Learning the right thing with visual attributes

Kristen Grauman
Department of Computer Science
University of Texas at Austin

With Chao-Yeh Chen, Aron Yu, and Dinesh Jayaraman
Beyond image labels

What does it mean to understand an image?

A lone cow grazes in a bright green pasture near an old tree, probably in the Scottish Highlands.

Labels

Cow
Tree
Grass

vs.

The story of an image
Attributes

- Mid-level semantic properties shared by objects
- Human-understandable and machine-detectable

Using attributes: Visual search

Suspect #1: Male, sunglasses, black and white hat, blue shirt

“Like this...but more ornate”

Person search
[Kumar et al. 2008, Feris et al. 2013]

Relative feedback
[Kovashka et al. 2012]
Using attributes: Interactive recognition

American Goldfinch?

[Cone-shaped beak? yes]

American Goldfinch?

[Branson et al. 2010, 2013]
Using attributes: Semantic supervision

Band-tailed pigeons:
- White collar
- Yellow feet
- Yellow bill
- Red breast

Zero-shot learning
[Lampert et al. 2009]

Mules:
- Shorter legs than donkeys
- Shorter tails than horses

Training with relative descriptions
[Parikh & Grauman 2011, Shrivastava & Gupta 2012]

Annotator rationales
[Donahue & Grauaman 2011]
Problem

With attributes, it’s easy to learn the wrong thing.

• Incidental correlations
• Spatially overlapping properties
• Subtle visual differences
• Partially category-dependent
• Variance in human-perceived definitions

…yet applications demand that correct meaning be captured!
Goal

Learn the right thing.

• How to decorrelate attributes that often occur simultaneously?

• Are attributes really class-independent?

• How to detect fine-grained attribute differences?
The curse of correlation

What will be learned from this training set?

Object Learning

Cat
The curse of correlation

What will be learned from this training set?

Attribute Learning

Forest animal? ✓ Brown? ✓ Has ears? ✓ Combinations? ✗

**Problem:** Attributes that often co-occur cannot be distinguished by the learner
The curse of correlation

**Problem:** Attributes that often co-occur cannot be distinguished by the learner

- **Forest animal**
  - Lion
  - Koala
  - Bear
  - Dolphin
  - Bird

- **Brown**
  - Lion
  - Koala
  - Bear
  - Dolphin
  - Bird
Idea: Resist the urge to share

Problem: Attributes that often co-occur cannot be distinguished by the learner
Semantic attribute groups

- Closely related attributes may share features.
- Assume attribute “groups” from external knowledge.
Standard approach: learning separately

Loss function: \[
L = \sum_{m} \sum_{(x,y^{(m)}) \in \text{samples}} \log \left( \left( 1 + e^{-y^{(m)}(\mathbf{x}^T w^{(m)})} \right) \right)
\]

\( x \): feature vector

\( w_1 \), \( w_2 \), ..., \( w_m \): learned weights

\( y^{(m)} \): label \((\pm 1)\)

\( m \): attribute index

JAYARAMAN ET AL., CVPR 2014
Proposed group-based formulation

$$\arg\min_{W} L(W|X,Y) + \sum_{d} \sum_{l} \| w_{d}^{S_{l}} \|_{2}$$

$S_{1}$ motion $S_{2}$ color $S_{3}$ texture

Group-wise weight matrix

Compute row $L_{2}$ norms

Penalize row $L_{1}$ norms

(by-group sharing)

(feature competition)
Formulation effect

Sparse features (no relationships among attributes)

Ours (inter-group competition, in-group sharing)

Standard multi-task learning (sharing and conflation across groups)

\[ \sum_{i} \sum_{j} |w_{i}^{j}| \]

\[ \sum_{d} \sum_{l} \| w_{d}^{S_{l}} \|_{2} \]

\[ \sum_{d} \| w_{d} \|_{2} \]
Results – Attribute detection

By decorrelating attributes, our attribute detectors generalize much better to novel unseen categories.

(*) Argyriou et al, Multi-task Feature Learning, NIPS 2007
(~) Farhadi et al, Describing Objects by Their Attributes, CVPR 2009
Attribute detection example

**Success cases**

- Not brown underparts
- No eye
- Not boxy
- No mouth
- No ear

**Failure cases**

- No feather
- Not furry
- Eyeline
- Black breast
- Not vegetation
Attribute localization examples

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Standard</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue back</td>
<td><img src="#" alt="Image" /></td>
<td><img src="#" alt="Image" /></td>
</tr>
<tr>
<td>Brown wing</td>
<td><img src="#" alt="Image" /></td>
<td><img src="#" alt="Image" /></td>
</tr>
<tr>
<td>Olive back</td>
<td><img src="#" alt="Image" /></td>
<td><img src="#" alt="Image" /></td>
</tr>
<tr>
<td>Crested head</td>
<td><img src="#" alt="Image" /></td>
<td><img src="#" alt="Image" /></td>
</tr>
</tbody>
</table>

Our method avoids conflation to learn the correct semantic attribute.
Goal

Learn the right thing.

- How to decorrelate attributes that often occur simultaneously?
- Are attributes really class-independent?
- How to detect fine-grained attribute differences?
Problem

Are attributes really category-independent?

Fluffy dog  ?  Fluffy towel
An intuitive but impractical solution

- Learn category-specific attributes?

Impractical! Would need examples for all category-attribute combinations…
Idea: Analogous attributes

- Given sparse set of category-specific models, infer “missing” analogous attribute classifiers.

1. Learned category-sensitive attributes

   - Spotted
   - Striped
   - Brown

2. Inferred attribute

   ?? = ???

3. Prediction

   A striped dog? Yes.

Chen & Grauman, CVPR 2014
Transfer via tensor completion

Construct sparse object-attribute classifier tensor

Discover low-d latent factors and infer missing classifiers (the analogous attributes)

Bayesian probabilistic tensor factorization [Xiong et al., SDM 2010].

\[ W \approx \sum_{k=1}^{K} O^k \circ A^k \circ C^k \]
Datasets

- **ImageNet attributes**
  - 9600 images
  - 384 object categories
  - 25 attributes
  - 1498 object-attribute pairs available

- **SUN attributes**
  - 14340 images
  - 280 object categories
  - 59 attributes
  - 61118 object-attribute pairs available

[Russakovsky & Fei-Fei 2010]
[Patterson & Hays 2012]
Inferring class-sensitive attributes

Our approach infers all 18K "missing" classifiers → savings of 348K labeled images

Category-sensitive outperforms status quo 76% of the time, average gain of 15 points in AP

Chen & Grauman, CVPR 2014
Which attributes are analogous?

Chen & Grauman, CVPR 2014
Goal

Learn the right thing.

• How to decorrelate attributes that often occur simultaneously?

• Are attributes really class-independent?

• How to detect fine-grained attribute differences?
Problem: Fine-grained attribute comparisons

Which is more comfortable?
Relative attributes

Use ordered image pairs to train a ranking function:

\[ O_m = \{ (\text{image}_i, \text{image}_j) | \text{“smiling more than”} \} \]

\[ w_m^T x_i > w_m^T x_j \quad \forall (i, j) \in O_m \]

[Parikh & Grauman, ICCV 2011; Joachims 2002]
Relative attributes
Rather than simply label images with their properties,
We can compare images by attribute’s “strength”
**Idea:** Local learning for fine-grained relative attributes

- Lazy learning: train query-specific model on the fly.
- Local: use only pairs that are similar/relevant to test case.

Yu & Grauman, CVPR 2014

Test comparison

Relevant nearby training pairs

Yu & Grauman, CVPR 2014
**Idea**: Local learning for fine-grained relative attributes.

Yu & Grauman, CVPR 2014
UT Zappos50K Dataset

Large shoe dataset, consisting of 50,025 catalog images from Zappos.com

- 4 relative attributes
- High quality pairwise labels from mTurk workers
- 6,751 ordered labels + 4,612 “equal” labels
- 4,334 twice-labeled fine-grained labels (no “equal” option)

Yu & Grauman, CVPR 2014
Results: Fine-grained attributes

Accuracy of comparisons – all attributes

<table>
<thead>
<tr>
<th></th>
<th>Zap50K-1</th>
<th>Zap50K-2</th>
<th>OSR</th>
<th>PubFig</th>
</tr>
</thead>
<tbody>
<tr>
<td>RelTree [2]</td>
<td>–</td>
<td>–</td>
<td>90.41</td>
<td>83.37</td>
</tr>
<tr>
<td>Global [3]</td>
<td>89.57</td>
<td>61.62</td>
<td>88.80</td>
<td>80.56</td>
</tr>
<tr>
<td>RandPair</td>
<td>84.34</td>
<td>57.98</td>
<td>86.93</td>
<td>72.46</td>
</tr>
<tr>
<td>FG-LocalPair</td>
<td>91.64</td>
<td>66.43</td>
<td>92.37</td>
<td>89.72</td>
</tr>
</tbody>
</table>

Accuracy on the 30 hardest test pairs

Yu & Grauman, CVPR 2014
Predicting useful neighborhoods

- Most relevant points = most similar points?
- Pose as large-scale multi-label classification problem

\[
\mathbf{y}_n = [1, 0, 1, 1, 0, ..., 0, 1] \\
\mathbf{\hat{y}}_q = [0, 0, 0, 1, 1, ..., 1, 0] \\
\]

\[
\mathbf{x}_n \xrightarrow{f} \mathbf{z}_n \xrightarrow{\phi} \mathbf{y}_n \\
\mathbf{\hat{y}}_q \xrightarrow{\phi} \mathbf{z}_n \xrightarrow{f} \mathbf{x}_q \\
\]

[Reconstruct] [Yu & Grauman NIPS 2014]
Predicting useful neighborhoods

SUN Attribute Dataset: 14,340 images, 707 classes

“hiking”

“eating”

“exercise”

“clouds”

Yu & Grauman, NIPS 2014
Summary

- Attribute learning is more nuanced than object learning
- Essential that language and visual concepts align

- Ideas:
  - Explicitly decorrelate attribute classifiers
  - Transfer between analogous attribute-object models
  - Fine-grained comparisons via lazy local learning