Decorrelating Semantic Visual Attributes by Resisting the Urge to Share

— **Source link** ⧉

Dinesh Jayaraman, Fei Sha, Kristen Grauman

Institutions: University of Texas at Austin, University of Southern California

Related papers:

- Describing objects by their attributes

- Learning to detect unseen object classes by between-class attribute transfer

- Relative attributes

- Attribute-Based Classification for Zero-Shot Visual Object Categorization

- An embarrassingly simple approach to zero-shot learning

Decorrelating Semantic Visual Attributes by Resisting the Urge to Share

Supplementary material for CVPR 2014 submission ID 0824

In this document, we provide supplementary material for our CVPR 2014 submission "Decorrelating Semantic Visual Attributes by Resisting the Urge to Share"(Paper ID 0824). Sec 1 gives additional details for our experimental setup (Sec 4 of the paper). Sec 1.1 lists the groups used in all three datasets in our experiments. Sec 1.2 discusses the details of the image descriptors used for each dataset. Sec 2 discusses how attributes are localized for our experiments in Sec 4.1 in the paper. Sec 3 discusses how it is posible to set parameters that generalize well to novel test sets, using only training data. Sec 4 discusses the details of the optimization of our formulation (Eq 4 in the paper).

1 Datasets

1.1 Groups

(see para on Semantic groups in Sec 4 in the paper)

Fig 1, 2 and 3 show the attribute groups used in our experiments on the CUB, AwA and aPY datasets respectively. The 28 CUB groups come pre-specified with the dataset [6]. The groups on AwA match exactly the groups specified in [5]. Those on aPY also match the groups outlined in [5] on the 25 attributes (see paper) used in our experiments (aPY-25). In each figure, attribute groups are enclosed in shaded boxes, and phrases in larger font labeling the boxes indicate the rationale for the grouping.

1.2 Features

(see also Sec 3.2 and para on Features in Sec 4 in the paper)

Base level features (before PCA) : For aPascal, we extract features with the authors' code [3] for each of 1x1+3x2=7 windows (again as in [3]). We append histograms of color (128 words), texture (256 words), HOG (1000 words), edge (9 words). For CUB, we use the same feature extraction code but instead of dividing the image in a rectangular grid as for aPY, we use windows centered at each of the 15 parts ((x, y) annotations provided with the dataset). The window dimensions at each part are set to 1/3rd (in both width and height) of the bird bounding box. Here, our histogram lengths are 150, 250 and 500 respectively for color, texture and HOG on each window. For AwA, we use the features supplied with the dataset ((1) global bag-of-words on SIFT, RG-SIFT, SURF, local self-similarity, together with (2) 3-level pyramid with 4×4+2×2+1=21 windows on PHOG, RGB color histograms).

2 Experiments: Learning the right thing for attribute detection

(see also Sec 4.1, para "Evidence of learning the right thing" in the paper)

We now explain the details of the attribute localization experiment described in the paper. To find locations on instance number n that contribute to positive detection of attribute m, we take the absolute value of the element-wise product of descriptor $\mathbf{x}_n$ with the attribute weight vector $\mathbf{w}^m$—denote this $\mathbf{h}$. Each feature dimension is mapped onto the bird part it was computed from, in a mapping f. For each part p, we then compute its weight as $l_p = \sum_{f(i)=p} |h_i|$. These part weights are visualized as highlights in Fig 10 in the paper.
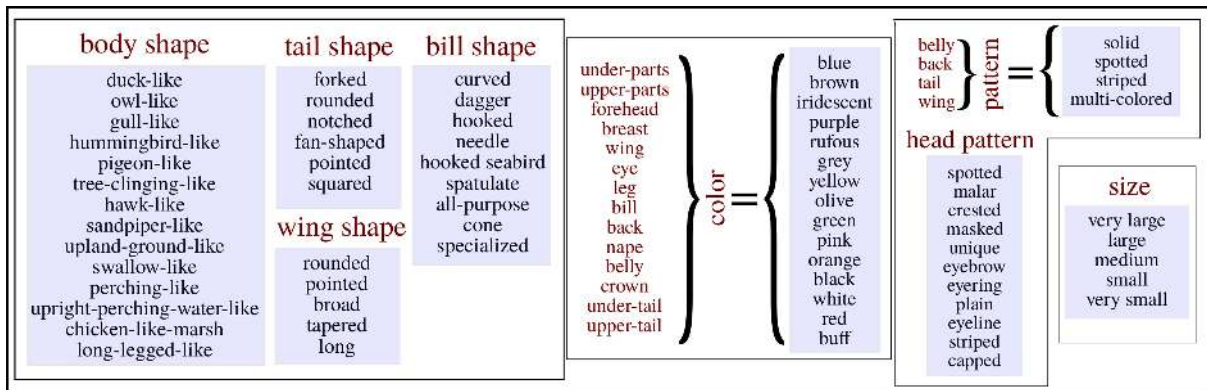
1

Fig 1: Attribute grouping on CUB. The color and pattern groups, condensed above, are to be interpreted as follows. Each part on the left, coupled with the term in the middle (color/pattern) represents the title of an attribute group. The predicates on the right applied to each part constitute the attributes in the its group, *e.g.*, the "belly-color" attribute group has attributes "belly-color-blue", "belly-color-brown" *etc.*
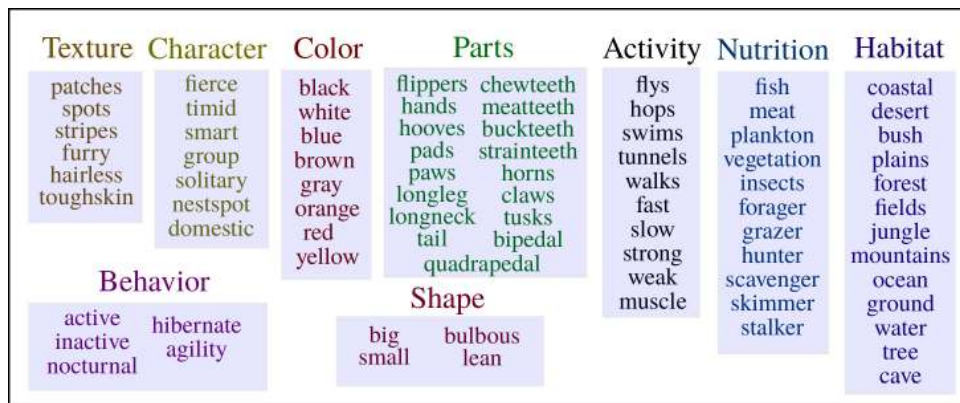

Fig 2: Attribute grouping on AwA

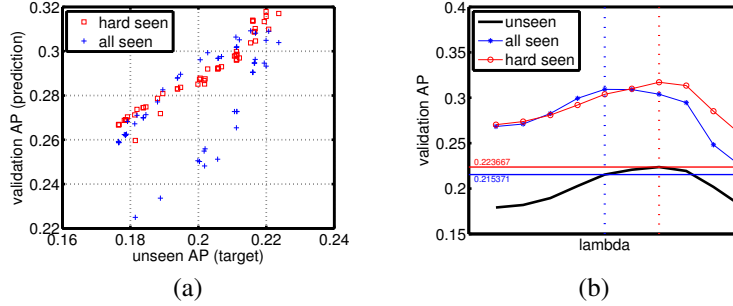

Fig 3: Attribute grouping on aPY-25

Fig 4: **(a)** "Hard seen" performance is a much better predictor of "unseen" set performance than "all seen" performance. **(b)** The optimal parameter λ on hard seen data (unlike all seen data) is also nearly optimal on unseen data. The vertical dashed lines pass through the peaks of the two performance curves. Solid horizontal lines through the intersection of these with the generalization performance curve mark the generalization performance of the corresponding models.

3 Experiments: Target-agnostic parameter selection

As mentioned in the paper, the aim of decorrelating attributes while learning attribute classifiers is to be able to generalize better to datasets different in distribution from the training dataset. In practice though, traditional parameter validation as a method of model selection requires access to (a subset of) the target dataset. Specifically, one selects parameters that produce best results when tested on a validation set drawn from the target distribution.

We argue that parameter selection tailored to a specific validation set distribution is not useful in real world situations where test distributions are (1) different from the training distribution and (2) unavailable during training and model selection. In this section, we study whether it is possible to select parameters for attribute-learning methods that generalize well to unknown novel distributions (different from the training distribution) even without any prior knowledge of/access to the test dataset distribution. Specifically, we show empirically that such target-agnostic parameter-selection is viable for our attribute decorrelation approach.

In Fig 4, we plot unseen test set performance at various regularization weights λ for the proposed method on CUB-200-2011 against equal-sized subsets (200 instances) of the hard seen and all-seen sets. While all-seen performance correlates poorly with generalization performance, hard-seen set performance is a much better indicator of unseen set performance. We verified that empirically determined optimal λ on the hard seen set is almost exactly optimal over the unseen set, for all three datasets - CUB, AwA and aPY-25. Further, we observed that such target-agnostically selected parameters over hard seen data for *other* methods were often highly *suboptimal* over the unseen set.[1]

These observations are explained as follows. For a learning method that decorrelates attributes, the best model in terms of generalization to an unknown test set is one that resolves correlations in the training set—this is not *adapted* to the test distribution, but instead only avoids tailoring to the *training* distribution. This is as true for the unseen test set distribution as it is for the hard-seen distribution which is constructed (from a carefully biased subsampling of the training distribution) to be different from the training distribution itself. On the other hand, for methods that do not resolve training set correlations, best model selection (for a specified target) consists of tailoring the parameters to that same target set.

Thus, our observations constitute further evidence that the proposed technique succeeds in resolving training data correlations better than our baselines, and further demonstrate that target-agnostic parameter validation may be used successfully in conjunction with our method.

4 Optimization details

(See also Sec 3 of the paper for notation and formulation)

We now describe in detail the optimization of our objective, closely following the scheme suggested in [4] for

---

[1] For all other results in the experiments in the paper and in this document, we use "traditional" parameter selection with the validation set drawn from the test distribution.

the tree-guided group lasso. Our objective is (Eq 4 from the paper):

$$L(\mathbf{W}|\mathbf{X}, \mathbf{Y}) + \lambda \sum_d \sum_g \|\mathbf{w}_d^{S_g}\|_2 \qquad (1)$$

where L is the logistic regression objective $\sum_m \sum_n \log(1 + \exp(-y_n^a(\mathbf{x}_n^T \cdot \mathbf{w}^m)))$.

Specifically, the non-smooth mixed-norm regularizer poses a non-trivial optimization problem. In the context of the group lasso, Bach *et. al.* [2] suggest squaring a similar mixed-norm regularizer before optimization. Since the regularizer is positive, squaring represents a smooth monotonic mapping. As such it preserves the same path of solutions during optimization while making the objective easier to optimize:

$$L(\mathbf{W}|\mathbf{X}, \mathbf{Y}) + \lambda \left(\sum_d \sum_g \|\mathbf{w}_d^{S_g}\|_2 \right)^2. \qquad (2)$$

This objective can be approximated using the following upper bound on the squared $\ell_{2,1}$ norm, due to Argyriou *et. al.* [1]:

$$\left(\sum_d \sum_g \|\mathbf{w}_d^{S_g}\|_2 \right)^2 \leq \sum_d \sum_g \frac{\left(\|\mathbf{w}_d^{S_g}\|_2 \right)^2}{\delta_{d,g}}, \qquad (3)$$

where the dummy variables $\delta_{d,g}$ are positive and satisfy $\sum_{d,g} \delta_{d,g} = 1$. Equality holds when $\delta_{d,g}$ are set to:

$$\delta_{d,g} = \|\mathbf{w}_d^{S_g}\|_2 / \sum_{d,g} \|\mathbf{w}_d^{S_g}\|_2. \qquad (4)$$

Our objective function now becomes:

$$L(\mathbf{W}|\mathbf{X}, \mathbf{Y}) + \sum_d \sum_g \frac{\left(\|\mathbf{w}_d^{S_g}\|_2 \right)^2}{\delta_{d,g}}. \qquad (5)$$

Similar to [4], this final objective is then minimized by alternatively optimizing with respect to (1) the dummy variables $\delta_{d,g}$ (update equation is simply the equality condition in Eq 3 for the bound in Eq 2) and (2) the weight vector $\mathbf{w}$ (updated by computing the appropriate gradient descent derivative).

References

[1] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-Task Feature Learning. In *NIPS*, 2007.

[2] F. Bach. Consistency of the group lasso and multiple kernel learning. In *JMLR*, 2008.

[3] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing Objects by Their Attributes. In *CVPR*, 2009.

[4] S. Kim and E. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML*, 2010.

[5] C. Lampert. Semantic Attributes for Object Categorization (slides). http://ist.ac.at/ chl/talks/lampert-vrml2011b.pdf, 2011.

[6] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. 2011.