Designing Category-Level Attributes for Discriminative Visual Recognition

Felix X. Yu+ Liangliang Cao* Rogerio S. Feris* John R. Smith* Shih-Fu Chang+

+Columbia University *IBM T.J. Watson Research Center

1. Attributes in Computer Vision

"Attribute" often refers to human nameable properties (e.g., furry, striped, black) that are shared across categories.

- nocturnal
- live in trees
- have large eyes
- have long middle fingers

- not nocturnal
- do not live in trees
- (Maybe) large eyes
- no long middle fingers

Application 1: Describing images/ semantic mid-level features

Application 2: Zero-shot Learning

Define novel categories in terms of the existing attributes (by user or animal specialist): Aye-ayes are • nocturnal • live in trees • have large eyes • have long middle fingers

More Applications: Image Retrieval, face verification, action recognition, rate event detection…

*Picture credit: C. Lampert

The "Classic" steps for attribute-based recognition:

- Manually picking (designing) a set of words as attributes.
- Label those attributes on some images.
- Train the attribute classifiers.

The Problems:

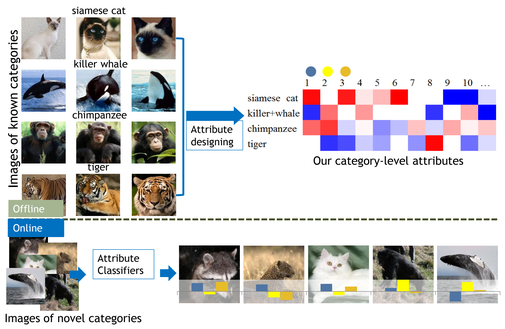
- Expensive!
- The manually designed attributes may not be discriminative.

2. Summary of Our Approach

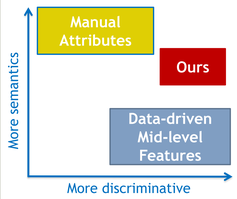
In this paper, we propose a **scalable** approach of **automatically** designing attributes for **discriminative** visual recognition.

Summary of the Idea:

- Define attributes as associations with a set of known categories (a category-attribute matrix), *e.g., dogs and cats have it but sharks and whales don't.*
- Optimize the category-attribute matrix to make the representation discriminative.



3. Related Works



☺**Fully automatic**, with a multi-class dataset

☺**Discriminative**

☺State-of-the art results in novel category recognition including zero-shot learning, and learning with few examples.

☹No clear semantic meanings (but they can be described in terms of known categories).

4. A Learning Framework for Visual Recognition with Category-Level Attributes

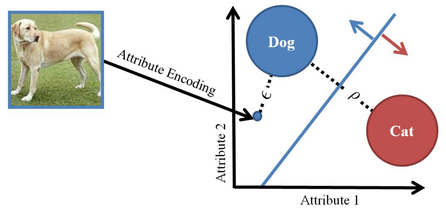
- The category-attribute matrix is denoted as $\mathbf{A} \in \mathbb{R}^{k \times l}$ (k categories, and l attributes).

Definition 1. *For an input image $\mathbf{x} \in \mathcal{X}$ (as low-level features), we define the following two steps to utilize attributes as mid-level cues to predict its category label $y \in \mathcal{Y}$.*

***Attribute Encoding**: Compute l attributes by attribute classifiers $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), ..., f_l(\mathbf{x})]^T$ in which $f_i(\mathbf{x}) \in \mathbb{R}$ models the strength of the i-th attribute for $\mathbf{x}$.*

***Category Decoding**: Choose the closest category (row of $\mathbf{A}$) in the attribute space (column space of $\mathbf{A}$):*

$$\arg\min_i \parallel \mathbf{A}_{i\cdot} - \mathbf{f}(\mathbf{x})^T \parallel .$$



Definition 2. *Define ϵ as the average encoding error of the attribute classifiers $\mathbf{f}(\cdot)$, with respect to the category-attribute matrix $\mathbf{A}$.*

$$\epsilon = \frac{1}{m}\sum_{i=1}^{m} \parallel \mathbf{A}_{y_i\cdot} - \mathbf{f}(\mathbf{x}_i) \parallel .$$

Definition 3. *Define ρ as the minimum row separation of the category-attribute matrix $\mathbf{A}$*

$$\rho = \min_{i \neq j} \parallel \mathbf{A}_{i\cdot} - \mathbf{A}_{j\cdot} \parallel .$$

Theorem 1. *The empirical error of multi-class classification is upper bounded by $2\epsilon/\rho$.*

Characters of good attributes:

- **Category-separability:** Large ρ.
- **Learnability:** Small ϵ. This also implies that attributes should be shared across "similar" categories.
- **Non-redundancy:** $r = \frac{1}{l} \parallel \mathbf{A}^T\mathbf{A} - \mathbf{I} \parallel_F^2$.

5. The Attribute Designing Algorithm

5.1 Learning the Category-attribute Matrix

$$\max_{\mathbf{A}} J(\mathbf{A}) = J_1(\mathbf{A}) + \lambda J_2(\mathbf{A}) + \beta J_3(\mathbf{A})$$

- $J_1(\mathbf{A}) = \sum_{i,j} \parallel \mathbf{A}_{i\cdot} - \mathbf{A}_{j\cdot} \parallel_2^2$
- $J_2(\mathbf{A}) = -\sum_{i,j} S_{ij} \parallel \mathbf{A}_{i\cdot} - \mathbf{A}_{j\cdot} \parallel_2^2$
- $J_3(\mathbf{A}) = - \parallel \mathbf{A}^T\mathbf{A} - \mathbf{I} \parallel_F^2$

The visual proximity matrix: $\mathbf{S} \in \mathbb{R}^{k\times k}$: $S_{ij} = e^{-D_{ij}/\sigma}$.

- When nonlinear kernels are used, SVM margins, of $k(k-1)/2$ *one-vs-one* SVMs modeled on low-level features are used as distance measurement for categories.
- When linear kernels are used we use the distances of category centers (category mean of the low-level features) as distance measurements (complexity linear to # images).

We propose to incrementally learn the columns of $\mathbf{A}$. Given an initialized $\mathbf{A}$, optimizing an additional column $\mathbf{a}$:

$$\max_{\mathbf{a}} \quad \mathbf{a}^T\mathbf{R}\mathbf{a} \qquad \text{s.t.} \quad \mathbf{a}^T\mathbf{a} = 1,$$

in which $\mathbf{R} = \mathbf{Q} - \eta\mathbf{A}\mathbf{A}^T$, $\mathbf{Q} = \mathbf{P} - \lambda\mathbf{L}$, $\mathbf{P}$ is with diagonal elements being $k-1$ and all the other elements -1, and $\mathbf{L}$ is the Laplacian of $\mathbf{S}$, $\eta = 2\beta$. This is a Rayleigh quotient problem, with the optimal $\mathbf{a}$ as the eigenvector of $\mathbf{R}$ with the largest eigenvalue.

5.2 Learning the Attribute Classifiers: Weighted SVM

$$\min_{\mathbf{w}_i,\xi} \quad \parallel \mathbf{w}_i \parallel_2^2 + C\sum_{j=1}^{m} |A_{y_j,i}|\xi_j$$

$$\text{s.t.} \quad \text{sign}(A_{y_j,i})\mathbf{w}_i^T\mathbf{x}_j \geq 1 - \xi_j$$

$$\xi_j \geq 0, \quad j = 1...m$$

5.3 Efficiency

The computational complexity of designing an attribute (a column of $\mathbf{A}$) is as efficient as finding the eigenvector with the largest eigenvalue of matrix $\mathbf{R}$: 1 hour to design 2,000 attributes based on 950 categories on the large-scale ILSVRC2010 dataset.

6. Experiments

Datasets: AwA (30k images, 50 categories) and ILSVRC2010 (1M images, 1k categories)

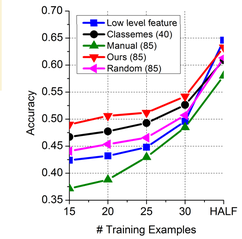
6.0 Verifying the Attribute Designing Criterion (AwA)

Measurement	Designed	Manual	Random
Encoding error ϵ	**0.03**	0.07	0.04
Minimum row separation ρ	**1.37**	0.57	1.15
Average row separation	**1.42**	1.16	1.41
Redundancy r	**0.55**	2.93	0.73

(# Attributes: 85)

6.1 The designed attributes are discriminative for novel, yet related categories.

AwA: Use 40 categories to design the attributes. Then use attributes as feature for recognizing the remaining 10 categories, with different # training images.



6.2 The designed attributes are discriminative for general novel categories (ILSVRC2010), if we can design a large amount of attributes based on a diverse set of known categories. We train attributes based on 950 categories.

Method	Precision@50
Low-level feature	33.40
Classeme (950)	39.24
Ours (500)	39.85
Ours (950)	42.16
Ours (2,000)	**43.10**

Category-level image retrieval result on 50 classes

	Percentage for training				
Method	1%	5%	10%	50%	100%
Low-level feature	35.55	52.21	57.11	66.21	**69.16**
Classemes (950)	38.54	51.49	56.18	64.31	66.77
Ours (500)	39.01	52.86	56.54	62.38	63.86
Ours (950)	41.60	55.32	59.09	65.15	66.74
Ours (2,000)	**43.39**	**56.51**	**60.36**	**66.91**	68.17

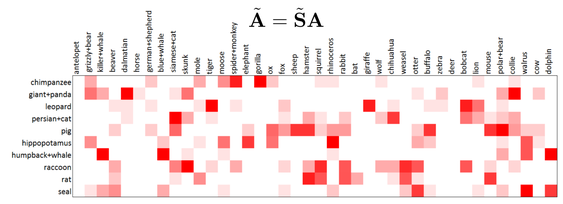
Image classification accuracy on 50 classes. The training set contains 54,636 images.

6.3 The attributes are effective for zero-shot learning

Given each novel category, and 40 known categories, we ask the user to find the top-5 visually similar categories: $\tilde{\mathbf{S}} \in \{0,1\}^{p\times k}$, in which $\tilde{S}_{ij}$ is the binary similarity of the i-th novel category and the j-th known category.

The novel categories are related to the designed attributes by the simple weighted sum:

$$\tilde{\mathbf{A}} = \tilde{\mathbf{S}}\mathbf{A}$$



Similarity matrix by averaging the results of 10 graduate students

We can then follow the proposed framework to do zero-shot learning:

Method	# Attributes	Accuracy
Lampert *et al.* [13]	85	40.5
Yu and Aloimonos [35]	85	40.0
Rohrbach *et al.* [23]	-	35.7
Kankuekul *et al.* [11]	-	32.7
Ours	10	40.52 ± 4.58
Ours	85	42.27 ± 3.02
Ours	200	42.83 ± 2.92
Ours (Fusion)	200	**46.94**
Ours (Adaptive)	200	45.16 ± 2.75
Ours (Fusion + Adaptive)	200	**48.30**

Zero-shot learning result on AwA (40 categories for training, 10 categories for testing)

Adaptive Attribute Design: $\tilde{J}_1(\mathbf{A}) = J_1(\tilde{\mathbf{S}}\mathbf{A})$.

More information: http://www.ee.columbia.edu/~yuxinnan/category.html