

Attribute-Based Transfer Learning for Object Categorization with Zero/One Training Example

Xiaodong Yu and Yiannis Aloimonos

University of Maryland, College Park, MD, USA
xdyu@umiacs.umd.edu, yiannis@cs.umd.edu

Abstract. This paper studies the one-shot and zero-shot learning problems, where each object category has only one training example or has no training example at all. We approach this problem by transferring knowledge from known categories (a.k.a *source categories*) to new categories (a.k.a *target categories*) via object attributes. Object attributes are high level descriptions of object categories, such as color, texture, shape, etc. Since they represent common properties across different categories, they can be used to transfer knowledge from source categories to target categories effectively. Based on this insight, we propose an attribute-based transfer learning framework in this paper. We first build a generative attribute model to learn the probabilistic distributions of image features for each attribute, which we consider as attribute priors. These attribute priors can be used to (1) classify unseen images of target categories (zero-shot learning), or (2) facilitate learning classifiers for target categories when there is only one training examples per target category (one-shot learning). We demonstrate the effectiveness of the proposed approaches using the *Animal with Attributes* data set and show state-of-the-art performance in both zero-shot and one-shot learning tests.

1 Introduction

In this paper, we focus on the one-shot learning [1] and the zero-shot learning [2] of object categories where there is only one training example per category or even no training example. Under these circumstances, conventional learning methods can not function due to the lack of training examples. To solve this problem, knowledge transfer becomes extremely important [3]: by transferring prior knowledge obtained from *source categories* (i.e. known categories) to *target categories* (i.e. unknown categories), we equivalently increase the number of training examples of the target categories. Thus, the difficulties raised by the scarcity of training examples can be greatly alleviated.

This paper present a transfer learning framework that utilizes the semantic knowledge of the object attributes. Object attributes are high-level descriptions about properties of object categories such as color, texture, shape, parts, context, etc. Human beings have a remarkable capability in recognizing unseen objects purely based on object attributes. For example, people who have never seen a zebra still could reliably identify an image of zebra if we tell them that “a zebra

is a wild quadrupedal with distinctive white and black strips living on African savannas”. Since they have prior knowledge about the related object attributes, e.g., *quadrupedal*, *white and black strips*, *African savannas*, they can transfer them to facilitate prediction of unseen categories. The attribute-based transfer learning framework is motivated by this insight. Figure 1 compares different learning process of conventional learning approaches and attribute-based transfer learning approaches: while conventional approaches treat each category individually and train each classifier from scratch, the attribute-based transfer learning approaches can help improve the learning of target classifiers using the attribute prior knowledge learned from source categories. Therefore, we are able to learn target classifiers with much fewer training examples, or even no examples. In the following, we will explore three key components in an attribute-based transfer learning system: attribute models, target classifiers and methods to transfer attribute priors. The main contributions of our paper are:

- 1) We present a generative attribute model that offers flexible representations for attribute knowledge transfer.
- 2) We propose two methods that effectively employ attribute priors in the learning of target classifiers and combine the training examples of target categories when they are available. Thus the attribute priors can help improving performance in both zero-shot and one-shot learning task.
- 3) We show state-of-the-art performance of our transfer learning system on the *Animal with Attributes* [2] data set.



Fig. 1. Comparison of the learning process between conventional learning approaches (a) and attribute-based transfer learning approaches (b)

The rest of this paper is organized as follows: Section 2 discusses the related work; Section 3 describes the attribute model, the target classifier and two approaches of knowledge transfer in details; we present the experimental results in Section 4 and conclude this paper in Section 5.

2 Related Work

Roughly, the methods of knowledge transfer for object categorization can be divided into three groups [3]: knowledge transfer by sharing either *features* [4,5],

model parameters [1,6] or *context information* [7]. Most of the early work relies on bootstrap approaches to select features or parameters to be transferred [4,5,1]. A very recent study [6] suggests that an explicit and controllable transfer of prior knowledge can be achieved by considering the ontological knowledge of object similarity. For example, *horse* and *giraffe* are both quadrupeds and share common topologies, so a full model can be transferred from horse to giraffe. The work presented in this paper integrates a broader ontological knowledge, i.e., object attributes, which can transfer knowledge either among similar categories (e.g., horse and giraffe), or among different categories that share common attributes (e.g., both German shepherd and giant panda have the attribute *black*).

Several recent studies have investigated the approach employing the object attributes in recognition problems [2,8,9,10]. Among them, our work is most related to [2,10]. However, as both studies focused on attribute prediction for zero-shot learning task, they did not attempt to combine attribute priors with the training examples of target categories. Thus, although useful, their applications in one-shot learning task are still limited. Since the framework presented in this paper (Figure 1.b) includes the route for both attribute priors and the training examples of target categories, we can benefit from these two domains whichever is available in learning a new target category. Compared to the existing work in [2,10], our contribution is a more complete framework for attribute-based transfer learning, which enables us to handle both zero-shot learning and one-shot learning problems. The approaches in [8,9] are also related to ours. However, their methods need attributes annotated for each image. Although this type of image-level attribute annotation will benefit intra-class feature selection [8] and object localization [9], it requires substantially human efforts to label each image. Thus their scalability to a large number of categories is greatly restricted compared to the category-level attribute annotations advocated in [2,10] and this paper.

3 Algorithms

3.1 Background

In the proposed approaches, the category-attribute relationship is represented by a category-attribute matrix \mathcal{M} , where the entry at the m -th row and the ℓ -th column is a binary value indicating whether category m has the ℓ -th attribute. Figure 3.a illustrates an example of \mathcal{M} . Each object category thus has a list of attributes whose corresponding values in \mathcal{M} equal to “yes”. Given an object category, the list of associated attributes \mathbf{a} is deterministic. Take the category *cow* in Figure 3 for example, we have $\mathbf{a} = \{black, white, brown, spots, furry, domestic\}$. This information is supposed to be available for both source categories and target categories.

In our approach, the attribute model and the target classifier belong to an extension of topic models, which constitute an active research area in the machine learning community in recent years [11,12,13]. Computer vision researchers have extended them to deal with various vision problems [14,15,16,17]. In a topic

model, a document \mathbf{w} is modeled by a mixture of topics, z 's, and each topic z is represented by a probability distribution of words, w 's. In the computer vision domain, a quantized image feature is often analogous to a word (a.k.a “visual words” [14]), a group of co-occurred image features to a topic (a.k.a “theme” [17]), and an image to a document. In Section 4, we will visualize visual words and topics using examples in the test data set. In this paper, we use the bag-of-features image representation [18]: the spatial information of image features is discarded, and an image is represented as a collection of orderless visual words.

3.2 Attribute Model and Target Classifier

The attribute model we employed is the Author-Topic (AT) model (Figure 2.a) [13]. The AT model is originally designed to model the interests of authors from a given document corpus. In this paper, we extend the AT model to describe the distribution of image features related to attributes. To our best knowledge, this is the first attempt of this kind. Indeed, authors of a document and attributes of an object category have many similarities, which allow us to analogize the latter to the former: a document can have multiple authors and an object category can have multiple attributes; an author can write multiple documents and an attribute can be presented in multiple object categories. Nevertheless, there is also noticeable difference between them: each document can have a distinct list of authors, while all images within an object category share a common list of attributes.

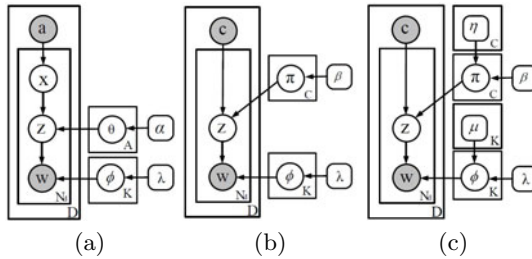


Fig. 2. Graphical representations of the Author-Topic (AT) model (a), the Category-Topic (CT) model (b) and the CT model with informative Dirichlet priors over π and ϕ (c). See text for detailed discussions of these models.

The AT model is a generative model. In this model, an image j has a list of attributes, denoted by \mathbf{a}_j . An attribute ℓ in \mathbf{a}_j is modeled by a discrete distribution of K topics, which parameterized by a K -dim vector $\theta_\ell = (\theta_{\ell 1}, \dots, \theta_{\ell K})$ with topic k receiving weight $\theta_{\ell k}$. The topic k is modeled by a discrete distribution of W codewords in the lexicon, which is parameterized by a W -dim vector $\phi_k = (\phi_{k1}, \dots, \phi_{kW})$ with codeword v receiving weight ϕ_{kv} . Symmetric Dirichlet priors are placed on θ and ϕ , with $\theta_\ell \sim \text{Dirichlet}(\alpha)$, and $\phi_k \sim \text{Dirichlet}(\lambda)$, where α and λ are hyperparameters that affect the sparsity of these distributions. The generative process is outlined in Algorithm 1.

Algorithm 1. The generative process of the Author-Topic model

-
- 1: given the attribute list \mathbf{a}_j and the desired number of visual words in image j , N_j
 - 2: **for** $i = 1$ to N_j **do**
 - 3: conditioning on \mathbf{a}_j , choose an attribute $x_{ji} \sim \text{Uniform}(\mathbf{a}_j)$
 - 4: conditioning on x_{ji} , choose a topic $z_{ji} \sim \text{Discrete}(\theta_{x_{ji}})$, where θ_ℓ defines the distribution of topics for attribute $x = \ell$
 - 5: conditioning on z_{ji} , choose a visual word $w_{ji} \sim \text{Discrete}(\phi_{z_{ji}})$, where ϕ_k defines the distribution of visual words for topic $z = k$
 - 6: **end for**
-

Given a training corpus, the goal of inference in an AT model is to identify the values of ϕ and θ . In [13], Rosen-Zvi et al. presented a collapsed block Gibbs sampling method. The ‘‘collapse’’ means that the parameters ϕ and θ are analytically integrated out, and the ‘‘block’’ means that we draw the pair of (x_{ji}, z_{ji}) together. The pair of (x_{ji}, z_{ji}) is drawn according to the following conditional distribution

$$p(x_{ji} = \ell, z_{ji} = k | w_{ji} = v, \Omega) \propto \frac{\alpha/K + N_{\ell, \setminus ji}^k}{\alpha + \sum_{k'=1}^K N_{\ell, \setminus ji}^{k'}} \frac{\lambda/W + C_{k, \setminus ji}^v}{\lambda + \sum_{v'=1}^W C_{k, \setminus ji}^{v'}}, \quad (1)$$

where $\Omega \equiv \{\mathbf{a}_j, \mathbf{z}_{\setminus ji}, \mathbf{x}_{\setminus ji}, \mathbf{w}_{\setminus ji}, \alpha, \lambda\}$, the subscript ji represents the i -th visual word in image j , $x_{ji} = \ell$ and $z_{ji} = k$ represent the assignments of current visual word to attribute ℓ and topic k respectively, $w_{ji} = v$ represents the observation that the current visual word is the v -th codeword in the lexicon, $\mathbf{z}_{\setminus ji}$ and $\mathbf{x}_{\setminus ji}$ represent all topic and attribute assignments in the training corpus excluding the current visual word, $N_{\ell, \setminus ji}^k$ is the total number of visual words that are assigned to attribute ℓ and topic k , excluding w_{ji} , and $C_{k, \setminus ji}^v$ is the total number of visual words with value v that are assigned to topic k , excluding w_{ji} .

To run the Gibbs sampling algorithm, we first initialize \mathbf{x} and \mathbf{z} with random assignments. In each Gibbs sampling iteration, we draw samples of x_{ji} and z_{ji} for all visual words in the training corpus according to the distribution in Equation (1) in a randomly permuted order of i and j . The samples of \mathbf{x} and \mathbf{z} are recorded after the burn-in period. In experiments, we observe 200 iterations are sufficient for the sampler to be stable. The posterior means of θ and ϕ can then be estimated using the recorded samples as follows:

$$\hat{\theta}_{\ell k} = \frac{\alpha/K + N_{\ell}^k}{\alpha + \sum_{k'=1}^K N_{\ell}^{k'}}, \quad \hat{\phi}_{kv} = \frac{\lambda/W + C_k^v}{\lambda + \sum_{v'=1}^W C_k^{v'}}, \quad (2)$$

where N_{ℓ}^k and C_k^v are defined in a similar fashion as in Equation (1), but without excluding the instance indexed by ji .

If there is only one attribute in each image and the attribute is the object category label, the AT model can be used in object categorization problems [16]. In this paper, we call this approach Category-Topic (CT) model (Figure 2.b) and use it as the target classifier in the proposed transfer learning framework.

It worth to note that the proposed transfer learning framework as illustrated in Figure 1.b is an open framework in that we can also employ other type of attribute models and target classifiers. For example, we evaluate SVM as a target classifier in this paper. Nevertheless, our experiments show that the CT model can outperform discriminative classifiers such as SVM by a large margin.

The inference of a CT model can be performed in a similar way to the AT model. In the Gibbs sampling, we draw samples z_{ji} according to the following conditional distribution

$$p(z_{ji} = k | w_{ji} = v, c_j = m, \Omega) \propto \frac{\beta/K + M_{m, \setminus ji}^k}{\beta + \sum_{k'=1}^K M_{m, \setminus ji}^{k'}} \frac{\lambda/W + C_{k, \setminus ji}^v}{\lambda + \sum_{v'=1}^W C_{k, \setminus ji}^{v'}}, \quad (3)$$

where $\Omega \equiv \{\mathbf{z}_{\setminus ji}, \mathbf{w}_{\setminus ji}, \beta, \lambda\}$, $M_{m, \setminus ji}^k$ is the number of visual words in images of category m assigned to topic k , excluding the current instance. The posterior mean of π can be estimated as follows:

$$\hat{\pi}_{mk} = \frac{\beta/K + M_m^k}{\beta + \sum_{k'=1}^K M_m^{k'}}, \quad (4)$$

and the posterior mean of ϕ is the same as in Equation (2).

After learning a CT model, we can use it to classify a test image $\mathbf{w}_t = \{w_{t1}, \dots, w_{tN_t}\}$ by choosing the target classifier that yields the highest likelihood, where the likelihood for category $c = m$ is estimated as

$$p(\mathbf{w}_t | c = m, \mathcal{D}^{\text{train}}) \approx \prod_{i=1}^{N_t} \sum_{k=1}^K \hat{\phi}_{kw_{ti}} \hat{\pi}_{mk}. \quad (5)$$

If the attribute list is unique in each category, an AT model can also be used to classify a new image by the maximum likelihood criterion. Suppose we have learned θ_ℓ for every $\ell = 1, \dots, A$ from the source categories, we can then use them in classifying an image of a target category using the approximate likelihood

$$p(\mathbf{w}_t | c = m, \mathbf{a}_m, \mathcal{D}^{\text{train}}) \approx \prod_{i=1}^{N_t} \sum_{k=1}^K \hat{\phi}_{kw_{ti}} \left(\frac{1}{A_m} \sum_{\ell \in \mathbf{a}_m} \hat{\theta}_{\ell k} \right) \equiv \prod_{i=1}^{N_t} \sum_{k=1}^K \hat{\phi}_{kw_{ti}} \tilde{\pi}_{mk}, \quad (6)$$

where \mathbf{a}_m is the attribute list associated to a target category $c = m$, A_m the length of \mathbf{a}_m . In the above equations, we have constructed a pseudo weight for the category-specified topic distribution of a new category from $\hat{\theta}_\ell$, i.e., $\tilde{\pi}_{mk} \equiv \left(\frac{1}{A_m} \sum_{\ell \in \mathbf{a}_m} \hat{\theta}_{\ell k} \right)$. This pseudo weight can be viewed as the prior of π_m before we see the real training examples of the new category. Although the unique-attribute-list assumption does not hold in general, it is necessary for attribute-only classifiers, including the AT model discussed in this paper and the approaches in [2,8], to predict unseen categories. The data set tested in this paper satisfies this assumption.

While the AT model can be used to deal with the zero-shot learning problem, it is ineffective for the one-shot learning problem. One may conjecture to add the

training examples of target categories to those of source categories and then re-train the AT model. However, this naive approach will not work well in practice because the number of training examples of source categories is usually higher than the one of target categories by several orders. Consequently the AT model can not well represent the new observations in the training examples of target categories. Thus we need approaches to control the balance between the prior information from source categories and the new information in target categories. We will propose two approaches to achieve this goal in the rest of this section.

3.3 Knowledge Transfer by Synthesis of Training Examples

The first knowledge transfer approach is to synthesize training example for target categories. The idea is as follows: first, we learn the attribute model from the training examples of the source categories; second, for each target category, we run the generative process in Algorithm 1 to produce S synthesized training examples using the estimated $\hat{\theta}$ and $\hat{\phi}$ as well as the attribute list associated to this target category. Each synthesized training example contains \bar{N} visual words, where \bar{N} is the mean number of visual words per image in the source categories. In this procedure, the number of synthesized training example, S , represent our confidence about the attribute priors. We can use it to adjust the balance between the attribute priors and new observations from the training images of target categories.

Since we adopt the bag-of-features representation, the synthesized example is actually composed of a set of image features without spatial information. So they are indeed “artificial” examples in that we can not visualize them like a real image. This is different from the image synthesis approaches in the literature [19,20], which output viewable images. Nevertheless, since our goal is to generate training examples for the target categories to assist the learning process, this is not an issue providing the classifiers take these bag-of-features as inputs.

3.4 Knowledge Transfer by Informative Parameter Priors

The second knowledge transfer approach is to give parameters of the CT model in the target classifiers informative priors. Figure 2.c illustrates the complete CT model, where π and ϕ are given Dirichlet distributions as priors. In these Dirichlet distributions, μ and η are base measurements that represent the mean of ϕ and π , and λ and β are scaling parameters that control the sparsity of the samples drawn from the Dirichlet distribution. When we have no clue about the prior of ϕ and π , we usually give symmetric Dirichlet priors, whose base measures are uniform distributions. The graphical representations of CT models often neglect such uniform distributed base measures and only retain the scaling parameters λ and β , as shown in Figure 2.b. This rule also applies to the AT model. In this paper, these scaling parameters are given vague values when doing Gibbs sampling, $\lambda = W$, $\alpha = \beta = K$.

However, after we learn the attribute model from source categories, our uncertainty about the ϕ and π of target categories will be greatly reduced. Our

knowledge on these parameters are represented by the estimated $\hat{\phi}$ in Equation (2) and $\tilde{\pi}$ in Equation (6). Since $E(\phi_k) = \mu_k$ and $E(\pi_m) = \eta_m$, now we can give informative priors to ϕ and π by setting $\mu_k = \hat{\phi}_k$ and $\eta_m = \tilde{\pi}_m$. The basic equation of Gibbs sampling of the CT model with informative prior the becomes

$$p(z_{ji} = k | w_{ji} = v, \Omega) \propto \frac{\beta \tilde{\pi}_{mk} + M_{m, \setminus ji}^k}{\beta + \sum_{k'=1}^K M_{m, \setminus ji}^{k'}} \frac{\lambda \hat{\phi}_{kv} + C_{k, \setminus ji}^v}{\lambda + \sum_{v'=1}^W C_{k, \setminus ji}^{v'}}, \quad (7)$$

where $\Omega \equiv \{c_j = m, \mathbf{z}_{\setminus ji}, \mathbf{w}_{\setminus ji}, \beta\eta, \lambda\mu\}$. The posterior means of π and ϕ in Equation (4) and (2) are updated accordingly. The value of λ and β represent our confidence on these priors, which can be used to control the balance between attribute priors and the new observations of training images of target categories. In the experiments, we set $\lambda = \beta = \bar{N}S$, where \bar{N} and S are defined as in Section 3.3.

By comparing Equation (7) and Equation (3), we can appreciate the importance of informative priors for the zero-shot learning task. If we have no prior knowledge about π , we can only give it a symmetric Dirichlet prior where $\eta_{mk} = 1/K$. In this scenario, the CT model have to see some training examples of target categories; otherwise, π_{mk} will be assigned to vague value $1/K$, which is useless for categorization tasks. Thus the CT model can not be used in zero-shot learning task. With the attribute knowledge, we can give π informative priors $\eta_{mk} = \tilde{\pi}_{mk}$, which permits us to perform zero-shot learning task using the CT model. Similar impact of the informative priors can be observed in the one-shot learning task.

4 Experiments

4.1 Data Set and Image Features

In the experiments, we use the ‘‘Animals with Attributes’’ (AwA) data set described in [2]. This data set includes 30475 images from 50 animal categories, and 85 attributes to describe these categories. The category-attribute relationship is labeled by human subjects and presented in a 50×85 matrix \mathcal{M} . Figure 3.a illustrates a subset of this matrix. 40 categories are selected as source categories and the rest 10 categories are used as target categories. The division of source and target categories is the same as in [2]. The 85 attributes can be informally divided into two groups: visual attributes such as *black*, *furry*, *big*, *arctic*, etc., and non-visual attributes such as *fast*, *weak*, *fierce*, *domestic*, etc. Totally there are 38 non-visual attributes (attribute No.34 to No.64 and attribute No.79 to No.85) and 47 visual attributes. While non-visual attributes are not directly linked to visual features, it turns out that the non-visual attributes have strong correlation to the visual attributes, as shown in Figure 3.b. Take the attribute *fast* as an example, the top three most related visual attributes are *furry* ($P(\text{furry}|\text{fast}) = 0.833$), *tail* ($P(\text{tail}|\text{fast}) = 0.833$) and *ground* ($P(\text{ground}|\text{fast}) = 0.786$).

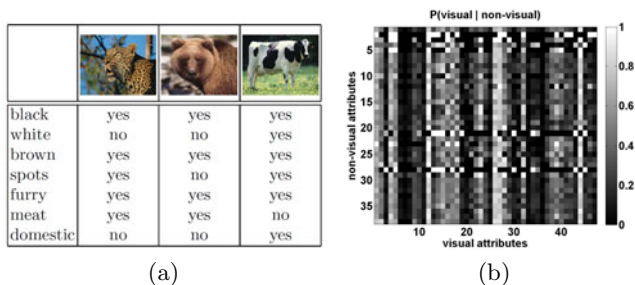


Fig. 3. (a): examples of ontological knowledge represented by the binary category-attribute values; (b): the probability of nonvisual attributes conditioned on visual attributes measured by $P(\text{visual}|\text{non-visual}) \equiv N(\text{visual}, \text{non-visual})/N(\text{non-visual})$, where $N(\cdot)$ denote the number of categories that have the particular attributes in the given data set. Images and attributes are from the “Animals with Attributes” data set [2].

All images are resized such that the longest side has 300 pixels. From each image, we extract four types of image features: SIFT [21], rgSIFT [22], local color histogram and local self-similarity histogram (LSS) [23]. Then for each type of feature, we build a visual lexicon of size 1000 by applying K-means clustering algorithms over features from 250 images randomly selected from source categories. Codewords from four type of features are combined into a single lexicon with 4000 codewords. Features in all images are quantized into one of the codewords in this lexicon. On average, there are about 5000 features in each image. So we set $\bar{N} = 5000$ in the approaches of attribute knowledge transfer in Section 3.3 and Section 3.4. In [2], color histogram (CH) and PHOG features are also extracted from 21 cells of a 3-level spatial pyramids. In our experiments, we did not use these features because the topic model can not discover sensible patterns of co-occurrence of CH/PHOG from the sparse 21 CH/PHOG features in each image.

4.2 Experiment Setup and Implementation Details

Baseline Algorithms. In the experiments, we use Direct Attribute Prediction (DAP) [2] and SVM as baselines in the zero/one-shot learning tasks.

The DAP is selected as a baseline because it is the state-of-the-art approach for zero-shot learning on the AWA data set. DAP uses a SVM classifier that is trained from source categories to predict the presence of each attribute in the images of target categories. Then the attribute predictions are combined into a category label prediction in an MAP formulation. The original DAP can only perform zero-shot learning. For one-shot learning, we use predicted attributes as features and choose a 1NN classifier following the idea in [8]. We call this classifier as “DAP+NN” in this paper.

When we use the synthesized training examples to transfer attribute knowledge, many existing classifiers can be used as the target classifier. We choose SVM as a baseline in this case, mainly because SVM is one of the state-of-the-art classifiers with bag-of-features image representation [24].

Implementation Details. The AT model has $K_0 = 10$ unshared topics per attribute in all tests. When using synthesized training examples, the CT model has 100 topics; when using informative priors, the number of topics in the CT model is the same as the total topics in the AT model. The SVM in the target classifiers is implemented using the C-SVC in LIBSVM with a χ^2 kernel. The kernel bandwidth and the parameter C are obtained by cross-validation on a subset of the source categories.

Evaluation Methodology. In the zero-shot learning scenario, both AT and DAP are trained using the first 100 images of each source category. Then we use the AT model to generate $S = \{10, 20, 100\}$ synthesized examples for each target category. The CT and SVM classifiers will be trained using these synthesized examples. We denote them as “CT+S” and “SVM+S” respectively in the reported results. Also we use the learned $\hat{\phi}$ and $\tilde{\pi}$ in the AT model as informative priors for the CT model as described in Section 3.4, where we set $S = \{2, 5, 10\}$. We denote it as “CT+P” in the reported results.

In the one-shot learning scenario, CT and SVM classifiers are trained with the synthesized training examples/informative priors obtained in the zero-shot learning test plus the first $M = \{1, 5, 10\}$ images of each target category. The AT model is trained with the first 100 images of each source category plus the first M images of each target category. DAP+NN uses the attribute predictions of the first M images of each target category as training data points to classify new images of target categories based on the nearest neighbor criterion.

In both zero-shot and one-shot learning tests, all classifiers are tested over the last 100 images of each target category and the mean of the diagonal of the confusion matrix is reported as the measurement of performance.

4.3 Results

Test 1: Overall Performance of Zero/One-Shot Learning. The overall performance of zero/one-shot learning are presented in the top row of Figure 4. These results show that the proposed approach outperforms the baseline algorithms in the following three aspects:

1. *We have proposed a better attribute model for knowledge transfer.* In both zero/one-shot learning tests, the AT model surpasses DAP and DAP+NN by 5.9% to 7.9%. Furthermore, all target classifiers that employ the prior knowledge from the AT model (SVM+S, CT+S and CT+P) achieve higher accuracy than DAP and DAP+NN. These results clearly show the advantages of the AT model in the attribute-based transfer learning framework.

2. *We have proposed better methods of knowledge transfer for one-shot learning.* In the one-shot learning test, the performance of the AT model does not improve compared to the zero-shot learning test. It is not a surprise: there are total 4000 images of source categories while only 10 images of target categories in training the AT model, thus the learned AT model will be almost the same as the one trained only with the 4000 source images. This result shows that the naive method of knowledge transfer will not work for the one-shot learning

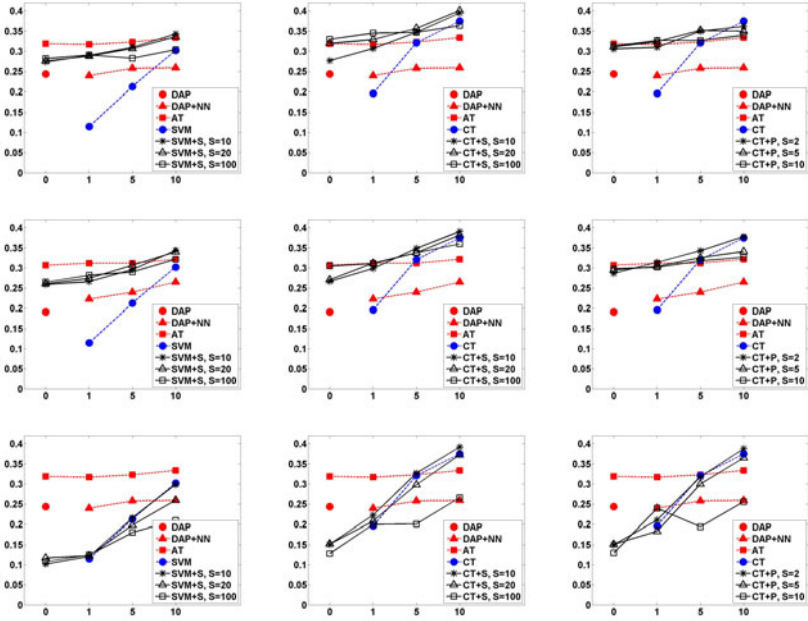


Fig. 4. Results of zero-shot and one-shot learning in Test 1 (top row, using all attributes), Test 2 (middle row, using visual attributes only) and Test 3 (bottom row, using randomly selected attributes) for SVM+S (column 1), CT+S (column 2) and CT+P (column 3) respectively. The x-axis represents the number of real examples, M , and the y-axis represents the mean classification accuracy, i.e., the mean of the diagonal of the confusion matrix.

task. The proposed CT+S and CT+P approaches achieve better balance between the prior attribute knowledge and the real example of target categories, and the additional single training example improves their accuracies by 0.9%-3% (CT+S) and 0.4%-1.4% (CT+P) respectively compared to their zero-shot learning results.

3. *We have proposed a better target classifier.* In both the zero-shot and one-shot learning tasks, the CT models (CT+S and CT+P) consistently exceed the baseline SVM classifier and thus the advantage of CT over SVM in the zero/one-shot learning tasks is confirmed.

In addition to the above conclusion, we also have the following observations.

4. *CT+S generally outperforms CT+P.* CT+P can be viewed as an online version of CT+S, where the informative priors are equivalent to the initial values estimated from the synthesized examples in the initialization stage. Thus, samples drawn with CT+P are not distributed according to the true posterior distribution $P(z_{ji}|z_{\setminus ji}, \mathbf{w})$, which includes all the synthesized and real training examples. As a result, the categorization performance is degraded.

5. *With the increasing number of real training examples, the improvement on classification due to the prior knowledge decreases accordingly.* This suggests that

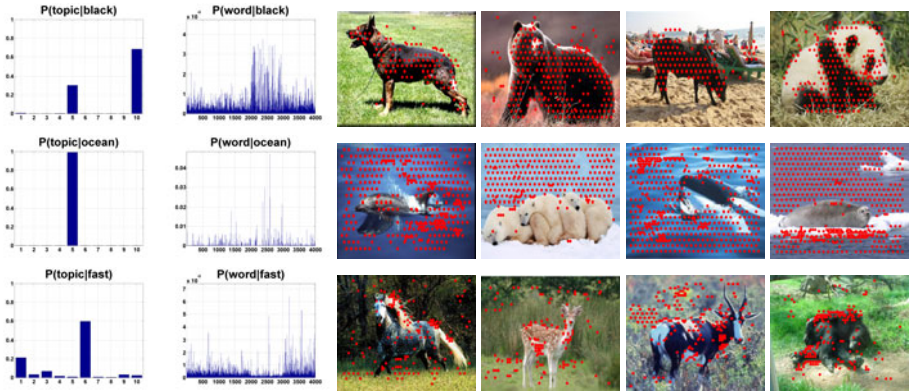


Fig. 5. Illustrations of three attribute models for *black*, *ocean* and *fast* from the top to the bottom. Column 1: the distribution of the 10 topics assigned to a particular attribute; Column 2: the distribution of codewords for a particular attribute; Column 3-6: examples of images from source categories (Column 3-5) and target categories (Column 6), superposed with the top 100 most likely codewords (solid red dots) for the attributes of the same row. Figures are best viewed in color.

the attributes do not contain all the information in target categories. Furthermore, some attributes may be difficult to learn and some are less informative to the categories. Thus when we have sufficient number of real training examples, the prior knowledge behaves more and more like noise and inevitably degrades the classification performance. We can thus derive a practical guideline from this observation to select an appropriate parameter S : when there is no or only one real training example, we can set a large value of S , e.g., 100; when more and more real training examples are available, we then gradually reduce the value of S to zero.

Illustrations of the Attribute Models. We show three attribute models for *black*, *ocean* and *fast* in Figure 5. Though we employ the bag-of-features image representation and discard the spatial information in the image representation, the visual features related to two visual attributes, *black* and *ocean*, roughly localize the regions of interest. As discussed in Section 4.1, the non-visual attribute, *fast*, is most correlated to visual attributes *furry*, *tail* and *ground*. So the visual features related to these visual attributes are implicitly linked to *fast*. Visual examples in Figure 5 support this assumption. The influence of the non-visual attributes on the classification performance will be evaluated quantitatively in Test 2.

Test 2: The Influence of the Non-visual Attributes in the Transfer Learning. In this experiment, we remove the non-visual attributes from the class-attribute matrix and repeat the above tests. Results are illustrated in the middle row of Figure 4. Clearly, the absence of non-visual attributes degrades the classification performance enormously for all classifiers in both zero-shot and

one-shot learning scenarios. This test illustrates the importance of the non-visual attributes in the transfer learning approaches.

Test 3: The Effectiveness of the Knowledge of Attribute in the Transfer Learning. In this experiment, we use the AT model learned from source categories to generate synthesized training examples or compute informative priors following **randomly selected** attributes for each target category, where the number of random attributes are the same as that of the true attributes in each target category. The results show that the classification performance is at the chance level in the zero-shot learning tasks. In the one-shot learning task, the prior knowledge from the randomly selected attributes does not improve the classification performance compared to those not using attribute priors. This experiment highlights the effectiveness of the knowledge of the attribute.

5 Conclusion and Future Work

In this paper, we proposed a transfer learning framework that employs object attributes to aid the learning of new categories with few training examples. We explore a generative model to describe the attribute-specified distributions of image features and two methods to transfer attribute priors from source categories to target categories. Experimental results show that the proposed approaches achieve state-of-the-art performance in both zero-shot and one-shot learning tests.

There are several areas to improve this work. First, we will evaluate our approaches using more data sets in the future, especially the FaceTracer data set [10] and the PASCAL+Yahoo data set [10]. We will also compare the attribute-based transfer learning approaches to those not using attributes, such as [4,5,1]. Second, we employ the bag-of-features image representation in this work, which discards valuable spatial information. In the future work, we will enhance the current model by including spatial constraints, such as regions [15] or vicinity [16]. By this way, we can localize attributes more accurately and subsequently improve the categorization performance. Finally, it would be highly valuable to formally study the influence of different visual attributes and select informative attributes for particular categories.

Acknowledgement

The support of the Cognitive Systems program (under project POETICON) is gratefully acknowledged.

References

1. Fei-Fei, L., Fergus, R., Perona, P.: One-Shot Learning of Object Categories. PAMI 28, 594–611 (2006)
2. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning To Detect Unseen Object Classes by Between-Class Attribute Transfer. In: CVPR (2009)

3. Fei-Fei, L.: Knowledge Transfer in Learning to Recognize Visual Object Classes. In: International Conference on Development and Learning (2006)
4. Bart, E., Ullman, S.: Cross-Generalization: Learning Novel Classes from a Single Example by Feature Replacement. In: CVPR, pp. 672–679 (2005)
5. Torralba, A., Murphy, K.P., Freeman, W.T.: Sharing Features: Efficient Boosting Procedures for Multiclass Object Detection. In: CVPR, vol. 2, pp. 762–769 (2004)
6. Stark, M., Goesele, M., Schiele, B.: A Shape-Based Object Class Model for Knowledge Transfer. In: ICCV (2009)
7. Murphy, K., Torralba, A., Freeman, W.T.: Using the Forest to See the Trees: A Graphical Model Relating Features, Objects, and Scenes. In: NIPS (2003)
8. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing Objects by Their Attributes. In: CVPR (2009)
9. Wang, G., Forsyth, D.: Joint Learning of Visual Attributes, Object Classes and Visual Saliency. In: CVPR (2009)
10. Kumar, N., Belhumeur, P.N., Nayar, S.K.: FaceTracer: A Search Engine for Large Collections of Images with Faces. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part IV. LNCS, vol. 5305, pp. 340–353. Springer, Heidelberg (2008)
11. Blei, D.M., Ng, A.Y., Jordan, M.I., Lafferty, J.: Latent Dirichlet Allocation. *JMLR* 3 (2003)
12. Griffiths, T.L., Steyvers, M.: Finding Scientific Topics. *Proceedings of the National Academy of Sciences* 101(suppl. 1), 5228–5235 (2004)
13. Rosen-Zvi, M., Chemudugunta, C., Smyth, P., Steyvers, M.: Learning author-topic models from text corpora. *ACM Transactions on Information System* (2009)
14. Sivic, J., Russell, B., Efros, A.A., Zisserman, A., Freeman, B.: Discovering Objects and Their Location in Images. In: ICCV, pp. 370–377 (2005)
15. Russell, B.C., Efros, A.A., Sivic, J., Freeman, W.T., Zisserman, A.: Using Multiple Segmentations to Discover Objects and their Extent in Image Collections. In: CVPR (2006)
16. Sudderth, E.B., Torralba, A., Freeman, W.T., Willsky, A.S.: Learning Hierarchical Models of Scenes, Objects, and Parts. In: ICCV, vol. 2, pp. 1331–1338 (2005)
17. Fei-Fei, L., Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories. In: CVPR, pp. 524–531 (2005)
18. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual Categorization with Bags of Keypoints. In: Workshop on Statistical Learning in Computer Vision, ECCV (2004)
19. Sun, N., Haas, N., Connell, J.H., Pankanti, S.: A Model-Based Sampling and Sample Synthesis Method for Auto Identification in Computer Vision. In: IEEE Workshop on Automatic Identification Advanced Technologies, Washington, DC, USA, pp. 160–165 (2005)
20. Jiang, D., Hu, Y., Yan, S., Zhang, L., Zhang, H., Gao, W.: Efficient 3D reconstruction for face recognition. *Pattern Recognition* 38, 787–798 (2005)
21. Lowe, D.G.: Distinctive Image Features from Scale-invariant Keypoints. *IJCV* 20, 91–110 (2004)
22. van de Sande, K.E., Gevers, T., Snoek, C.G.: Evaluation of Color Descriptors for Object and Scene Recognition. In: CVPR (2008)
23. Shechtman, E., Irani, M.: Matching Local Self-Similarities across Images and Videos. In: CVPR, pp. 1–8 (2007)
24. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local Features and Kernels for Classification of Texture and Object Categories: A Comprehensive Study. *IJCV* 73, 213–238 (2007)