prediction in the following form:

$$Label(\mathbf{z}) = \underset{i=1,\ldots,L}{\operatorname{argmax}} \prod_{m=1}^{M} \frac{p(a_m^{u_i}|\mathbf{z})}{p(a_m^{u_i})} \tag{15}$$

The DAP model enables us to transfer the attribute–class association of known classes to unseen ones, and facilitates the realization of zero-shot learning system.

4.2. Proposed zero-shot classification Schemes

In Section 3, we have proposed the ARL model to discover the underlying relation between attributes automatically from data. After obtaining such relation, we also want to reuse it to boost the attribute prediction performances of traditional attribute classifiers that do not consider such relation. In this sub-section, based on different ways of utilizing the attribute relation, we propose two types of schemes for attribute-based zero-shot classification under the DAP model, including: (1) ARL (i.e. ARL+DAP) scheme, and (2) ARL+SVM/KRR (i.e. ARL+SVM/KRR+DAP) scheme. The first scheme is to consider the attribute relation during the learning of attribute classifiers. And the second one is to modify the outputs of traditional attribute classifiers using the discovered attribute relation learned from the ARL model.

4.2.1. ARL scheme

The ARL scheme uses our ARL model to learn attribute classifiers, and the DAP technique to realize the attribute-label mapping. Fig. 4 illustrates the overview of this scheme. As illustrated in Fig. 4, we first use all the training images and their corresponding attributes to train M attribute classifiers jointly using the proposed ARL model. Then, the trained classifiers are employed to predict the attribute values of unseen test images, and each test image will be given M attributes of real values. It is worth noting that the trained classifiers have considered the attribute–attribute relation implicitly in the optimization process. With specific inventory of attributes for each object class, the predicted attribute values will be mapped into class labels through DAP technique.

4.2.2. ARL+SVM/KRR scheme

Given the attribute–attribute relation learnt from the proposed ARL model, we also want to reuse it to improve the performance of traditional classifiers that are learned separately, e.g. support vector machine (SVM) and kernel ridge regression (KRR). Hence, the second attribute-based zero-shot classification scheme utilizes SVM and KRR to perform attribute prediction, and uses the proposed ARL model to learn the relation between attributes. Then we modify the predicted attribute values using the learnt attribute relation explicitly, with more details described in the following. Similar to the first scheme, the DAP technique is used to perform the attribute-label mapping. Accordingly, the overview of such scheme is illustrated in Fig. 5.

There might be many ways to incorporate the attribute–attribute relation into the attribute values predicted by traditional classifiers. In this work, we utilize a simple and natural way, with the underlying intuition that attributes can benefit from each other according to their correlation relation. Specifically, the attribute value vector *Score* ($Score \in \mathbb{R}^M$) for a test image predicted by SVM or KRR is modified by the attribute–attribute correlation in the following form:

$$Score^* = (Score^T \times \mathbf{Cor})^T \tag{16}$$

where $Score^* \in \mathbb{R}^M$, and $\mathbf{Cor} \in \mathbb{R}^{M \times M}$ is learnt from our ARL model as described in Algorithm 1. Experiments in Section 5 demonstrate the effectiveness of the proposed method.

5. Experiments

In this section, we perform experiments to evaluate the proposed methods on three challenging data sets (*Animals with Attributes* [9], *aPascal* and *aYahoo* [3]) and our proposed *Texture with Attributes* data set. Two sets of experiments are carried out: (1) attribute prediction experiments, which evaluate the attribute classifiers of our ARL method learnt jointly incorporated with attribute relation; (2) zero-shot classification experiments, which evaluate the performances of proposed two types of zero-shot classification schemes with attribute–attribute relation considered. Section 5.1 introduces the data sets used in the experiments. The experiment setup is

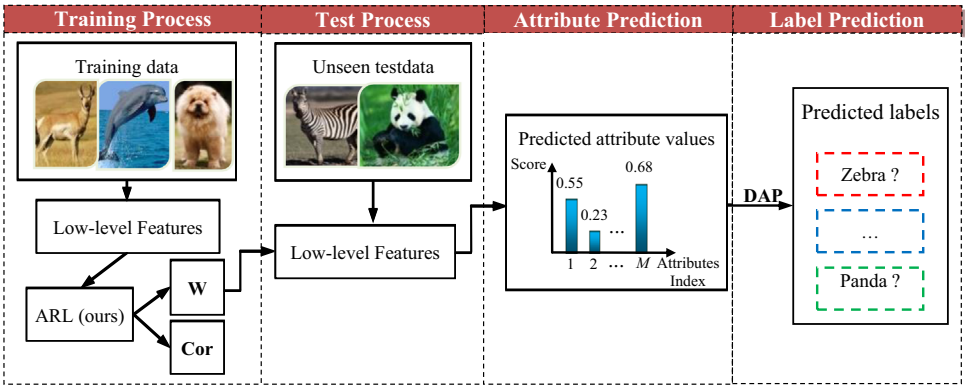


Fig. 4. Overview of the proposed ARL scheme for zero-shot classification.

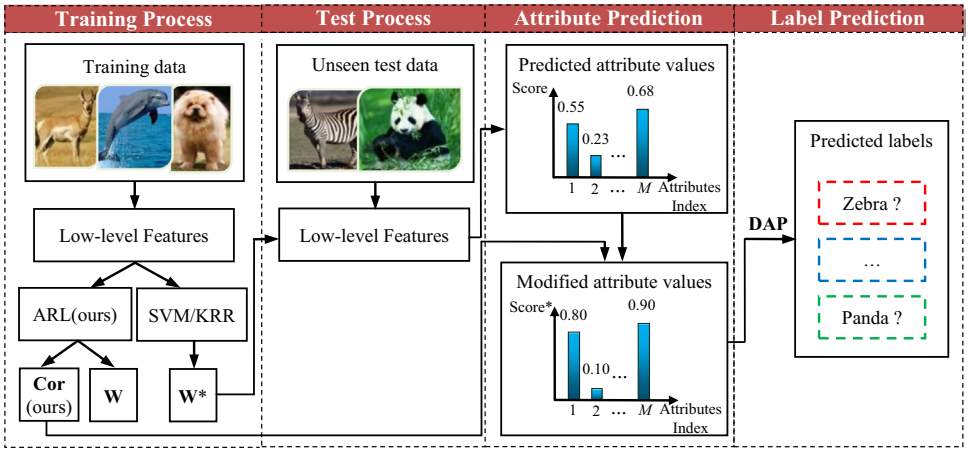


Fig. 5. Overview of the proposed ARL+SVM/KRR scheme for zero-shot classification.

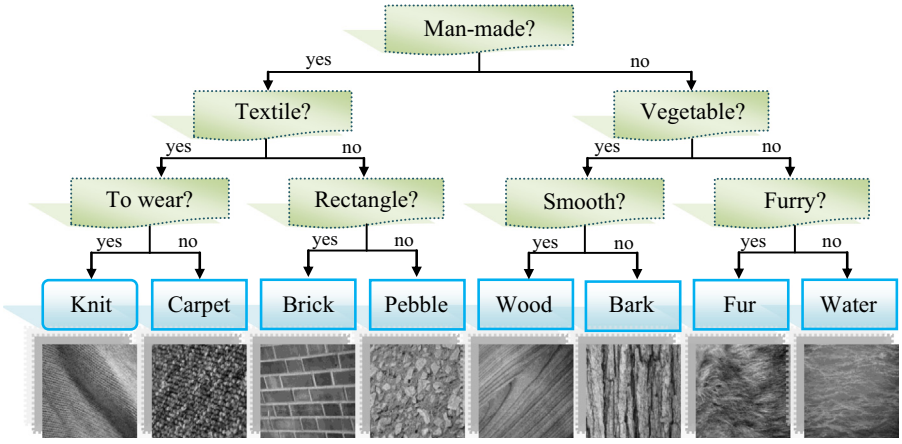


Fig. 6. Attributes for the *Texture with Attributes* data set.

described in Section 5.2. Finally, we report the experimental results in Section 5.3 and Section 5.4.

5.1. Data sets

In this sub-section, we will introduce the data sets with attribute description used in the experiments and their corresponding feature representation.

5.1.1. Animals with Attributes (AWA)

We use a subset of the *Animals with Attributes* (*AWA*) dataset [9], which consists of 30,475 images of 50 animals categories and 85 attributes. These attributes describe properties of animals such as color, patterns, size, anatomy, behavior etc. Since some attributes are not discriminate based on the given features, as is reported in [9], only 59 attributes are employed in our experiments. For computational reasons, we down-sample the training set to 100 images per category as the data set for our experiments, with an exception that "mole" has only 92 images. As given in [9], six types of pre-extracted feature representation, i.e. RGB color histograms (Cq), Sift, rgbShift, Phog, Surf and local self-similarity histograms (LSS) are used as descriptors for these images. And 10 classes are selected as the test set and the other 40 classes as the training set with the same division in [9]. Both attribute prediction and zero-shot classification experiments are performed on this data set.

5.1.2. aPascal and aYahoo

We also use the *aPascal-train* and *aYahoo-test* data sets introduced by Farhadi et al. [3]. The former consists of 20 classes and the latter 12, including animals, vehicles, and household items etc. There are 64 attributes for these two data sets, including shape, textures, atomy and parts. We use color features as image descriptor as is given in [3]. Similarly, we use 100 images for each class to form the data set used in our experiments. Two different groups of these two data sets are used. One employs all samples in *aPascal* as training set and those in *aYahoo* as test set, with "*aPascal* & *aYahoo*" for short. The other uses 5 classes (*bus, car, cat, dog and motorbike*) as test data while the other 15 ones in *aPascal* as training data, with "*aPascal*" for short. Note that instances of a specific class in these two data sets may have different inventories of attributes, i.e., a unique inventory of attributes for each class cannot be obtained. Therefore, we just perform attribute prediction experiments other than zero-shot classification experiments on these two data sets, which are different from that of the *AWA* data set.

5.1.3. Texture with Attributes (TWA)

To facilitate study of a wide range of texture analysis problems, researchers have collected a large number of textures, both in the form of surface textures and natural scenes. Currently, there are several well-known texture data sets, such as *Brodatz* [55], *MeasTex* [56], *VisTex* [57] and *Outex* [58]. Among them it is revealed that surface textures in *Outex* exhibit well defined

variations to a given reference in terms of illumination, rotation and spatial resolution [58]. However, these data were not useable in attribute-based texture classification context so far, because no attributes are defined for each texture category at all. To overcome this problem, we construct the *Texture with Attributes* (*TWA*) data set and perform zero-shot classification experiments on it.

Motivated by the hierarchical structure of the *ImageNet* Large Scale Visual Recognition Competition 2010 (ILSRC10) [59], we design a set of seven semantic or descriptive attributes: "man-made", "textile", "vegetable", "to wear", "rectangle pattern", "smooth" and "furry". The attribute tree is constructed to ensure that each texture class is described by a unique attribute description set, as illustrated in Fig. 6.

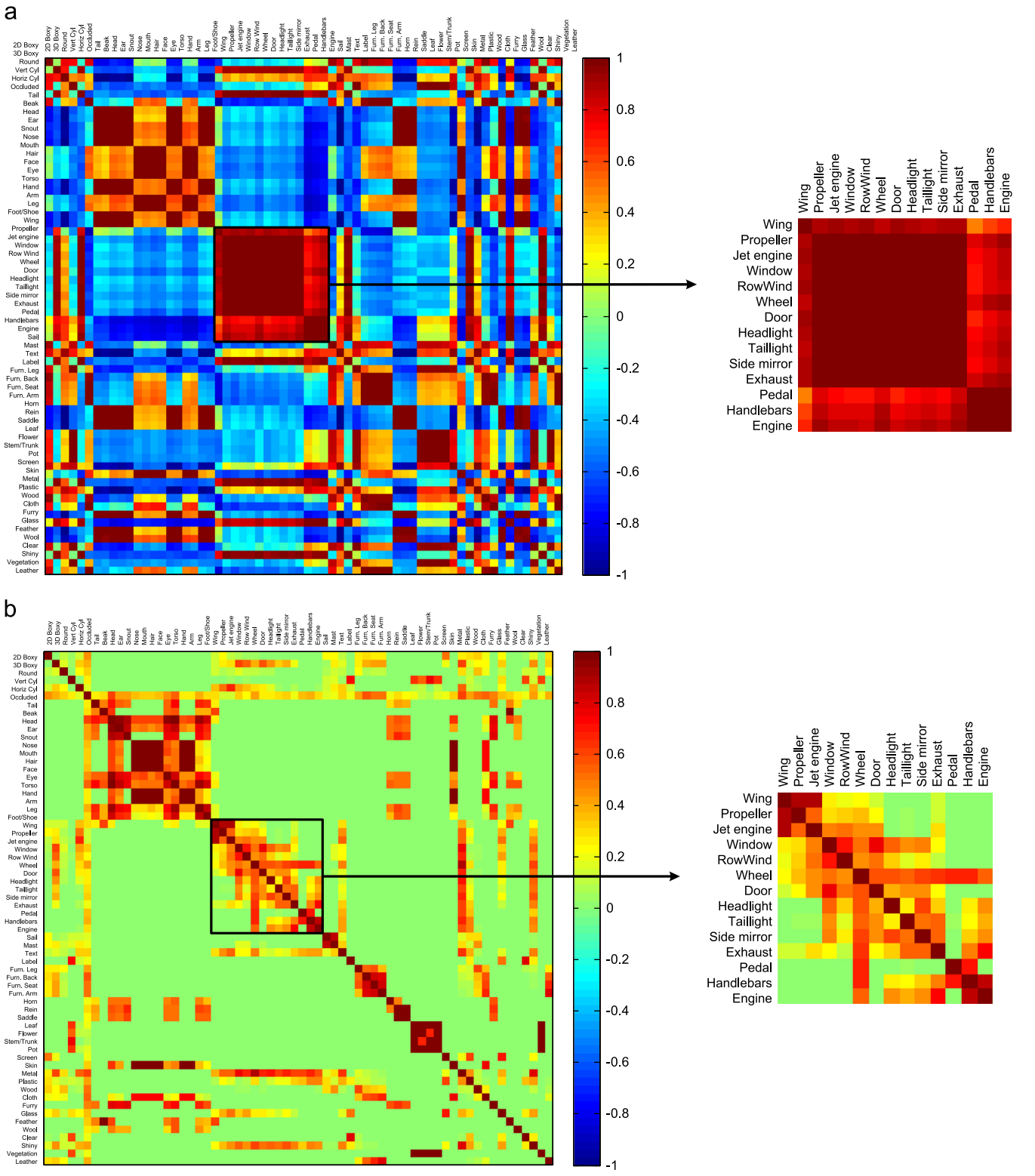


Fig. 7. Attribute correlation coefficient matrix earned by different methods on *aPascal* (a) the proposed ARL and (b) NormMI. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

These attributes generally correspond to basic surface texture commonly found in daily life. The values of these attributes are binary. If a texture class is positive or negative in some attribute, its value is 1 or 0 respectively. Example images of eight texture classes (*Knit, Carpet, Brick, Pebble, Wood, Bark, Fur and Water*) are collected from surface textures in the *Outex* data set, which range from man-made textures to artificial ones. The remaining collection is composed of 320 images with 40 ones for each class. Zero-shot classification experiments are performed on this data set.

For the *AWA, aPascal* and *aYahoo* data sets, features of objects are represented by histograms of local image, which are usually high dimensional. In this paper, we compute kernel matrixes which are based on χ^2-kernel [60] of individual feature type (sum of individual χ^2-kernel when use all feature types), as done in [9]. For the bandwidth parameters of χ^2-kernels, we first compute the median of the χ^2-distances over the training samples, and then set bandwidth as the five times inverse of such median. Among various texture feature extraction methods, filter bank methods are most commonly used, such as wavelet transforms [61], Gabor filters [62] and contoured transforms [63] with the aim to enhance edges and lines of different orientations and scales. For the *TWA* data set, a 4-level decomposition structure for the Wavelet Package Transform [61] is utilized, which is a classical method for texture feature extraction. The mean and the variance of each coefficients matrix in the 4-th level form the 34-dimensional feature vector for each texture images.

5.2. Experimental setup

For the first set of experiments, i.e. attribute prediction, we compare the proposed ARL model with SVM as in [4,9]and KRR that are trained separately without considering any attribute–attribute relation. As for the second one, i.e. zero-shot classification experiments, we compare the proposed ARL scheme, "ARL+SVM" and "ARL+KRR" methods with SVM and KRR schemes, where SVM and KRR schemes use traditional SVM and KRR to train attribute classifiers respectively while DAP is used to perform the attribute–class mapping.

For attribute–attribute relation learning, we also compare the proposed ARL method to normalized mutual information (NormMI), which is a commonly used method [14,19]. Accordingly, we incorporate the attribute relation mined by NormMI into SVM and KRR, and get two zero-shot classification scheme called "NormMI+SVM" and "NormMI+KRR" respectively. In the zero-shot classification experiments, "NormMI+SVM" and "NormMI+KRR" are compared with our proposed ARL, "ARL+SVM" and "ARL+KRR" methods.

For the proposed ARL method, we can select the optimal regularization parameters λ_1 andλ_2 through grid search in the range (0, 1]. For computational reasons, we empirically set the regularization parameters λ_1 and λ_2 of the proposed ARL method as 0.1 and 0.05, respectively. Similar to [9], the parameter C of SVM is set to 10. The parameter λ of KRR is selected from {0.001, 0.01, 0.1, 1, 10, 100} on the validation set of training classes. As in [9], a sigmoid transform [64] maps the outputs of attribute classifiers into probabilistic scores for DAP, where optimal values for parameter A and B are set on the same validation set. We adopt the classification accuracy to evaluate the attribute prediction and the zero-shot object classification performances of different methods.

5.3. Results of attribute prediction experiments

In this sub-section, we first show the attribute correlation learned from our proposed ARL method compared with that of normalized mutual information (NormMI), and then report the performances of the learned attribute classifiers with comparison to those of SVM and KRR.

5.3.1. Learnt attribute relation

Firstly, in Fig. 7, we depict the correlation coefficient matrix among attributes learnt from the proposed ARL method and NormMI on the *aPascal* data set. Note that the attributes of this data set are grouped according to parts of object classes (e.g. people, vehicles and horses). Here, red and yellow indicate high correlation coefficients while blue and green denote low ones.

From Fig. 7(a), one can find that some attributes, such as"Wheel", "Door", "Headlight", "Taillight" and "Side mirror", are highly positively correlated. These attributes are all relevant to the object "vehicles", demonstrating the attribute relation we learn is reliable. However, for the relation learnt by NormMI shown in Fig. 7(b), this trend is not so distinct. In addition, on this data set, the negative correlation between attributes can be reflected by the proposed ARL method, but cannot be reflected by NormMI. In addition, in Section 5.4, we will further investigate whether all kinds of correlation can help improve the performance of attbibute-based zero-shot classification.

In addition, we analyze the convergence of the proposed Algorithm 1. The change of the objective function value defined in Eq. (8) on the *aPascal* data set is given in Fig. 8. From Fig. 8, one can find that objective function value decreases rapidly that takes within 100 iterations. It illustrates the fast convergence of the proposed Algorithm 1. In the following experiments, we set the iteration number as 100 empirically.

5.3.2. Results of attribute prediction

In attribute-based zero-shot classification problems, higher accuracy of attribute prediction in the image–attribute layer will promote the performance of zero-shot classification in the attribute–class layer. Now, we investigate the performance of the proposed ARL method in attribute prediction in comparison with traditional SVM and KRR that consider no attribute relation. Note that individual attribute classifiers are trained on the training set, and are evaluated on the novel test set. The overall average attributes prediction accuracies among all attributes on different

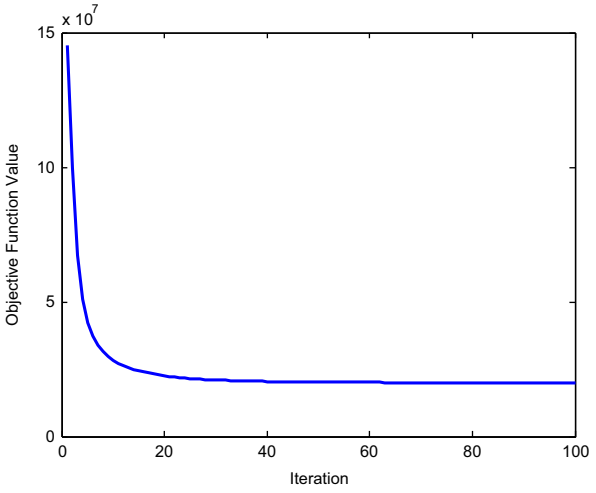


Fig. 8. Convergence of objective function value on *aPascal* data.

Table 1
Average accuracy of different attributes on three data sets (%).

Methods	*aPascal*	*aPascal* & *aYahoo*	*AWA*
SVM	89.67	88.04	75.26
KRR	86.75	85.62	73.04
ARL (ours)	**90.53**	**89.86**	**78.79**

data sets are reported in Table 1, where bold results illustrate the winning method.

From Table 1, one can see that the proposed ARL model consistently achieves better performance comparing to SVM and KRR on three data sets. For example, on the *aPascal* data set, the average attribute prediction accuracy obtained by the proposed ARL method is 90.53%, which is much higher than those of SVM (89.67%) and KRR (86.75%). This validates our intuition that learning the attribute–attribute relation automatically from data can promote the attribute prediction performance.

In addition, we report the prediction accuracies of all attributes on *AWA* data set in Fig. 9. From Fig. 9, one can see that, on most attributes, the proposed ARL method with attribute relation incorporated usually outperforms SVM and KRR that do not considering the relation between attributes. In addition, Fig. 9 shows that attributes such as "white" and "group" cannot generalize well from training classes to novel test ones, while others such as "hand" and "ground" can generalize well. It suggests that some attribute characterization is less informative for novel test classes.

5.4. Results of zero-shot classification

In the following experiments, we first compare the proposed ARL scheme with SVM and KRR in attribute-based zero-shot classification. Then, we incorporate the attribute–attribute relation learnt from the proposed ARL model into SVM and KRR, in comparison with the relation learnt by normalized mutual information (NormMI) as in [19]. The zero-shot classification experiments are performed on the *AWA* and the *TWA* data sets.

5.4.1. Results on AWA data set

Now we first perform zero-shot classification on the *AWA* data set by comparing SVM, KRR and the proposed ARL schemes. Table 2 reports the average classification accuracy of 10 test classes using different types of features.

From the results shown in Table 2, one can see that the proposed ARL method outperform SVM and KRR in predicting novel object classes, with the highest accuracy (37.40%) using all six features. These results further validates that mining the relation between attributes from data indeed boosts the performances of zero-shot object classification tasks.

Furthermore, to boost traditional attribute classifiers that are trained separately without considering the relationship between attributes, we reuse the attribute relation learned from the proposed ARL method to modify the predicted attribute values. To be specific, as mentioned in Section 4.2, the proposed "ARL+SVM" and "ARL+KRR" schemes, which are incorporated by the attribute relation discovered by the proposed ARL model, are compared with "NormMI+SVM" and "NormMI+KRR" where the attribute relation is discovered by normalized mutual information (NormMI). Fig. 10 shows the zero-shot classification accuracy of SVM and KRR as well as their corresponding counterparts with different attribute relation incorporated on *AWA* data set. Note the term "$*_{\_all}$" means all correlation coefficients (positive, negative and uncorrelated ones) are incorporated into the outputs of attribute classifiers, while the term "$*_{\_pos}$" means the negative correlations are ignored.

From Fig. 10(a), one can see that, in most cases, incorporating the positive attribute correlation other than all correlation into traditional SVM usually boost the zero-shot classification results,

Table 2
Zero-shot learning ccuracy on *AWA*(%).

Methods	SVM	KRR	ARL(ours)
Cq	28.84	24.63	**29.98**
Lss	27.37	26.69	**32.84**
Sift	27.82	27.97	**35.24**
rgbSift	24.90	24.62	**33.05**
Phog	27.81	31.53	**33.71**
Surf	29.94	26.60	**35.03**
All features	32.12	30.54	**37.40**

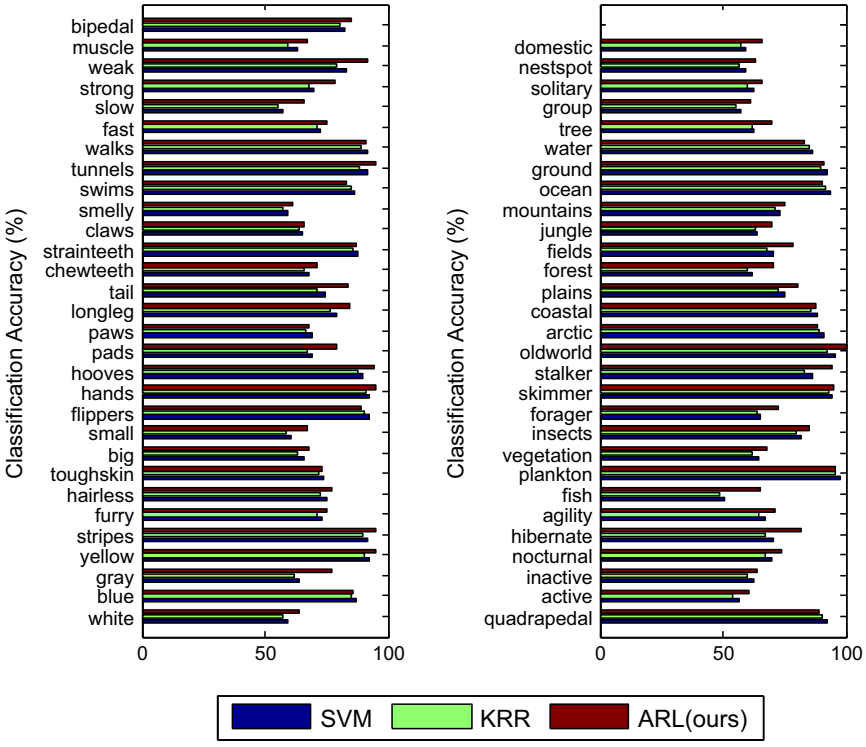


Fig. 9. Attribute prediction accuracy obtained by SVM, KRR and ARL on *AWA*.

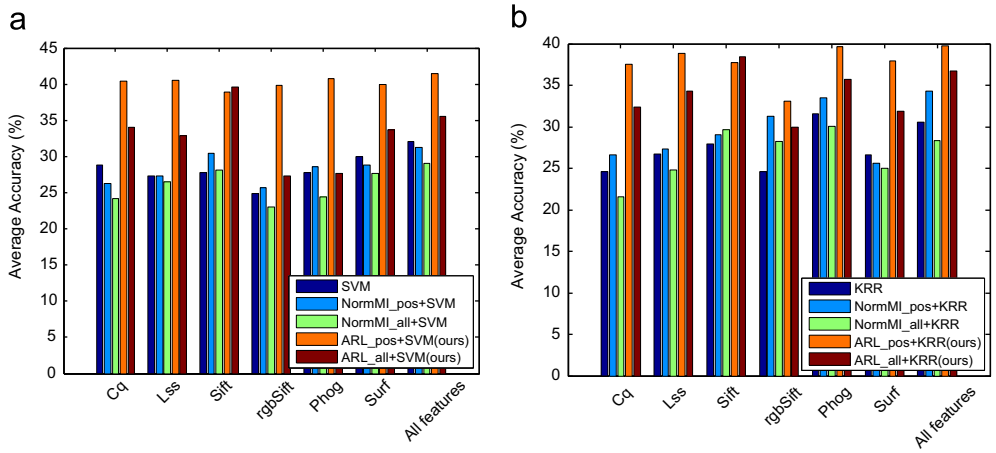


Fig. 10. Zero-shot learning accuracy of SVM and KRR with attribute relation incorporated on *AWA* (a) SVM and (b) KRR.

Table 3
Zero-shot learning ccuracy on *TWA* (%).

Methods	Average accuracy
SVM	62.45
$\text{NormMI}_{\_pos}+\text{SVM}$	60.32
$\text{ARL}_{\_pos}+\text{SVM}$ (ours)	**66.65**
KRR	60.21
$\text{NormMI}_{\_pos}+\text{KRR}$	63.17
$\text{ARL}_{\_pos}+\text{KRR}$ (ours)	**65.05**
ARL (ours)	65.89

whatever the attribute relation is learnt by NormMI and the proposed ARL method. To be specific, the"$\text{ARL}_{\_pos}+\text{SVM}$"scheme usually performs better than SVM and "$\text{ARL}_{\_all}+\text{SVM}$", while "$\text{NormMI}_{\_pos}+\text{SVM}$" usually outperforms SVM and "$\text{NormMI}_{\_all}+\text{SVM}$". On the other hand, one can find that the "$\text{ARL}_{\_pos}+\text{SVM}$" achieves better performance than "$\text{NormMI}_{\_pos}+\text{SVM}$" in most cases, illustrating that the attribute relation learnt from the proposed ARL model is more likely to reflect the true structure among attributes. The similar trend can be found in Fig. 10(b) as in Fig. 10(a), i.e. the proposed "$\text{ARL}_{\_pos}+\text{KRR}$" is superior to the other four methods in most cases. These results validate the effectiveness of the proposed method we incorporate the attribute relation to zero-shot learning in Eq. (15).

5.4.2. Results on TWA data set

Moreover, we perform zero-shot classification experiments on *TWA* data set. In Section 5.4.1, it has been shown that positive attribute relation can largely promote the performances of zero-shot classification comparing to all relations. Hence, we incorporate the positive relation learned by the proposed ARL as well as the NormMI to the "ARL+SVM" and "ARL+KRR" schemes on this data set. Due to the relatively smaller class number of this data set, a leave-two-class-out strategy is employed to perform zero-shot classification experiments. To be specific, each time samples of two classes are selected as the test data while the others are used as the training data, ensuring that the training and the test data are disjoint. Respect to eight texture classes, the number of combination is 28. The average accuracy is computed as the final results, reported in Table 3.

From Table 3, it can be seen that the proposed ARL method considering the attribute relation is superior to SVM and KRR on the *TWA* data set. In addition, the proposed "$\text{ARL}_{\_pos}+\text{SVM}$" method outperforms "$\text{NormMI}_{\_pos}+\text{SVM}$", and the proposed "$\text{ARL}_{\_pos}+\text{KRR}$" outperforms "$\text{NormMI}+\text{KRR}_{\_pos}$", respectively. It demonstrates that the attribute relation learned by the proposed ARL model is more suitable to boost the performances zero-shot classification, compared to that learned by NormMI.

6. Conclusion and future works

In this paper, we mainly focus on learning the relationship between attributes automatically from data for attribute-based zero-shot object classification. We first formulate the attribute–attribute relation learning and the attribute classifiers' training in a unified objective function, which can be casted as a multi-task learning problem. In this way, the attribute–attribute relation can be discovered automatically and explicitly, with no need for external linguistic knowledge base and human intervention. Then, we propose two types of attribute-based zero-shot object classification schemes considering the attribute–attribute relation, where the learnt relation is reused to improve traditional classifiers that are trained separately without considering the underlying the relation between attributes. Finally, we evaluate the proposed methods through attribute prediction experiments and zero-shot classification experiments on three challenging data sets.

In the proposed ARL model, we using the least square loss function and L2-norm regularizer in the objective function. It may be promising to use the other forms of loss function and regularization, which will be one of our future works. Meanwhile, the semantic attribute space can be expanded by generating new attributes based on the attribute relation we learned, which may further boost the zero-shot learning performance. Also it is interesting to investigate whether using preferable attributes for specific categories can improve the performance of zero-shot learning. In addition, feature selection for attribute classifiers may be necessary and meaningful when feature dimension is high, and will also be our future work.

Acknowledgment

The authors would like to thank all reviewers and the associate editor for their constructive comments which significantly improved this paper. This work is supported in part by the Jiangsu Natural Science Foundation for Distinguished Young Scholar (No. BK20130034), the Specialized Research Fund for the Doctoral Program of Higher Education (No. 20123218110009), the NUAA Fundamental Research Funds (No. NE2013105), the Fundamental Research Funds for the Central Universities of China

(No. NZ2013306), the Funding of Jiangsu Innovation Program for Graduate Education (No. CXZZ13_0173), and the National Natural Science Foundation of China (No. 61379015), China.

References

[1] H. Larochelle, D. Erhan, Y. Bengio, Zero-data learning of new tasks, in: Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, Illinois, 2008, pp. 646–651.

[2] M. Palatucci, G. Hinton, D. Pomerleau, T.M. Mitchell, Zero-shot learning with semantic output codes, Advances in Neural Information Processing Systems, Curran Associates Inc., Vancouver B.C., Canada, 2009.

[3] A. Farhadi, I. Endres, D. Hoiem, D. Forsyth, Describing objects by their attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 1778–1785.

[4] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, B. Schiele, What helps where and why? Semantic relatedness for knowledge transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 910–917.

[5] H. Lang, H. Ling, Classifying covert photographs, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 1178–1185.

[6] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Describable visual attributes for face verification and image search\, IEEE Trans. Pattern Analy. Machine Intell. 33 (2011) 1962–1977.

[7] D. Parikh, K. Grauman, Relative attributes, in: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, 2011, pp. 503–510.

[8] M. Rohrbach, M. Stark, B. Schiele, Evaluating knowledge transfer and zero-shot learning in a large-scale setting, in: Proceedings of the *IEEE* Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2011, pp. 1641–1648.

[9] C.H. Lampert, H. Nickisch, S. Harmeling, Learning to detect unseen object classes by between-class attribute transfer, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, 2009, pp. 951–958.

[10] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and smile classifiers for face verification, in: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, 2009, pp. 365–372.

[11] A. Farhadi, I. Endres, D. Hoiem, Attribute-centric recognition for cross-category generalization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, 2010, pp. 2352–2359.

[12] O. Russakovskyl, L. Fei-Fei, Attribute learning in large-scale datasets, in: Workshop on Parts and Attributes,European Conference on Computer Vision, Heraklion, Crete, Greece, 2010, pp. 10–11.

[13] B. Hariharan, S.V.N. Vishwanathan, M. Varma, Efficient max-margin multi-label classification with applications to zero-shot learning, Mach. Learn. 88 (2012) 127–155.

[14] A. Kovashka, S. Vijayanarasimhan, K. Grauman, Actively selecting annotations among objects and attributes, in: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, 2011, pp. 1403–1410.

[15] N. Kumar, P.N. Belhumeur, S.K. Nayar, FaceTracer: a search engine for large collections of images with faces, in: Proceedings of the European Conference on Computer Vision, Marseille, France, 2008, pp. 340–353.

[16] A. Kovashka, D. Parikh, K. Grauman, WhittleSearch: image search with relative attribute feedback, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 1973–2980.

[17] Y.-H. Lei, Y.-Y. Chen, B.-C. Chen, L. Iida, W. Hsu, Where is who: large-scale photo retrieval by facial attributes and canvas layout, in: Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval, Portland, Oregon, USA, 2012, pp. 701–710.

[18] W.J. Scheirer, N. Kumar, P.N. Belhumeur, T.E. Boult, Multi-attribute spaces: calibration for attribute fusion and similarity search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2933–2940.

[19] Y. Wang G. Mori, A discriminative latent model of object classes and attributes, in: Proceedings of the European Conference of Computer Vision, Heraklion, Crete, Greece, 2010, pp. 5–11.

[20] B. Siddiquie, R.S. Feris, L.S. Davis, Image ranking and retrieval based on multi-attribute queries, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2011, pp. 801–808.

[21] D. Mahajan, S. Sellamanickam, V. Nair, A joint learning framework for attribute models and object descriptions, in: Proceedings of the IEEE International Conference on Computer Vision, Barcelona, 2011, pp. 1227–1234.

[22] M. Rohrbach, M. Stark, G. Szarvas, B. Schiele, Combining language sources and robust semantic relatedness for attribute-based knowledge transfer, in: Workshop on Parts and Attributes, European Conference on Computer Vision, Heraklion, Crete, Greece, 2010, pp. 10–11.

[23] M. Fink, Object classification from a single example utilizing class relevance metrics, Advances in Neural Information Processing Systems, MIT Press, British Columbia, Canada, 2004.

[24] E. Bart, S. Ullman, Cross-generalization: learning novel classes from a single example by feature replacement, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005, pp. 672–679.

[25] L. Fei-Fei, R. Fergus, P. Perona, One-shot learning of object categories, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 594–611.

[26] M. Stark, M. Goesele, B. Schiele, A shape-based object class model for knowledge transfer, in: Proceedings of the IEEE International Conference on Computer Vision, Kyoto, 2009, pp. 373–380.

[27] M. Marszałek, C. Schmid, Semantic hierarchies for visual object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, 2007, pp. 1–7.

[28] A. Zweig, D. Weinshall, Exploiting object hierarchy: combining models from different category levels, in: Proceedings of the IEEE International Conference on Computer Vision, Rio de Janeiro, 2007, pp. 1–8.

[29] D. Parikh, A. Kovashka, A. Parkash, K. Grauman, Relative attributes for enhanced human-machine communication, in: Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.

[30] T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.-M. Chang, V.L. Malave, R.A. Mason, M.A. Just, Predicting human brain activity associated with the meanings of nouns, Science 320 (2008) 1191–1195.

[31] V. Ferrari, A. Zisserman, Learning visual attributes, Advances in Neural Information Processing Systems, Curran Associates Inc., Vancouver, B.C., Canada, 2007.

[32] F. Giunchiglia, A.K. Pandey, N. Sebe, J.R.R. Uijlings, J. Stottinger, (Unseen) event recognition via semantic compositionality, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3061–3068.

[33] D. Parikh, K. Grauman, Interactively building a discriminative vocabulary of nameable attributes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2011, pp. 1681–1688.

[34] S.J. Hwang, F. Sha, K. Grauman, Sharing features between objects and their attributes" in: Proocedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2011, pp. 1761–1768.

[35] F.X. Yuy, R. Ji, M.-H. Tsai, G. Yey, S.-F. Changy, Weak attributes for large-scale image retrieval, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2949–2956.

[36] O. Hasegawa, A. Kawewong, S. Tangruamsub, P. Kankuekul, Online incremental attribute-based zero-shot learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 3657–3664.

[37] R. Caruana, Multitask learning, Mach. Learn. 28 (1997) 41–75.

[38] J. Baxter, A Bayesian/information theoretic model of learning to learn via multiple task sampling, Mach. Learn. 28 (1997) 7–39.

[39] S. Thrun, Is learning the n-th thing any easier than learning the first, Advances in Neural Information Processing Systems, MIT Press, 1996.

[40] E. Bonilla, K.M. Chai, C. Williams, Multi-task gaussian process prediction, Advances in Neural Information Processing Systems, Curran Associates Inc., Vancouver, B.C., Canada, 2007.

[41] Y. Zhang, D.-Y. Yeung, A convex formulation for learning task relationships in multi-task learning, in: Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence, 2010.

[42] A.K. Gupta, D.K. Nagar, Matrix Variate Distributions, CRC Press, 1999.

[43] Y. Zhang, D.-Y. Yeung, Multi-task boosting by exploiting task relationships, Machine Learning and Knowledge Discovery in Databases, Springer, Berlin Heidelberg, 2012, pp. 697–710.

[44] H. Fei, J. Huan, Structured feature selection and task relationship inference for multi-task learning, Knowledge and Information Systems 35 (2013) 345–364.

[45] Y. Zhang, D.-Y. Yeung, Probabilistic multi-task feature selection, Advances in Neural Information Processing Systems, Curran Associates Inc., Vancouver, B.C., Canada, 2010.

[46] A. Saha, P. Rai, H. Daumé III, S. Venkatasubramanian, Online learning of multiple tasks and their relationships, in: Proceedings of AISTATS, 2011.

[47] F. Dinuzzo, Learning output kernels for multi-task problems, Neurocomputing 118 (2013) 119–126.

[48] P. Rai, A. Kumar, H. Daumé III, Simultaneously leveraging output and task structures for multiple-output regression, Advances in Neural Information Processing Systems (2012).

[49] Y. Zhang, D.-Y. Yeung, "Transfer metric learning with semi-supervised extension", ACM Trans. Intell. Syst. Technol. 3 (2012) 1–28.

[50] Q. Gu, Z. Li, J. Han, Correlated multi-label feature selection, in: Proceedings of the ACM International Conference on Information and Knowledge Management, Glasgow, Scotland, UK, 2011, pp. 1087–1096.

[51] Y. Guo, D. Schuurmans, Adaptive large margin training for multilabel classification, in: Association for the Advancement of Artificial Intelligence, AI Access Foundation, San Francisco, CA, 2011.

[52] J. Zhang, J. Huan, Inductive multi-task learning with multiple view data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2012, pp. 543–551.

[53] T. Van Gestel, J. Suykens, B. Baesens, S. Viaene, J. Vanthienen, G. Dedene, B. De Moor, J. Vandewalle, Benchmarking least squares support vector machine classifiers, Mach. Learn. 54 (2004) 5–32.

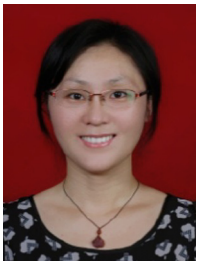
[54] S.S. Keerthi, S.K. Shevade, SMO algorithm for least-squares SVM formulation, Neural Comput. 15 (2003) 487–507.

[55] P. Brodatz, Textures: a Photographic Album for Artists and Designers, Dover Publications, New York, 1996.

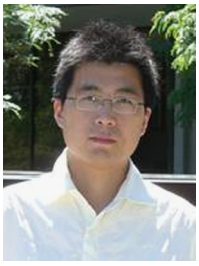
[56] G. Smith, I. Burns, Measuring texture classification algorithms, Pattern Recognit. Lett. 18 (1997) 1495–1501.

[57] V. a. M. G. a. t. M. M. Lab, Vision Texture (Online), Available: at: 〈http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html〉.

[58] T. Ojala, T. Maenpaa, M. Pietikainen, J. Viertola, J. Kyllonen, S. Huovinen, Outex-new framework for empirical evaluation of texture analysis algorithms, in: Proceedings of the International Conference on Pattern Recognition, 2002, pp. 701–706.
[59] A. Berg, J. Deng, F.-F. Li. (2010). ImageNet: Large Scale Visual Recognition Challenge.
[60] O. Teytaud, R. Jalam, Kernel-based text categorization, in: Proceedings of the International Joint Conference on Neural Networks, 2001, pp. 1891–1896.
[61] S. Liapis, N. Alvertos, G. Tziritas, Maximum likelihood texture classification and Bayesian texture segmentation using discrete wavelet frames, in: Proceedings of the International Conference on Digital Signal Processing Proceedings, Santorini, 1997, pp. 1107–1110.
[62] S.E. Grigorescu, N. Petkov, P. Kruizinga, Comparison of texture features based on Gabor filters, IEEE Trans. Image Process. 11 (2002) 1160–1167.
[63] A.L. Da Cunha, J. Zhou, M.N. Do, The nonsubsampled contourlet transform: theory, design, and applications, IEEE Trans. Image Process. 15 (2006) 3089–3101.
[64] J.C. Platt, Probabilistic outputs for support vector machines and comparison to regularized likelihood methods, Adv. Large Margin Classifiers 10 (1999) 61–74.



Mingxia Liu, received the B.S. degree and M.S. degree from Shandong Normal University, Shandong, China, in 2003 and 2006, respectively. In 2006, she joined the School of Information Science and Technology of Taishan University as a Lecturer. She is currently a Ph.D. candidate in Computer Science from Nanjing University of Aeronautics and Astronautics, Nanjing, China. Her research interests include machine learning, pattern recognition, computer vision, and image analysis.



Daoqiang Zhang, received the B.S. degree, and Ph.D. degree in Computer Science from Nanjing University of Aeronautics and Astronautics (NUAA), China, in 1999, and 2004, respectively. He joined the Department of Computer Science and Engineering of NUAA as a Lecturer in 2004, and is a professor at present. His research interests include machine learning, pattern recognition, data mining, and medical image analysis. In these areas, he has published over 100 scientific articles in refereed international journals such as Neuroimage, Pattern Recognition, Artificial Intelligence in Medicine, IEEE Trans. Neural Networks; and conference proceedings such as IJCAI, AAAI, SDM, ICDM. He was nominated for the National Excellent Doctoral Dissertation Award of China in 2006, won the best paper award at the 9th Pacific Rim International Conference on Artificial Intelligence (PRICAI'06), and was the winner of the best paper award honorable mention of Pattern Recognition Journal 2007. He has served as a program committee member for several international and native conferences such as IJCAI, SDM, CIKM, PAKDD, PRICAI, and ACML, etc. He is a member of the Machine Learning Society of the Chinese Association of Artificial Intelligence (CAAI), and the Artificial Intelligence & Pattern Recognition Society of the China Computer Federation (CCF).

Songcan Chen, received the B.Sc. degree in mathematics from Hangzhou University (now merged into Zhejiang University), Zhejiang, China, the M.S. degree in computer applications from Shanghai Jiaotong University, Shanghai, China, and the Ph.D. degree in communication and information systems from the Nanjing University of Aeronautics and Astronautics (NUAA), Nanjing, China, in 1983, 1985, and 1997, respectively. He was at NUAA in January 1986. Since 1998, he has been a full-time Professor with the Department of Computer Science and Engineering, NUAA. He has authored or co-authored over 130 scientific peer-reviewed papers. His current research interests include pattern recognition, machine learning, and neural computing.