Learning Predictable and Discriminative Attributes for Visual Recognition

Yuchen Guo and **Guiguang Ding** and **Xiaoming Jin** and **Jianmin Wang**
School of Software, Tsinghua University, Beijing 100084, China
yuchen.w.guo@gmail.com, {dinggg,xmjin,jimwang}@tsinghua.edu.cn,

Abstract

Utilizing attributes for visual recognition has attracted increasingly interest because attributes can effectively bridge the semantic gap between low-level visual features and high-level semantic labels. In this paper, we propose a novel method for learning predictable and discriminative attributes. Specifically, we require the learned attributes can be reliably predicted from visual features, and discover the inherent discriminative structure of data. In addition, we propose to exploit the intra-category locality of data to overcome the intra-category variance in visual data. We conduct extensive experiments on Animals with Attributes (AwA) and Caltech256 datasets, and the results demonstrate that the proposed method achieves state-of-the-art performance.

Introduction

Attributes, which reflect the properties shared by objects, have been widely used as the mid-level representation for images to bridge the semantic gap between low-level visual features and high-level semantic labels. Actually, it's always difficult to directly construct models between visual features and labels. Thus we can find some attributes as the intermediary which are predictable from visual features and lead to natural models of categories. Attribute-based representation has shown promising results in many applications, such as object recognition (Rastegari, Farhadi, and Forsyth 2012; Yu et al. 2013), large-scale image retrieval (Siddiquie, Feris, and Davis 2011), and facial verification (Kumar et al. 2009).

Designing attributes manually (Farhadi et al. 2009) may lead to interpretable attributes. However, it's burdensome and costly to do so and it has been widely observed that the performance of obtained attributes is sometimes worse than random attributes because they are neither predictable nor discriminative. Recently, learning *latent* attributes automatically from data has attracted increasingly attention (Rastegari, Farhadi, and Forsyth 2012; Yu et al. 2013; Akata et al. 2013). With some proper learning criteria, latent attributes are more predictable which means they can be generated reliably from visual features by some models like SVM classifier, and more discriminative such that superior visual recognition performance can be achieved with them.

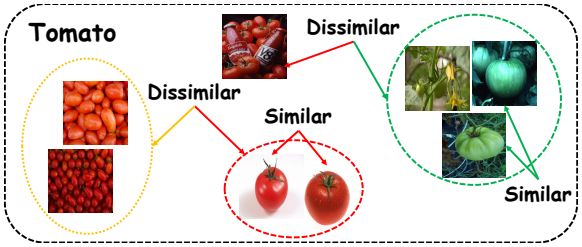


Figure 1: All images are in "Potato" category, from Caltech256 dataset. Images of the same category may have different attributes because of the large intra-category variance.

Based on how the latent attributes are learned, we can summarize recent methods into *category-based* methods and *sample-based* methods. The category-based methods learn attributes by considering the relationship between categories, like in (Yu et al. 2013). They aims to design attributes for each category such that large category-separability can be achieved. Though high discriminability can be achieved, they still suffer from low predictability. On the contrary, sample-based methods, such as in (Rastegari, Farhadi, and Forsyth 2012), mainly focus on constructing effective models between samples and attributes such that the learned attributes can be reliably predicted. Although they also incorporate some criteria into the learning function to improve the discriminability, such as small intra-category distance and large inter-category distance, the discriminability is guaranteed just in an indirect way. Consequently, the learned attributes aren't discriminative enough. Actually, the learned attributes should be predictable and discriminative simultaneously such that accurate visual recognition from visual features to categories can be achieved with attributes as intermediary. However, we can see sample-based methods and category-based methods ignore either of them respectively.

Moreover, existing methods require images of the same category to have similar or identical attributes. However, this requirement is too strict for real-world dataset such that recognition performance can be severely affected. Actually, there is large intra-category variance in real-world dataset. As illustrated in Figure 1, we can observe that some images of the same category seem very dissimilar. Therefore, it's not reasonable to assign similar attributes to all images of the same category. On the contrary, we just need to assign similar attributes to *similar* images of the same category, which is termed as exploiting *intra-category locality* in this paper.

Motivated by this observation, in this paper, we propose a novel method for learning **P**redictable and **D**iscriminatve **A**ttributes (PDA) for visual recognition. Specifically, we model the learning problem by a max margin framework which jointly optimizes the predictability and discriminability such that both can be explicitly achieved. Then we can obtain two groups of max-margin classifiers. One group is for reliably generating attribute representation for images from visual features and the other is for performing recognition based on the attributes. Furthermore, we regularize this framework with intra-category locality. This regularization is quite important because it can address the large intra-category variance and noise in visual data. Because our attributes are both predictable and discriminative, and the intra-category variance is taken into account, it's expected to achieve better recognition performance on them. We conduct image classification on AwA dataset. Moreover, the learned attributes by PDA are represented by binary codes, so it can support highly efficient content-based image retrieval (CBIR) as Hashing (Gionis, Indyk, and Motwani 1999). Thus we also carried out CBIR experiments on Caltech256. Both tasks are quite typical for visual recognition. Experimental results validate the effectiveness of our PDA in comparison with several state-of-the-art related methods.

Related Work

As mentioned above, the attributes should be predicable and discriminative, while the interpretable isn't necessary. Actually, manually designed attributes shows unsatisfactory performance because they are neither predictable nor discriminative. Fortunately, we can discover attributes from data automatically. Yu *et al.* propose a method to design discriminative *category*-level attributes (CLA). Given a matrix $\mathbf{S} \in \mathbb{R}^{C\times C}$ where S_{ij} is the similarity between the i-th and the j-th categories, CLA learns attributes $\mathbf{A} \in \mathbb{R}^{C\times k}$ for *categories* by minimizing the objective function as follows,

$$\mathcal{O}_{\text{CLA}} = tr(\mathbf{A}^T(\mathbf{P} - \lambda\mathbf{L})\mathbf{A}) + \beta\|\mathbf{A}^T\mathbf{A} - \mathbf{I}\|_F^2 \quad (1)$$

where $\mathbf{P}$ is with diagonal elements being $k-1$ and all others -1, and $\mathbf{L}$ is the Laplacian of $\mathbf{S}$ (von Luxburg 2007). Here each category is represented by different attributes respectively and their representations are expected to be similar for related categories and dissimilar for unrelated categories. However, they represent each category by one attribute representation, i.e., all images of one categories should have the same attributes, which ignores the intra-category variance in visual data. And they train a group of SVM classifiers to predict attributes for to-be-classified samples **after** they obtain the category attributes, which may lead to low predictability.

On the other hand, Rastegary *et al.* propose to represent samples by discriminative binary codes (DBC). Different from CLA, they design sample-level attributes. Images of the same category can have different attributes. Furthermore, they incorporate the attribute classifier learning into the attribute designing, such that the attributes are highly predictable. And they require images of the same category to have similar attributes while images of different categories to have dissimilar attributes, which may lead to discriminative attributes. However, they ignore the intra-category variance and the discriminability is achieved in an indirect way.

Table 1: Notations and descriptions in this paper

Notation	Description	Notation	Description
$\mathbf{X}$	data matrix	n	#samples
$\mathbf{Y}$	label matrix	d	#dimension
$\mathbf{A}$	attribute matrix	k	#attributes
$\mathbf{S}$	similarity matrix	C	#categories
$\mathbf{W}$	NN matrix	p	#NN
$\mathbf{w}_{va}$, $\mathbf{w}_{ac}$	classifiers	α, λ	parameters

Other than the global information of data, the local structure has drawn considerable attention because there is evidence that data is often drawn from a low-dimensional manifold embedded in ambient space (Roweis and Saul 2000; Belkin and Niyogi 2001). Theoretically, the manifold can be discovered by exploiting the local geometric structure, i.e., considering the relationship between near points rather than all points (Belkin, Niyogi, and Sindhwani 2006). Exploiting locality of data leads to much better learning performance (Cai et al. 2011). Actually, when facing large intra-category variance, global information is useless because it can't capture the intrinsic relationship of data. And it will be influenced by noise, which is very common in real-world dataset.

The Proposed Method

A Unified Framework

Given a set of images $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n] \in \mathbb{R}^{d\times n}$, and the corresponding label matrix $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_n] \in \{0, 1\}^{C\times n}$ where $y_{ij} = 1$ if the j-th image belongs to the i-th category and $y_{ij} = 0$ otherwise, we aims to learn predictable and discriminative attributes $\mathbf{A} = [\mathbf{a}_1, ..., \mathbf{a}_n] \in \{-1, 1\}^{k\times n}$ and two groups of classifiers, $\mathbf{w}_{va}^i (i = 1, ..., k)$ between visual features and attributes, and $\mathbf{w}_{ac}^j$ $(j = 1, ..., C)$ between attributes and categories. Below we will introduce our learning method in detail. Some notations appearing in this paper and their corresponding descriptions are summarized in Table 1.

Following the work in (Rastegari, Farhadi, and Forsyth 2012), we regard each attribute as a split of feature space. Actually, it's intuitive to assume samples at one side of the split may share some properties, i.e., they should have similar attributes. Each split is modeled as a hyperplane of a linear classifier separating all samples. Hence different splits represent different hyperplanes thus they focus on different properties of samples. As mentioned above, we want the attributes to be predictable and discriminative simultaneously. In this paper, we propose a unified optimization framework to **explicitly** take both of them into consideration as follows,

$$\begin{aligned}\min_{f_{va}, f_{ac}, \mathbf{A}} \; & \ell(f_{va}(\mathbf{X}), \mathbf{A}) + \mathcal{R}_1(f_{va}) \\ & + \lambda(\ell(f_{ac}(\mathbf{A}), \mathbf{Y}) + \mathcal{R}_2(f_{ac}))\end{aligned} \quad (2)$$

where $\ell(\cdot, \cdot)$ is a loss function and $\mathcal{R}(f)$ is the regularization term to control the complexity of classifier f to avoid overfitting. The first line in Eq. (2) considers the predictability of attributes while the second line focuses on the discriminability, and λ is balance parameter. Actually, the most predictable attributes (i.e., $\ell(f_{va}(\mathbf{X}), \mathbf{A}) = 0$), e.g., obtained

by setting attributes to the output of f_{va}, may contains little information of categories leading to poor discriminability. On the other hand, the most discriminative attributes (i.e., $\ell(f_{ac}(\mathbf{A}), \mathbf{Y}) = 0$), e.g., obtained by assigning identical attributes to all images of the same category, are difficult to predict from visual feature. Previous works mainly focus on one while ignoring the other, resulting in unsatisfactory attribute representation. In addition, the purpose of imposing attributes as the intermediary is for better visual recognition (or label prediction) from visual features. Our unified framework **explicitly** achieves this purpose, which is an essential difference from previous attribute learning works, such as (Rastegari, Farhadi, and Forsyth 2012) and (Yu et al. 2013).

The classifiers f_{va} and f_{ac} should not only fit the training data well, but also have generalization ability, i.e., they have satisfactory visual recognition performance on test data. Theoretically, classifiers with large margin always generalize well, such as SVM (Vapnik 1998). Therefore, we utilize this idea and guarantee the predictability by minimizing classification error and maximizing classification margin as

$$\begin{aligned} \min_{f_{va}} \ell(f_{va}) + \mathcal{R}_1 = \min_{\mathbf{w}_{va}, \xi} \sum_{i=1}^{k} (\|\mathbf{w}_{va}^i\|^2 + C_1 \sum_{j=1}^{n} \xi_{ij}) \\ \text{s.t. } a_{ij}(\mathbf{x}_j' \cdot \mathbf{w}_{va}^i) \geq 1 - \xi_{ij},\ \xi_{ij} \geq 0 \end{aligned} \tag{3}$$

where ξ_{ij} is the slack variable and C_1 is the regularization parameter. On the other hand, we can define a multi-class classification loss for $\ell(f_{ac}(\mathbf{A}), \mathbf{Y})$ on the labeled samples following the work in (Tsochantaridis et al. 2005) as below,

$$\ell(f_{ac}) = \sum_{i=1}^{n} \max_{j \neq j_i} \mathbf{a}_i' \cdot \mathbf{w}_{ac}^j - \mathbf{a}_i' \cdot \mathbf{w}_{ac}^{j_i} \tag{4}$$

where the i-th sample belongs to the j_i-th category. Moreover, we can define $\mathcal{R}_2(f_{ac}) = \sum_{i=1}^{C} \|\mathbf{w}_{ac}^i\|^2$. Now we have the specific objective function of our framework as follows,

$$\begin{aligned} &\min_{\mathbf{w}_{va}, \mathbf{w}_{ac}, \mathbf{A}, \xi} \sum_{i=1}^{k} (\|\mathbf{w}_{va}^i\|^2 + C_1 \sum_{j=1}^{n} \xi_{ij}) \\ &+ \lambda (\sum_{i=1}^{C} \|\mathbf{w}_{ac}^i\|^2 + C_2 \sum_{i=1}^{n} \max_{j \neq j_i} \mathbf{a}_i' \cdot \mathbf{w}_{ac}^j - \mathbf{a}_i' \cdot \mathbf{w}_{ac}^{j_i}) \\ &\text{s.t. } a_{ij}(\mathbf{x}_j' \cdot \mathbf{w}_{va}^i) \geq 1 - \xi_{ij},\ \xi_{ij} \geq 0 \end{aligned} \tag{5}$$

where C_1 and C_2 are the regularization parameters. Thus the predictability and the discriminability are explicitly considered for the learned attributes under this unified framework.

Intra-category Locality

Actually, considering predictability and discriminability simultaneously is still not enough. On one hand, Eq. (5) indeed aims to find a **balance** between predictability and discriminability. A good balance can result in effective attributes. However, it's also likely to achieve a bad balance in real world if there is no proper regularization. On the other hand, the relationship between images is ignored in Eq. (5). For example, it's expected that similar images should have similar attributes. Previous works usually require images of the same category to have identical (Farhadi et al. 2009; Lampert, Nickisch, and Harmeling 2009) or similar (Rastegari, Farhadi, and Forsyth 2012) attributes. Yet, this requirement is indeed too strict for real-world dataset because there is large intra-class variance such that images of the same category may have different attributes, as shown in Figure 1.

To address these issues, we propose to impose *intra-category locality* regularization into Eq. (5). Specifically, we require similar images of the same category to have similar attributes. This regularization takes two perspectives into consideration. First, because of the large intra-category variance, it's meaningless and harmful to require all images of the same category to have similar attributes. Thus we have the constraint "*similar*". Second, if we require similar images, regardless of their category information, to have similar attributes, the learned attributes will be indiscriminative. Actually, we still wish images of different categories to have different attributes for more discriminability, even though they are visually similar. This is represented by the constraint "*the same category*". In addition, this regularization term makes our method generalize better, which is also a quite important property for real-world applications. Specifically, at first we need to construct a nearest neighbor matrix to capture the intra-category locality information as follows,

$$W_{ij} = \begin{cases} 1, & (\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_j) \vee \mathbf{x}_j \in \mathcal{N}(\mathbf{x}_i)) \wedge \mathbf{y}_i = \mathbf{y}_j \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $\mathcal{N}(\mathbf{x}_i)$ is the p nearest neighbor of $\mathbf{x}_i$. Therefore, the intra-category locality regularization is formulated as below,

$$\mathcal{R}_{\text{ICL}} = \frac{1}{2} \sum_{i,j=1}^{n} W_{ij} \|\mathbf{a}_i - \mathbf{a}_j\|^2 = tr(\mathbf{A}\mathbf{L}\mathbf{A}^T) \tag{7}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix of $\mathbf{W}$ and $\mathbf{D}$ is a diagonal matrix with diagonal element $D_{ii} = \sum_{j=1}^{n} W_{ij}$. Now we can incorporate this intra-category locality regularization into the learning framework as shown in Eq. (5). Thus we can obtain the overall objective function as follows,

$$\begin{aligned} &\min_{\mathbf{w}_{va}, \mathbf{w}_{ac}, \mathbf{A}, \xi} \sum_{i=1}^{k} (\|\mathbf{w}_{va}^i\|^2 + C_1 \sum_{j=1}^{n} \xi_{ij}) + \alpha \mathcal{R}_{\text{ICL}} \\ &+ \lambda (\sum_{i=1}^{C} \|\mathbf{w}_{ac}^i\|^2 + C_2 \sum_{i=1}^{n} \max_{j \neq j_i} \mathbf{a}_i' \cdot \mathbf{w}_{ac}^j - \mathbf{a}_i' \cdot \mathbf{w}_{ac}^{j_i}) \\ &\text{s.t. } a_{ij}(\mathbf{x}_j' \cdot \mathbf{w}_{va}^i) \geq 1 - \xi_{ij},\ \xi_{ij} \geq 0 \end{aligned} \tag{8}$$

where α is the regularization parameter. Thus our PDA can explicitly consider predictability and discriminability. And with the intra-category locality regularization, PDA can overcome the influence of large intra-category variance and noise in real-world visual dataset, and result in good balance between predictability and discriminability. Furthermore, our learning framework has an significant difference from previous works, i.e., images of the same category may have quite different attributes. We think this property is very reasonable and important because of the large intra-category variance. And as validated in our experiment, this property can indeed promote the recognition performance observably.

Algorithm 1 Learning Algorithm

Input:

Image data $\mathbf{X}$, labels $\mathbf{Y}$, #attributes k.

Output:

Two groups of classifiers, $\mathbf{w}_{va}$ and $\mathbf{w}_{ac}$.

1: Initialization: $\mathbf{A} \leftarrow$PCA$(\mathbf{X}, k)$,
2: Binarization: $\mathbf{A} \leftarrow \text{sign}(\mathbf{A})$.
3: Construct $\mathbf{W}$ by Eq. (6), a diagonal matrix with diagonal element $D_{ii} = \sum_{j=1}^{n} W_{ij}$ and $\mathbf{L} = \mathbf{D} - \mathbf{W}$.
4: Training classifiers: learn SVM classifiers $\mathbf{w}_{va}$ and $\mathbf{w}_{ac}$.
5: **repeat**
6: Optimize $\mathbf{A}$ greedily to minimize Eq. (8) by block coordinate descent algorithm.
7: Update classifiers $\mathbf{w}_{va}$ and $\mathbf{w}_{ac}$.
8: **until** Convergence
9: Return $\mathbf{w}_{va}$ and $\mathbf{w}_{ac}$.

Learning Algorithm

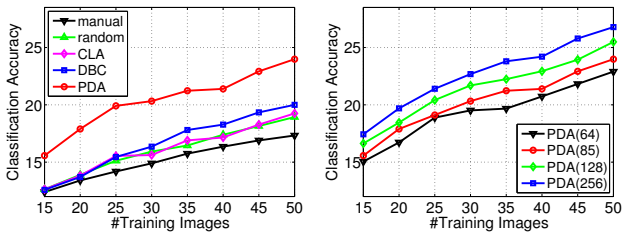
The optimization problem of Eq. (8) is difficult. Fortunately, we don't need to find the global minimum, and good local minimum can always result in satisfactory performance. Therefore we can employ an iterative strategy for optimization. The learning algorithm is summarized in Algorithm 1.

This algorithm has the following steps. Suppose the data matrix $\mathbf{X}$ has been centralized, i.e., $\sum_{i=1}^{n} \mathbf{x}_i = \mathbf{0}$. First we project $\mathbf{X}$ to a k-dimensional space by Principle Component Analysis (PCA). Then we can threshold the projected data to obtain the initial attribute $\mathbf{A}$. As the data is centralized, the attribute is balanced, i.e., $\mathbf{A1} \approx \mathbf{0}$. And the directions selected by PCA are orthogonal, thus attributes are uncorrelated, i.e., $\mathbf{A}\mathbf{A}^T \approx n\mathbf{I}_{k\times k}$. Then we can train initial classifiers $\mathbf{w}_{va}$ using $\mathbf{X}$ and $\mathbf{A}$, and $\mathbf{w}_{ac}$ with $\mathbf{A}$ and $\mathbf{Y}$. Second, we can construct $\mathbf{W}$ by Eq. (6) and $\mathbf{L}$. Third, we can iteratively adjust $\mathbf{A}$ in a greedy way by block coordinate descent (Richtárik and Takác 2014) and adjust classifier parameters by retrain linear SVM classifiers with $\mathbf{X}$, $\mathbf{A}$ and $\mathbf{Y}$. Theoretically, we need the algorithm to converge to reach local minimum. But in real-world scenarios, we find out that 5 to 10 iterations are enough to achieve satisfactory performance.

Experiment and Discussion

Image Classification on AwA Dataset

The Animal with Attributes (AwA) dataset (Lampert, Nickisch, and Harmeling 2009) contains $30,475$ images from 50 animal categories. It has 85 manually designed attributes, such as *black* and *big*, and each category is labeled with these attributes. We use the pre-computed low-level visual features consisting of RGB histogram, SIFT, rgSIFT, PHOG, SURF and local-similarity histogram. All features are first normalized to unit length individually and then concatenated into a single feature vector with $10,940$ dimensions. For the implementation efficiency, we apply PCA to reduce the feature dimension to $1,024$. Following the setting in previous works, we change the numbers of images from each category $(15, ..., 50)$ when training models, and select other 10 images per category for validation. Then all of the rest images form the test set. To verify the effectiveness of PDA, we compare it to the following baselines, manual attributes (Lampert, Nickisch, and Harmeling 2009), random attributes, Category-level Attributes (Yu et al. 2013), and Discriminative Binary Codes (Rastegari, Farhadi, and Forsyth 2012). For the comparison fairness, CLA, DBC and our PDA will learn 85 attributes. When implementing CLA and DBC, we carefully tuned their model parameters for each experiment and the best results are reported. And for PDA, we consistently set $\lambda = 1$ to balance predictability and discriminability, $\alpha = 0.1$ for intra-category locality regularization, p to 4 times of the number of training images per category for constructing nearest neighbor matrix $\mathbf{W}$, and the classifier regularization parameters are all set to 0.1.



(a) Compare to baselines (b) The effect of #attributes

Figure 2: Classification Accuracy on AwA

First, we compare PDA to all baselines, whose results are shown in Figure 2(a). We can observe that our PDA can significantly and consistently outperform all baselines with different sizes of training data, which verifies the effectiveness and superiority of PDA. The experiment results also reveal some important points as below. The performance of manual attributes is worst of all, which has been widely mentioned in previous works (Yu et al. 2013), because they are neither discriminative nor predictable. CLA and DBC can improve the performance to some extent because their attributes show more discriminability and predictability respectively. However, CLA almost ignores the predictability while DBC just achieves discriminability implicitly, therefore their performance is still unsatisfactory. In addition, all baseline methods neglect the large intra-category variance of image data, which markedly degrade their performance. In a summary, our PDA explicitly takes predictability and discriminability into account simultaneously, and addresses intra-category variance by exploiting the intra-category locality. Thus PDA can achieve best classification performance on AwA dataset.

Second, we investigate the effect of the number of attributes on the classification performance of PDA. The result is presented in Figure 2(b). With more attributes, PDA achieves better performance. This is reasonable because more attributes can encode more inherent discriminative information. In addition, the initial attributes are generated by binarizing the projection of PCA, thus the attributes are uncorrelated to each other initially. After just few iterations, the correlation between attributes is still quite small. Therefore more attributes can lead to better performance. However, if we keep increasing the number of attributes, the improvement will become less. This is because we use PCA for initialization, and it's well-known that for most real-world dataset, the variance (discriminative information) is always

Table 2: Classification accuracy of variants of PDA

Size	15	20	30	40	50
DBC	12.58	13.70	16.35	18.28	20.00
PDA-0	13.11	14.22	16.52	17.34	19.73
PDA-ALL	13.48	16.00	18.88	19.18	21.35
PDA	**15.57**	**17.89**	**20.33**	**21.39**	**23.98**

Table 3: Training time on AwA (in seconds)

Size	15	20	30	40	50
DBC	48.46	66.14	97.75	115.52	128.94
PDA	26.97	38.04	69.74	99.79	122.72

contained in first few projections. Hence more attributes will pick directions with low variance so fewer discriminative information can be obtained (Wang, Kumar, and Chang 2010).

Third, we conduct experiment to validate our claim that intra-category locality can overtly improve the performance. Here, we compare PDA to the best baseline, DBC, and two variants of PDA, PDA-0 by setting $p = 0$ or $\alpha = 0$ thus it considers no intra-category locality, and PDA-ALL by requiring all images of the same category to have similar attributes which ignores the "locality". The number of attributes is fixed to 85. The result is summarized in Table 2. We can observe the following points from this result. First, DBC can be regarded as a special case of PDA-ALL by setting $\lambda = 0$ which just implicitly considers the discriminability. We can observe that PDA-ALL achieves better performance than DBC, which validates the necessity of explicitly modeling the discriminability. Second, PDA-0, which ignores the relationship between samples, just achieve comparable result to DBC. Though the predictability and discriminability are explicitly considered in PDA-0, "bad" balance is finally obtained and it has low generalization ability. Therefore the performance is unsatisfactory. And at last, by comparing PDA to its two variants, we can observe that considering the relationship between samples can improve the generalization ability (PDA and PDA-ALL vs. PDA-0), and exploiting locality is more effective than global information (PDA vs PDA-ALL) because of the large intra-category variance of real-world data. By summarizing these observations, we can draw the conclusion that exploiting intra-category locality is indeed salutary for better recognition performance.

Fourth, to evaluate the efficiency of Algorithm 1, we compare the training time of PDA and DBC on AwA dataset with different training size. For PDA, the number of iterations is fixed to 5 as in all other experiments which can lead to satisfactory recognition performance and the other parameters are set as introduced above. The quantitative results are summarized in Table 3. Generally, it takes less time for PDA to train models than DBC, and the training time of PDA increases linearly with training size, validating the efficiency of Algorithm 1. All results above are obtained on a computer which has Intel Core i7-2600 3.40GHz CPU and 8GB RAM.

Last but not the least, we visualize some interesting and meaningful attributes (splits) learned by PDA in Figure 3, where each row corresponds to one attribute. For a specific attribute i, we select images with the largest (resp. smallest) value of $\mathbf{x}'_j \cdot \mathbf{w}^i_{va}$, i.e., images with highest attribute classification confidence, and put them on the left (resp. right) side in this figure. We can observe that the learned attributes are indeed discriminative, and even semantically meaningful and interpretable to some extent. For example, three attributes shown here are maybe corresponding to "Green", "Water" and "With human" respectively. Furthermore, as shown on the left side of the each row, these four images may belong to totally different categories, e.g.,four images in the first row are from "Wolf", "Deer", "Lion" and "Tiger" respectively. But actually they are visually similar and ought to have similar attributes. Therefore it's unreasonable to require images of the different categories to have different attributes, i.e., large inter-category distance, which is an important requirement in CLA and DBC. This is also an important reason why we just exploit the intra-category locality but dismiss the inter-category distance in learning function.

Figure 3: Visualization of the discovered attributes (learned splits) by PDA. Each row corresponds to one attribute (split).

Image Retrieval on Caltech256 Dataset

Besides image classification, the learned binary codes can also support large-scale image retrieval like Hashing which is highly efficient in both storage and time. Since each binary attribute can be represented by one bit, it takes just about 1GB memory to store all attribute representations of 32 million images even with 256 attributes. In addition, it's also quite efficient to compute the Hamming distance between a query and all images in database whose attribute representations are stored in memory, since only bit XOR operations are required, which is also frequently mentioned in previous research about Hashing, such as (Gong and Lazebnik 2011).

In this paper, we conduct image retrieval experiments with binary attributes (or hash codes) on Caltech256 dataset. This dataset contains 256 categories and $30,607$ images. The retrieve task on Caltech256 is very challenging because it has a large number of categories and the intra-category variance is also very large as shown in Figure 1. Moreover, there are only about 120 images per category, which further increases the difficulty. In our experiment, we represent each image by a $2,048$-dimension feature obtained from Locality-constrained Linear Coding (Wang et al. 2010) based on SIFT (Lowe 1999) local descriptor because it achieves state-of-the-art performance for image representation. In addition, we select 20 images per category (totally $5,120$) as the query images and the rest as the database. We can train models on database and generate binary representations for images in both database and query set with the learned models for distance measure. Finally the database will first return images with shorter distance to each query image as retrieval results.

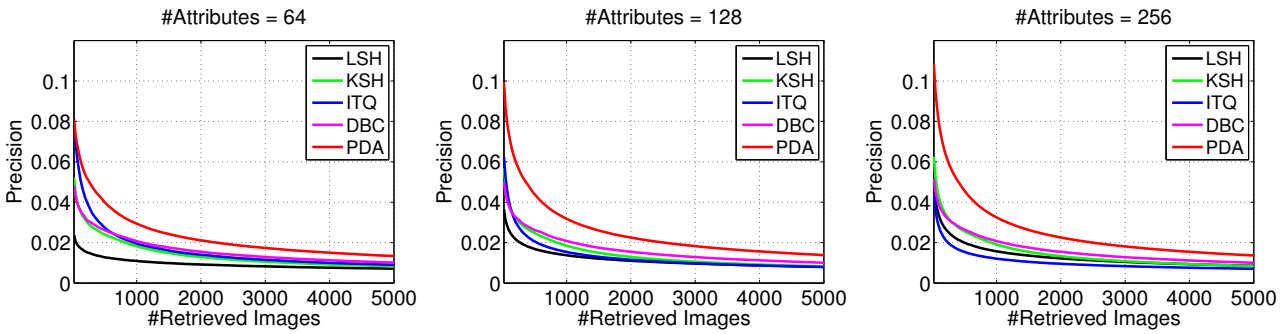


Figure 4: Quantitative Comparison between Methods for Image Retrieval on Caltech256 Dataset.

We compare our PDA with several binary representation learning methods. Locality Sensitive Hashing (LSH) (Andoni and Indyk 2006) is chosen as the base method. And we also choose two state-of-the-art supervised Hashing methods, Supervised Hashing with Kernels (KSH) (Liu et al. 2012), and Iterative Quantization (ITQ) (Gong and Lazebnik 2011) based on CCA (Hotelling 1936), because they all utilize the label information. Also, we compare PDA to DBC, which are both binary attribute learning methods. We adopt *precision curve*, which reflects the precision level (the ratio of relevant images in retrieved images) with respect to the number of retrieved images, as the evaluation metric as in (Gong and Lazebnik 2011). Moreover, images sharing the same label given by the dataset are regarded as relevant. For baselines, we carefully tuned their model parameters and the best results are reported. For PDA, we set $\lambda = 1$, $\alpha = 0.1$, $p = 1,000$, and classifiers' regularization parameters to 0.1.

First we quantitatively compare the performance of different methods with different number of attributes. The precision curves are plotted in Figure 4. We can observe that our PDA can significantly outperform all baseline methods regardless of the number of attributes, which validates the effectiveness of PDA for image retrieval. Furthermore, when we increase the number of attributes, all methods except ITQ show better performance, and PDA can achieve more promotion compared to other baselines because PDA can acquire more discriminative information with more attributes.

Actually, we can also observe an important point from the results, i.e., the performance of ITQ, DBC and PDA suggests that explicitly exploit the discriminative information and address the intra-category variance are both quite important for designing effective binary attribute representations. Specifically, ITQ adopts CCA to exploit the discriminative information. But it performs rotation to minimize the quantization error **after** CCA, which may result in worse discrimination. DBC proposes to exploit discriminative information together with attribute classifier learning (or Hash function learning). Hence more discrimination can be preserved by DBC compared to ITQ which may lead to better performance. However, DBC achieves this just in an **implicit** way and ignores the intra-category variance. Our PDA explicitly achieves this goal such that it can learn more discriminative attributes than DBC. In addition, PDA is able to address the intra-category variance by exploiting the intra-category locality. Consequently, PDA can achieve much better performance in comparison to baselines, including DBC and ITQ.

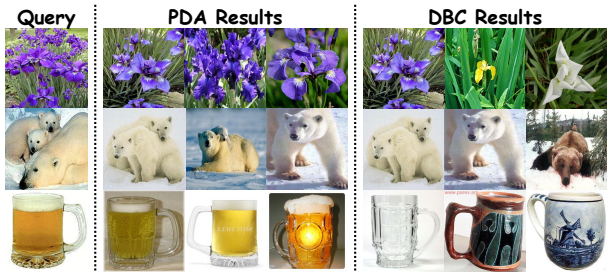


Figure 5: Qualitative comparison between PDA and DBC. We show the first three **relevant** results to the query images.

Moreover, we qualitatively compare PDA to the best baseline method DBC. Specifically, in the retrieved images, we select the first three **relevant** images to the query, i.e., images with the same label as and most similar attribute representations to the query image. Some results are shown in Figure 5. The results are meaningful. Actually, all images shown here have the same label as each query respectively, but some retrieved images by DBC are quite unsatisfactory. In real-world applications, it's expected to retrieve images with not only the same label, but also more visual similarity. However, DBC requires all images of the same category to have similar attributes such that visually dissimilar images may have similar attributes. Thus, though DBC and other methods may achieve high precision for image retrieval measured only by labels, their quality is still unsatisfactory. But our PDA focuses on the intra-category locality, making just similar images have similar attributes. Hence PDA can simultaneously achieve high precision and superior retrieval quality, which makes PDA a practical method for real world.

Conclusion

In this paper, we propose a novel method for learning predictable and discriminative attributes for visual recognition. We propose to explicitly and simultaneously take predictability and discriminability into consideration. More importantly, we propose to exploit intra-category locality to overcome the large intra-category variance which is ignored by most previous works. These ideas are modeled in a unified framework and can be solved efficiently by the learning algorithm. We conduct both image classification and efficient image retrieval experiments on AwA and Caltech256 datasets respectively. Extensive experimental results demonstrate the superiority of PDA compared to several state-of-the-art related methods, verifying the effectiveness of PDA.

Acknowledgements

This research was supported by the National Basic Research Project of China (Grant No. 2011CB70700), the National Natural Science Foundation of China (Grant No. 61271394), and the National HeGaoJi Key Project (No. 2013ZX01039-002-002). In the end, the authors would like to sincerely thank the reviewers for their valuable comments and advice.

References

Akata, Z.; Perronnin, F.; Harchaoui, Z.; and Schmid, C. 2013. Label-embedding for attribute-based classification. In *CVPR*.

Andoni, A., and Indyk, P. 2006. Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *FOCS*.

Belkin, M., and Niyogi, P. 2001. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, 585–591.

Belkin, M.; Niyogi, P.; and Sindhwani, V. 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research* 7:2399–2434.

Cai, D.; He, X.; Han, J.; and Huang, T. S. 2011. Graph regularized nonnegative matrix factorization for data representation. *IEEE Trans. Pattern Anal. Mach. Intell.* 33(8):1548–1560.

Farhadi, H.; Endres, I.; Hoiem, D.; and Forsyth, D. 2009. Describing objects by their attributes. In *CVPR*.

Gionis, A.; Indyk, P.; and Motwani, R. 1999. Similarity search in high dimensions via hashing. In *VLDB*.

Gong, Y., and Lazebnik, S. 2011. Iterative quantization: A procrustean approach to learning binary codes. In *CVPR*.

Hotelling, H. 1936. Relations between two sets of variables. *Biometrika* 28:312–377.

Kumar, N.; Berg, A.; Belhumeur, P.; and Nayar, S. 2009. Attribute and simile classifiers for face verification. In *CVPR*.

Lampert, C. H.; Nickisch, H.; and Harmeling, S. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*.

Liu, W.; Wang, J.; Ji, R.; Jiang, Y.; and Chang, S. 2012. Supervised hashing with kernels. In *CVPR*.

Lowe, D. G. 1999. Object recognition from local scale-invariant features. In *ICCV*.

Rastegari, M.; Farhadi, A.; and Forsyth, D. A. 2012. Attribute discovery via predictable discriminative binary codes. In *ECCV*.

Richtárik, P., and Takác, M. 2014. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Math. Program.* 144(1-2):1–38.

Roweis, S., and Saul, L. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500):2323–2326.

Siddiquie, B.; Feris, R.; and Davis, L. 2011. Image ranking and retrieval based on multi-attribute queries. In *CVPR*.

Tsochantaridis, I.; Joachims, T.; Hofmann, T.; and Altun, Y. 2005. Large margin methods for structured and interdependent output variables. *Journal of Machine Learning Research* 6:1453–1484.

Vapnik, V. 1998. *Statistical learning theory*. Wiley.

von Luxburg, U. 2007. A tutorial on spectral clustering. *Statistics and Computing* 17(4):395–416.

Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T. S.; and Gong, Y. 2010. Locality-constrained linear coding for image classification. In *CVPR*.

Wang, J.; Kumar, O.; and Chang, S. 2010. Semi-supervised hashing for scalable image retrieval. In *CVPR*.

Yu, F. X.; Cao, L.; Feris, R. S.; Smith, J. R.; and Chang, S.-F. 2013. Designing category-level attributes for discriminative visual recognition. In *CVPR*.