Learning Scalable Discriminative Dictionary with Sample Relatedness

Jiashi Feng[1,2], Stefanie Jegelka[2], Shuicheng Yan[1], Trevor Darrell[2]
[1]Department of ECE, National University of Singapore, Singapore
[2]Department of EECS & ICSI, UC Berkeley, USA
[1]{a0066331,eleyans}@nus.edu.sg, [2]{stefje,trevor}@eecs.berkeley.edu

Abstract

Attributes are widely used as mid-level descriptors of object properties in object recognition and retrieval. Mostly, such attributes are manually pre-defined based on domain knowledge, and their number is fixed. However, pre-defined attributes may fail to adapt to the properties of the data at hand, may not necessarily be discriminative, and/or may not generalize well. In this work, we propose a dictionary learning framework that flexibly adapts to the complexity of the given data set and reliably discovers the inherent discriminative middle-level binary features in the data. We use sample relatedness information to improve the generalization of the learned dictionary. We demonstrate that our framework is applicable to both object recognition and complex image retrieval tasks even with few training examples. Moreover, the learned dictionary also help classify novel object categories. Experimental results on the Animals with Attributes, ILSVRC2010 and PASCAL VOC2007 datasets indicate that using relatedness information leads to significant performance gains over established baselines.

1. Introduction

We often classify or search for images with complex contents, such as an image containing multiple objects. In addition, several practical scenarios demand to recognize novel objects. Such tasks can pose great challenges to current learning methods for image classification and/or retrieval, whose performance heavily depends on the sufficiency of training samples. In recent years, evidence has emerged that a promising solution to these problems may be approaches based on middle-level representations such as attributes [18, 36]. Attributes are mid-level descriptors of observable object properties such as *furry*, *striped* and *four-legged* for animals. Such attributes occur across different (related) categories [7] and can therefore greatly help recognize previously unseen objects, whose attributes are shared with related objects.

Most of the numerous methods that have been proposed for learning with attributes use a fixed-size attribute vocabulary and require it to be manually pre-defined [7, 18]. Such requirements for supervision come with three main drawbacks. (1) *Scalability.* Many approaches need the training data to be completely annotated with semantic attributes. This limits scalability and number of attributes that can be practically learned. Moreover, fixing the number of attributes in advance fails to adapt to varying complexity across different data sets. (2) *Discriminativeness.* Pre-defined attributes can be redundant and are not necessarily discriminative for similar object categories. For instance, the attribute *four-legged* cannot help distinguish *dog* from *cat*. (3) *Generalization.* Pre-defined attributes may not capture the common properties of different samples and thus may not generalize to novel samples. These limitations inevitably damage the performance of attribute-based object recognition and/or retrieval methods in practice. Although several methods have been proposed to remedy a specific one of the above limitations [18, 8, 27, 1], none of them

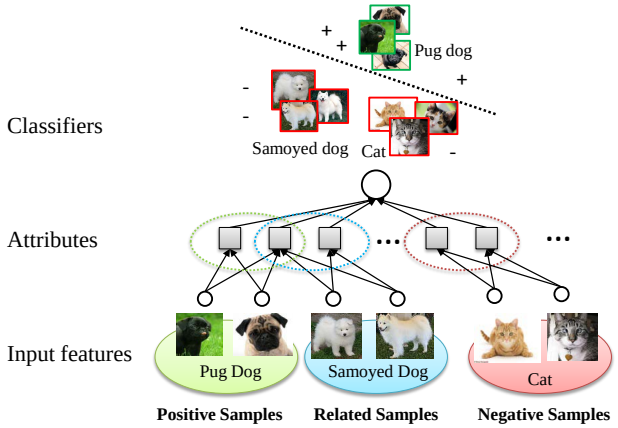

Figure 1. Illustration of the proposed dictionary learning method. It uses three types of samples for training: positive samples, samples related to the positive class and negative samples. "Attributes" of related samples (*pug dog* and *samoyed dog*) are encouraged to be shared, but the "attributes" of unrelated samples (*pug dog* and *cat*) may be different. The binary feature representations indicating existence of attributes are the input for classifiers. Note related samples are still classified to be negative.

overcomes all these limitations simultaneously.

Here, we propose a new dictionary learning method which encodes the image visual features into binary ones, and more importantly it effectively alleviates the above limitations. Our approach is motivated by the fact that humans flexibly adapt the number and nature of the attributes they use to the relatedness and variety of the observed objects, and to the complexity of the task. For example, from the great number of possible attributes to describe a set of animals, such as *furry*, *four-legged* and *can swim*, humans effectively only use a limited number. The principle to select attributes is simple: the chosen attributes should provide sufficient information to reflect shared and discriminative properties. Our method follows this principle and combines three main ingredients.

First, our model discovers binary features by factorizing low-level features of training images into a dictionary of arbitrary (infinite) size – the actual visual patterns present in the data form the dictionary, which adapts to the complexity of the data. Such ideas are captured by priors like the Indian Buffet Process (IBP) [11] or Beta Process (BP) [21]. We use an asymptotic limit of an IBP feature model that allows for fast inference [2]. The resulting AdaptiveDictionary algorithm (whose details are given in Algorithm 1) is practical even for large data sets.

Second, we use the AdaptiveDictionary algorithm in a discriminative framework that not only strives for good representations (with small reconstruction residue), but also biases towards learning dictionary which provides discriminative binary features. In the model, the dictionary, binary representations of training samples and classifiers are learned jointly in a max-margin framework [30].

Third, to enhance the generalization ability of dictionary, we utilize the knowledge about sample relatedness to guide the learned binary features to capture the relational structure between samples. In particular, we encourage closely related samples to have more similar binary features than less related ones. Hence, the dictionary generalizes by exploiting related examples while still being discriminative. Figure 1 shows a graphical illustration of our proposed method.

The comprehensive experiments in Section 5 suggest that the resulting learned dictionary is indeed discriminative and generalizes well. In short, our approach has the following benefits: (1) The size of learned dictionary automatically adapts to the complexity of the training data. Thus we need not bother to determine an appropriate number of basis in the dictionary – our regularization parameter works across a variety of data sets. (2) We also need not pre-define an attribute vocabulary and tediously annotate the attributes for the training samples. (3) The model can incorporate arbitrary levels of sample relatedness from a variety of sources. In this way, the structure captured by the learned dictionary and features can be tailored to specific needs and data.

2. Related Work

Farhadi *et al*. [7] are among the first to propose to use visual attributes to identify familiar objects, and to describe unfamiliar objects when new images are provided. Lampert *et al.* [18] showed that attributes can also be used to recognize object categories without any training image. Following these seminal works, many attribute-based object recognition methods have been proposed [36, 31]. Recently, Parikh *et al.* [23] introduced the concept of relative attributes that capture relations instead of being absolute binary. Kovashka *et al.* [16] integrated relative attributes into user feedback for retrieval. These works require manually labeled samples to learn the attribute classifiers. Moreover, their predefined attributes may not be discriminative.

Recent work moves towards automating the attribute learning process. Parikh *et al.* [22] involved humans to identify the discriminative attributes in an active learning framework. Berg *et al.* [1] proposed to automatically discover attributes by mining images and associated text from the Internet. Rastegari *et al.* [27] learned discriminative attributes by maximizing the separation of different classes. A similar idea was used to automatically design discriminative category-level attributes [8]. All of these works consider attributes as *outputs of a classifier* on low-level features, and the number of attributes is prespecified. IBP priors have been used to learn representations of data with a flexible, adaptive number of attributes [3, 12]. By themselves, they are however not necessarily discriminative. Hence, Quadrianto *et al*. [25] show a supervised IBP method to discover attributes that preserve the neighborhood structure of training data. This method is orthogonal to our approach in that it needs extensive relative ranking annotations of samples. Metric learning [32, 34, 4] and discriminative dictionary learning [15, 20] are conceptually close to attribute learning, but use linear projections and therefore do not generate efficient binary representations.

Semantic relatedness is exploited in several recent works to solve the problem of recognizing a large number of object categories. Most of the existing works only use relations at the level of categories [5, 13]. Yang *et al.* [35] used sample relatedness to help detect complicated events. They associate related but negative samples with soft continuous labels and reduce the classification problem to reduced.

3. Background

3.1. Binary Representation for Images

Given a set of training images $\{I_i\}_{i=1}^N$, which are represented in $\mathbb{R}^D$ by low-level feature vectors $\{\mathbf{x}_i\}_{i=1}^N$, we aim to learn a dictionary of basic visual patterns ("attributes") $\{\mathbf{a}_k\}_{k=1}^K$ in $\mathbb{R}^D$ whose linear combinations can represent

the image features, *i.e.*, $\mathbf{x}_i \approx \sum_{k=1}^{K} z_{i,k}\mathbf{a}_k$. Here, $z_{i,k} \in \{0,1\}$ is a binary indicator of whether the image $\mathbf{x}_i$ has the visual pattern $\mathbf{a}_k$ or not. We will use the collection $\mathbf{z}_i = [z_{i,1}, \ldots, z_{i,K}]^\top$ of the indicators for the K basic visual patterns as the binary representation of the image $\mathbf{x}_i$ for image classification and retrieval tasks. The above feature factorization of all the N images over K basis can be compactly written as $X \approx AZ$. The matrices $X \in \mathbb{R}^{D\times N}$, $A \in \mathbb{R}^{D\times K}$ and $Z \in \mathbb{R}^{K\times N}$ contain the features $\mathbf{x}_i$, dictionary of patterns $\mathbf{a}_k$ and binary representations $\mathbf{z}_i$, respectively.

Instead of learning the binary features as the output of classifiers (as done in recent work [7, 29]), we strive for a good reconstruction. The binary representation of a new image $\mathbf{x}$ will be $\mathbf{z} = \arg\min_{\hat{\mathbf{z}}\in\{0,1\}^K} \|\mathbf{x} - A\hat{\mathbf{z}}\|_2^2$ [1].

Sample relatedness measures the probability of two samples sharing common properties. As a key component, it is utilized to regularize the above binary representations in this work and thus regularize the dictionary, in order to discover basic patterns of good generalization ability. The details are given in Section 4.

As opposed to manual definitions of attributes [18, 7] and fixed-size attribute discovery methods [1, 27], we do not bound the size of our dictionary. Formally, from an infinite number of basis, K_+ basis are realized in the training data; *i.e.*, K_+ depends on the complexity of the data. Our model is derived from the Indian Buffet Process, described next.

3.2. IBP for Learning Flexible Dictionary

The Indian Buffet Process (IBP) [11] is a nonparametric prior for describing an infinite latent feature model. That means, we assume there are infinitely many latent dictionary basis (basic patterns), and we have the IBP prior on the distribution of these basis. Each image can have multiple basic patterns, and the number of basis in N images is almost surely finite [11]. Before specifying the prior, we turn to the other distributions – the prior on A and the likelihood. Both are often assumed to be Gaussian [11, 2, 25]:

$$p(A|0,\sigma_A^2) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{N}(a_{d,k};0,\sigma_A^2). \tag{1}$$

$$p(X|Z,A,\sigma_X^2) = \prod_{i=1}^{N}\mathcal{N}(\mathbf{x}_i; A\mathbf{z}_i, \sigma_X^2 I). \tag{2}$$

Here, σ_A^2 and σ_X^2 denote the variances. The likelihood in Eqn. (2) says that the observed feature $\mathbf{x}_i$ concentrates around the combination of its constituent basis $A\mathbf{z}_i$, with a deviation σ_X^2 caused by noises.

[1] If we were to restrict our basis to be orthogonal, *i.e.*, $\mathbf{a}_i^\top \mathbf{a}_j = 0, \forall i \neq j$, then we had linear transformations $\mathbf{z}_i = A^T\mathbf{x}_i$ as in [7, 29]. However, it is easy to show that with binary coefficients $\mathbf{z}$, orthogonal attributes have insufficient representational power.

Remark 1 (On the Gaussian assumption)**.** *Note that in our experiments, the extracted image features are mostly histograms (in $\mathbb{R}_+$ space) and the above Gaussian assumptions (over $\mathbb{R}$ space) may be a bit simplistic. However, theoretically, our model can integrate a wider range of exponential family distributions (see [14]). As to the current work, even though the true distributions are most likely not Gaussian, the model performs well. We would expect it to perform even better with more accurate assumptions and we leave this investigation in our future work.*

The IBP prior distribution for the binary indicator z_{ik} is a Bernoulli distribution $\text{Bernoulli}(\pi_k)$, whose parameter π_k is sampled from a Beta distribution $\text{Beta}(\alpha/K, 1)$, parameterized by a hyperparameter α and the number of basis K [11]. Thus,

$$p(Z|\alpha) = \prod_{k=1}^{K}\frac{\frac{\alpha}{K}\Gamma(m_k+\frac{\alpha}{K})\Gamma(N-m_k+1)}{\Gamma(N+1+\frac{\alpha}{K})},$$

where $m_k = \sum_{i=1}^{N} z_{k,i}$ denotes the total number of times the kth basis occurs in N samples. Taking $K \to \infty$, we obtain the final prior on the binary representations [11]:

$$\lim_{K\to\infty} p(Z|\alpha) = \frac{\alpha^{K_+}}{K_+!}e^{-\alpha H_N}\prod_{k=1}^{K_+}\frac{(N-m_k)!(m_k-1)!}{N!}. \tag{3}$$

Here $H_N = \sum_{i=1}^{N} 1/i$ is the Nth harmonic number, and $K_+ < \infty$ denotes the finite number of discovered basic patterns in the observations X.

Learning such models, however, can be computationally challenging and usually requires sampling or variational methods [11, 28]. As an alternative, Broderick *et al.* [2] derive the limit of the joint distribution $P(X, A, Z)$ as $\sigma_X \to 0$. In this case, each feature $\mathbf{x}_i$ is deterministically assigned to the set of basis that generate the best representation for it. The complexity of the model (number of basis) is indirectly controlled by a parameter λ that is connected to the hyperparameter α as $\alpha = \exp(-\lambda^2/2\sigma_X^2)$. In the end, asymptotically the joint probability becomes [2]:

$$-\log p(X,A,Z) \sim \|X - AZ\|_F^2 + \lambda^2 K_+ \tag{4}$$

up to constant terms. That means, MAP inference in the limit is equivalent to minimizing $\|X-AZ\|_F^2+\lambda^2K_+$ w.r.t. A, Z, K_+, where K_+ is the number of basis in the dictionary A. This is the objective we will work with.

4. Scalable Discriminative Dictionary

4.1. Learning Discriminative Dictionary in IBP

Let $\{X, \mathbf{y}\}$ be the training feature vectors and labels. Each element y_i in $\mathbf{y} \in \mathbb{R}^N$ is the category label of $\mathbf{x}_i$ and takes one value from a set $\mathcal{C} = \{1, \ldots, C\}$ indexing

C categories. For the cth category, we train a separate linear classifier $\mathbf{w}_c \in \mathbb{R}^{K_+}$ that uses the binary representations Z. That is, a test sample $\mathbf{x}$ with basis assignments $\mathbf{z}$ is classified as $y = \arg\max_{y' \in \mathcal{C}} \mathbf{w}_{y'}^\top \mathbf{z}$. We collect all classifiers in $W = [\mathbf{w}_1, \ldots, \mathbf{w}_C]$.

To learn dictionary that are useful for the ultimate task of classification, we simultaneously integrate representational (reconstruction of X from dictionary basis) and discriminative (labels and classifiers W) aspects in our model:

$$p(X, \mathbf{y}, A, Z, W) = p(W)p(A)p(Z)p(\mathbf{y}|Z, W)p(X|A, Z).$$

This model is in the spirit of [19] with an equality prior on the generative and discriminative parameters. The prior on A is Gaussian (Eqn. (1)), while for the representational part $p(A)p(Z)p(X|A, Z) = p(A, Z, X)$, we take the asymptotic model in Eqn. (4):

$$-\log\left(p(A)p(Z)p(X|A, Z)\right) \propto \|X - AZ\|_F^2 + \lambda^2 K_+. \quad (5)$$

For W, we use $p(W) = \exp\{-\|W\|_F^2\}$. The conditional distribution of the classes is discriminative [17]:

$$-\log\left(p(\mathbf{y}|Z, W)p(W)\right) \propto \sum_i f(y_i; W, \mathbf{z}_i) + \beta \sum_{i,j} g(\mathbf{z}_i, \mathbf{z}_j) + \|W\|_F^2. \quad (6)$$

In particular, the first term $f(\cdot)$ represents a multi-class classification loss on the labeled samples. We define $f(\cdot)$ as [30]:

$$f(y_i; W, \mathbf{z}_i) = \max_{\bar{y}_i \in \mathcal{C} \setminus y_i} \mathbf{w}_{\bar{y}_i}^\top \mathbf{z}_i - \mathbf{w}_{y_i}^\top \mathbf{z}_i, \quad (7)$$

where $\bar{y}_i$ denotes the incorrect class label for the sample $\mathbf{x}_i$. Minimizing this loss will encourage the attribute representations from different categories to be separated with a large margin. The second term $g(\cdot, \cdot)$ encourages related samples to have similar attribute representations:

$$g(\mathbf{z}_i, \mathbf{z}_j) = s_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2. \quad (8)$$

The coefficient s_{ij} is the strength of relatedness between the samples $\mathbf{x}_i$ and $\mathbf{x}_j$. For more closely related samples, the value of s_{ij} is larger and hence $g(\cdot, \cdot)$ enforces them to share more basis. As a result, the dictionary basis will describe properties that are shared between related samples. There are various ways to set the values of s_{ij}, for instance, using the semantic similarity of category labels, appearance similarity, or user feedback.

4.2. Attribute Learning Algorithm

Combining the probabilities in Eqn. (5) and Eqn. (6), and maximizing the obtained joint probability yield the following optimization problem:

$$\min_{W, A, Z, K_+} \sum_i \left\{ \max_{\bar{y}_i \in \mathcal{C} \setminus y_i} \mathbf{w}_{\bar{y}_i}^\top \mathbf{z}_i \right\} - \mathbf{w}_{y_i}^\top \mathbf{z}_i + \|W\|_F^2 + \beta \sum_{i,j} s_{ij} \|\mathbf{z}_i - \mathbf{z}_j\|_2^2 + \|X - AZ\|_F^2 + \lambda^2 K_+. \quad (9)$$

We alternatingly optimize the objective function with respect to W and A, Z as follows. Fixing the dictionary A, we obtain the representation Z for X in terms of A. Using the representations Z, we learn the classifiers W. Given the new W, we modify A and Z to reduce the classification loss. This process is repeated until convergence. Algorithm 1 summarizes the optimization procedure.

Algorithm 1: AdaptiveDictionary

```
Initialize K_+ = 1, A = [∑_i x_i / N].
while Objective function decreasing do
    for i = 1, ..., N do
        for k = 1, ..., K_+ do
            Set z_{i,k} ∈ {0, 1} to minimize the objective
            in Eqn. (9) greedily;
        end
    end
    A ← X Z^⊤ (Z Z^⊤)^{-1};
    Sample new basis a_{K_+ + 1} with probabilities
    P(a_{K_+ + 1} = x_i) ∝ ||x_i − A z_i||_2^2;
    A ← [A, a_{K_+ + 1}], K_+ ← K_+ + 1.
    Update classifiers W (Linear SVM).
end
```

The AdaptiveDictionary algorithm starts by assigning every sample to the first basis, and consequently infers that the first basis is the average of all sample points. As the algorithm proceeds, it computes the squared loss $\|\mathbf{x} - A\mathbf{z}\|^2$ for each sample, and selects a new candidate basis via importance sampling: point $\mathbf{x}$ becomes a new basis with probability proportional to its residual. We may stop adding the basis early if the number of attributes reaches a pre-defined value, or run the algorithm until convergence. In our experiments, we found the proposed algorithm to converge fast (see supplementary material).

4.3. Other Applications

Our framework does not only apply to classification; it is equally suitable for learning discriminative binary features for image retrieval [6] or the detection of complex events [35], considering the obtained representations are compact binary vectors and efficient for storage and calculating similarities. In these tasks, we replace the multi-class classification in Eqn. (7) by regression: $f(y_i; \mathbf{w}, \mathbf{z}_i) = |\mathbf{w}^\top \mathbf{z}_i - y_i|^2$. Here, the y_i's are annotated ranking scores

(for retrieval) or confidence scores (for event detection) of the ith sample. The ranking or confidence score of a new sample is predicted as $\mathbf{w}^\top \mathbf{z}$.

5. Experiments

5.1. Classification on the AwA Dataset

First, we address the classification of known object categories and investigate the effectiveness of using sample relatedness. We then extend the problem to recognizing novel categories. Finally, we look at the adaptivity and visual properties of learned dictionary. All the experiments in this subsection use the Animals with Attributes (AwA) dataset introduced in [18]. The dataset contains $30,475$ images from 50 animal categories, and 85 manually defined attributes for those animals, such as *black*, *big*, *strong*. Each category is labeled with respect to the 85 attributes. We follow the experimental protocol of [18] and use the provided sample split: 40 categories as known ($24,295$ images) and 10 categories as novel ($6,180$ images). We use the provided pre-computed low-level features, including RGB color histogram, SIFT, rgSIFT, PHOG, SURF and local self-similarity histograms. All the features are first normalized individually and then concatenated into a single $10,940$-dimensional vector. We then apply PCA to reduce the dimension of the feature vectors to 700 for the implementation efficiency.

We define the sample relatedness s_{ij} of two training images in Eqn. (8) via the semantic similarity of their (lowest) categories in the WordNet [9] hierarchy:

$$s_{ij} = \frac{\text{depth}(P_{ij})}{0.5 \times \text{length}(i,j) + \text{depth}(P_{ij})}. \qquad (10)$$

Here, P_{ij} denotes the nearest common ancestor of the samples $\mathbf{x}_i$ and $\mathbf{x}_j$, $\text{depth}(\cdot)$ denotes the depth in the hierarchy and $\text{length}(\cdot,\cdot)$ gives the path length between two nodes.

5.1.1 Classifying Known Categories

The first experiment studies the discriminative ability of the learned dictionary and binary features to classify the 40 known categories on the AwA dataset. Following the experimental setup in [8], we select different numbers of images per category for training ($15, 20, 25, 30, 50$ respectively), 25 images per category for testing and 10 images per category for validation. The values of parameters are tuned on the validation set and then fixed as $\lambda = 1 \times 10^{-3}$ and $\beta = 0.1$. When training on $2,000$ images (50 images per category), Algorithm 1 converges after around 150 iterations. We plot the convergence curve of the objective value in the supplementary material. It takes about 167 seconds to learn 204 basis from the 2,000 training images on a Matlab platform on a PC with 2.83 GHz Quad CPU. We

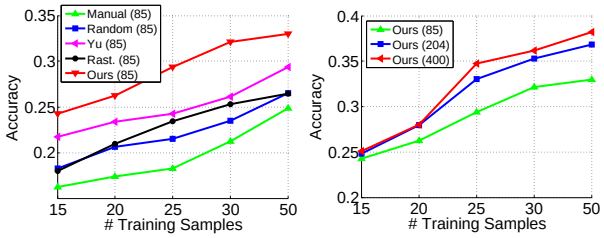

Figure 2. Average classification accuracy over 40 known categories. Left: comparison between our method and baselines with a fixed number of 85 attributes (basis). Right: accuracy of our method with varying numbers of attributes (basis).

compare our framework with four other types of attributes: (1) the category-level attribute annotations provided in the AwA dataset [18]; (2) random category-level attributes [7]; (3) the category-level discriminative attributes of [8]; (4) learning discriminative binary codes for each image [27]. For a fair comparison, the number of attributes for all the compared methods is fixed to 85. That means we stop our algorithm after having discovered 85 attributes. The classification accuracy for varying sizes of the training set is shown in Figure 2 (left).

The results suggest two observations. First and not surprisingly, all discriminative attribute learning methods (Yu [8], Rastegari [27] and ours) outperform manually and random attributes in classification. This implies the utility of learning attributes tuned for the ultimate task. Second, our method consistently outperforms all the compared methods for different training set sizes. Hence, it is beneficial for classification of known categories to shape the structure of the dictionary basis, via the reconstruction term and the regularization from sample relatedness.

Figure 2 (right) shows the performance of our adaptive dictionary learning for different (fixed) numbers of basis (85, 204 and 400). The number 204 is automatically determined by the algorithm. The 400 attributes are obtained via continuing to apply Algorithm 1 after achieving the stopped criterion. We see that increasing the number of attributes indeed helps the classifiers, but the gains diminish at some point: the automatically chosen number of 204 basis is basically as good as doubling this number. On the other hand, for a small training set, smaller dictionary is sufficient. This makes sense, as the complexity of the data mildly grows with the size of the training set. However, the size of the useful attribute representation (vocabulary) is clearly sublinear in the number of training samples, and levels off quickly (see also Section 5.1.4). Moreover, the limited size of dictionary imposed by our approach has an additional regularizing effect (smaller shattering coefficient).

5.1.2 Effectiveness of Using Sample Relatedness

Next, we investigate the effect of using related samples to enhance the classification performance when the training

Table 1. Classification accuracy (%) with different values of the weight β of using sample relatedness.

# labeled	$\beta = 0$	$\beta = 0.1$	$\beta = 0.5$
15	19.25	24.25	24.51
20	20.86	26.25	26.49
25	23.74	29.38	29.27
30	28.24	32.13	32.29
50	30.49	33.00	32.67

images are insufficient. We tune the value of the parameter β of the related sample term in Eqn. (9) to control the contribution of the related samples, and select three different values of $\beta = \{0, 0.1, 0.5\}$. Setting $\beta = 0$ is equivalent to not using the sample relatedness. We run the evaluations over 5 different splits of the training samples. Table 1 displays the performance (accuracy averaged over multiple splits) of classifying the 40 known categories for different training set sizes. The results imply that when the number of training samples is small (*e.g.*, 15, 20 training images per category), the performance gain induced by the related samples is significant (around 6%). This clearly demonstrates that when the training data are not sufficient for learning discriminative features and good classifiers, related samples can transfer useful common knowledge via the learned dictionary. However, when the training samples are sufficiently informative, the gain from related samples is reduced: if 50 training samples per category are provided, the related samples improves performance only by about 2%.

5.1.3 Classifying Novel Categories

Here, we test how the learned dictionary generalizes to the 10 novel categories (6,180 images in total) that are not present in the above dictionary learning. The images are split into the training set (50 images per category), the test set (200 images per category) and the validation set (the remaining images). We use the dictionary A learned from the 40 known animal categories in Section 5.1.1. The binary representation of any image $\mathbf{x}$ from novel categories is $\mathbf{z} = \arg\min_{\mathbf{z}\in\{0,1\}^\kappa} \|\mathbf{x} - A\mathbf{z}\|_2$. The category prediction on novel samples in the test set is via clustering (without using the training set) or linear classifiers (learned on the training set) as in [8]. Here we do not adopt the method of classifying novel categories in [18] as it relies on category-level attribute labels. We compare with five baseline methods: (1) using manually [18] and (2) randomly [7] defined attributes, (3) using low-level features provided in the dataset [18], (4) using classemes [29] and (5) the method proposed in [8]. The comparison results are plotted in the left panel of Figure 3. The results confirm that our method consistently outperforms the previous methods under varying numbers of training samples (including no training samples from the novel category). Here our approach benefits from using sample relatedness, which guides it to learn attributes capturing the shared knowledge. As a result, the learned attributes help generalization. We also investigate the performance for different sizes of the dictionary. The results are shown in Figure 3 (right). Here, 204 is the dictionary size determined by our method. As with known categories, we observe diminishing gains with respect to the number of basis: increasing from 30 to 85 basis boosts the performance significantly, while further increasing the number of basis does not improve the performance much.

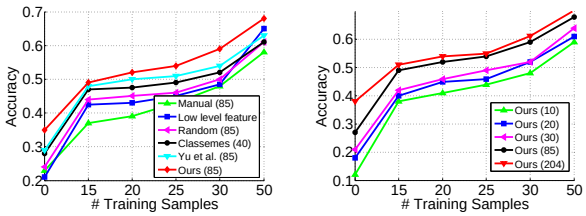

Figure 3. Average classification accuracy over 10 novel categories. Left: comparison with baselines with fixed numbers of basis. Right: results of our method with varying numbers of basis. When the number of training samples is zero, classification is achieved via clustering the test samples.

5.1.4 Properties of The Learned Dictionary

To gain insight beyond the quantitative experiments above, we explore other properties of our learned dictionary.

Adaptivity. First, as a proof of concept, we learn dictionary on (1) a very homogeneous data set (Set I) with images from three categories that have many common properties: *Dalmatian*, *German Shepherd* and *Chihuahua*; and (2) a very diverse data set (Set II) with images drawn from three very different categories: *Antelope*, *Beaver* and *Dalmatian*. One would expect that the diverse Set II requires more basis to be described than the homogeneous Set I. Figure 4 shows results with 50 images per category ($\beta = 0.01$ and $\lambda = 1$). Indeed, the model automatically learns more attributes for Set II, which has a wider variety of data. For both data sets, the number of attributes grows automatically with the size of the data.

Dictionary Visualization. As a second part of investigating properties of the learned dictionary, Figure 5 visualizes the dictionary basis via exemplar images[2]. The parameter setting for attribute learning was identical to the one in Section 5.1.1. Each row in Figure 5 shows 5 example images whose features are closest to a basis $\mathbf{a}$ (a column vector in the learned A). The exemplars suggest that the learned basis capture certain common properties across multiple categories. For example, the basis in row (a) may be described by the texture pattern of *furry*, and row (b) may be viewed close to the concepts of *dotted* and *striped* patterns.

[2]Use the example images provided in the dataset

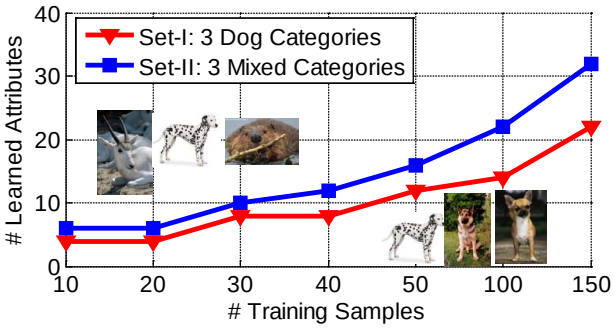

Figure 4. Size of learned dictionary vs. dataset size and complexity. The red (resp. blue) curve shows the number of learned dictionary basis from 3 dog categories (resp. mixture of 3 quite different animal categories). Example images of the two image sets are shown at the neighbor of the corresponding curve.

Figure 5. Visualization of learned dictionary. Each row shows the top 5 closest images to a certain basis, in terms of the Euclidean distance between their features.

5.1.5 Component-wise Model Evaluation

In the supplementary material, we provide further experiments evaluating the effect of the individual components of our model, *i.e.*, the discriminative (Eqn. (6)) and the generative part (Eqn. (5)), on image classification results.

5.2. Classification on ILSVRC2010

In addition to AwA, we learn adaptive dictionary on the large ILSVRC2010 dataset and use them to classify novel categories. ILSVRC2010 contains 1.2 million images from $1,000$ diverse categories. Following the experimental setting in [8, 10], 950 categories are randomly chosen as known categories for dictionary learning (1,192,481 images). We evaluate the classification performance on the remaining 50 disjoint categories (68,925 images) using the learned dictionary. The $4,096$-dimensional Fisher vectors in [10] are extracted to represent images. Considering the large number of training samples, we use a small weight $\beta = 10^{-5}$ on the sample relatedness regularization. We also control the size of dictionary to be relatively small ($\lambda^2 = 10^3$) for fair comparison with baselines. The sample relatedness is computed on the ImageNet hierarchy by Eqn. (10). We randomly sample $100,000$ training images to learn the dictionary. The dictionary learning takes around 3 hours on a server with 32 GB RAM and 8 CPU cores of 3.00 GHz and discovers $1,554$ attributes. For classification, we again use a linear SVM. We use 80% of the novel category samples for training ($55,140$ images), 10% for testing ($6,893$ images), and 10% for validation ($6,892$ images). Table 2 shows the resulting multi-class classification accuracy, using different numbers of training images from the training set of novel categories. Even though it uses fewer attributes than the best performing baseline [8] ($1,554$ vs. $2,000$), our method outperforms it by $4-7\%$.

Table 2. Image classification accuracy on 50 classes from ILSVRC2010. The numbers in parentheses indicate # attributes.

# training	1%	5%	10%	50%	100%
Low-level feature	33.15	50.50	55.12	64.31	66.70
Classemes (950)	36.46	49.82	54.07	62.22	64.55
Yu [8] (2,000)	40.06	53.14	57.30	63.54	66.91
Ours (1,554)	**44.23**	**58.68**	**64.38**	**69.71**	**71.24**

5.3. Image Retrieval on PASCAL VOC2007

Finally, we test the usefulness of our approach for retrieving images that contain *multiple* objects. For this experiment, we use the trainval subset of PASCAL VOC2007, which contains $5,011$ images of 20 object categories. Each image contains multiple objects. In a practical image retrieval system, users can manually specify the sample relatedness among certain samples, *i.e.*, s_{ij} in Eqn. (9), according to their personal preference. Such feedback can be directly integrated to tailor the attributes to a certain user's interests, improving his experience during the retrieval. Here, we simulate this feedback by specifying the relatedness of two images by the number of overlapping category labels.

We extract the same Fisher vector features as [24]. Four state-of-the-art methods in image retrieval [10] are adopted as comparison baselines: two unsupervised methods (Spectral Hashing [33] and SKLSH [26]), and two supervised methods (CCA based ITQ [10] and the approach in [8]). We randomly select 500 images as queries and use the remaining images for training. For the baselines of SH, SKLSH and ITQ, we use the implementations and parameter settings provided by the authors of [10]. The number of feature bits used by the four baselines is fixed as 200, and we tune our parameter λ to learn 196 attributes for fair comparison. The parameters are set as $\lambda = 0.2$ and $\beta = 0.1$. The samples having more than 1 overlapping category label with the query are treated as positive. Performance is evaluated by the precision and recall rate, and is shown in Figure 6. The results demonstrate that, as the case for classification, the supervised approaches generally outperform the unsupervised ones, because they benefit from the additional category information related to the end task. Moreover, our method significantly outperforms all four baselines.

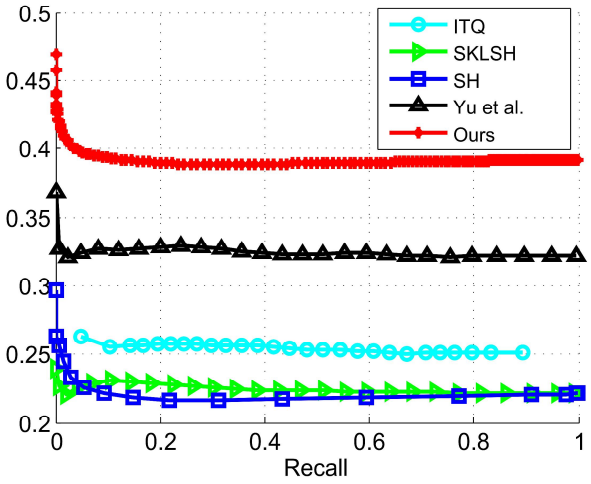

Figure 6. The image retrieval precision-recall curves of our method and other state-of-the-art methods on PASCAL VOC2007.

6. Conclusions and Future Work

We propose a novel dictionary learning model that is able to automatically discover basic patterns (non-semantic attributes) with both good discriminative and generalization properties. It incorporates related samples and semantic information to learn a representation well-aligned with the relations of categories. As a result, it captures shared properties of the related objects very well. In addition, the learned dictionary automatically adapts to the complexity of the data. Such flexible dictionary can be, as we show, effectively applied for image classification and retrieval. Our model is in structure similar to a three-layer neural network which has, however, fewer layers than popular Deep Neural Networks (DNN). Hence, it can be trained much more efficiently, while still performing very well. The performance is due to a regularization in the attribute layer that is targeted to the end task. Such regularization is usually not applied in DNNs. In addition, the number of "attribute" equivalents in DNNs is usually pre-set and fixed. It would be an interesting direction to further investigate the relationship between our structurally steered attributes and DNNs, and how such regularization ideas may help augment DNN models.

Acknowledgement J. Feng and S. Yan are supported by the Singapore National Research Foundation under its International Research Centre @Singapore Funding Initiative and administered by the IDM Programme Office. S. Jegelka is supported in part by the Office of Naval Research under contract/grant number N00014-11-1-0688. T. Darrell is supported in part by DARPA Mind's Eye and MSEE programs, by NSF awards IIS-0905647, IIS-1134072, and IIS-1212798, and by support from Toyota.

References

[1] T. Berg, A. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010. 1, 2, 3

[2] T. Broderick, B. Kulis, and M. Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *ICML*, 2013. 2, 3

[3] S. Changpinyo and E. Sudderth. Learning image attributes using the Indian Buffet Process. Technical report, Brown University, 2012. 2

[4] J. Davis, B. Kulis, P. Jain, S. Sra, and I. Dhillon. Information-theoretic metric learning. In *ICML*, 2007. 2

[5] J. Deng, A. Berg, and L. Fei-Fei. Hierarchical semantic indexing for large scale image retrieval. In *CVPR*, 2011. 2

[6] M. Douze, A. Ramisa, and C. Schmid. Combining attributes and Fisher vectors for efficient image retrieval. In *CVPR*, 2011. 4

[7] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3, 5, 6

[8] Y. Felix, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, 2013. 1, 2, 5, 6, 7

[9] C. Fellbaum. *WordNet*. Springer, 2010. 5

[10] Y. Gong and S. Lazebnik. Iterative quantization: A Procrustean approach to learning binary codes. In *CVPR*, 2011. 7

[11] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian Buffet Process. In *NIPS*, 2006. 2, 3

[12] Y. Hu, K. Zhai, J. Boyd-Graber, and S. Williamson. Modeling images using transformed Indian Buffet Processes. In *ICML*, 2012. 2

[13] Y. Jia, J. Abbott, J. Austerweil, T. Griffiths, and T. Darrell. Visual concept learning: Combining machine vision and Bayesian generalization on concept hierarchies. In *NIPS*, 2013. 2

[14] K. Jiang, B. Kulis, and M. I. Jordan. Small-variance asymptotics for exponential family dirichlet process mixture models. In *NIPS*, pages 3167–3175, 2012. 3

[15] N. Khan and M. Tappen. Stable discriminative dictionary learning via discriminative deviation. In *ICPR*, 2012. 2

[16] A. Kovashka, D. Parikh, and K. Grauman. Whittlesearch: Image search with relative attribute feedback. In *CVPR*, 2012. 2

[17] S. Kumar and M. Hebert. Discriminative random fields. *IJCV*, 2006. 4

[18] C. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2, 3, 5, 6

[19] J. Lasserre, C. Bishop, and T. Minka. Principled hybrids of generative and discriminative models. In *CVPR*, 2006. 4

[20] J. Mairal, F. Bach, and J. Ponce. Task-driven dictionary learning. *TPAMI*, 2012. 2

[21] J. Paisley and L. Carin. Nonparametric factor analysis with beta process priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 777–784. ACM, 2009. 2

[22] D. Parikh and K. Grauman. Interactively building a discriminative vocabulary of nameable attributes. In *CVPR*, 2011. 2

[23] D. Parikh and K. Grauman. Relative attributes. In *ICCV*, 2011. 2

[24] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *ECCV*, 2010. 7

[25] N. Quadrianto, V. Sharmanska, D. Knowles, and Z. Ghahramani. The supervised IBP: Neighbourhood preserving infinite Latent Feature Models. In *UAI*, 2013. 2, 3

[26] M. Raginsky and S. Lazebnik. Locality-sensitive binary codes from shift-invariant kernels. In *NIPS*, 2009. 7

[27] M. Rastegari, A. Farhadi, and D. Forsyth. Attribute discovery via predictable discriminative binary codes. In *ECCV*, 2012. 1, 2, 3, 5

[28] C. Reed and Z. Ghahramani. Scaling the Indian Buffet Process via submodular maximization. In *ICML*, 2013. 3

[29] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient object category recognition using classemes. In *ECCV*, 2010. 3, 6

[30] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. In *JMLR*, 2005. 2, 4

[31] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 2

[32] K. Weinberger and L. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009. 2

[33] Y. Weiss, A. Torralba, and R. Fergus. Spectral hashing. In *NIPS*, 2008. 7

[34] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *NIPS*, 2002. 2

[35] Y. Yang, Z. Ma, Z. Xu, S. Yan, and A. Hauptmann. How related exemplars help complex event detection in web videos? In *ICCV*, 2013. 2, 4

[36] X. Yu and Y. Aloimonos. Attribute-based transfer learning for object categorization with zero/one training example. In *ECCV*, 2010. 1, 2