The language of visual attributes

Kristen Grauman
Facebook AI Research
University of Texas at Austin
Attributes vs. objects

Physical entity

- Red
- Round
- Ripe
- Fresh

Visual properties
Value of attributes

Visual search

Zero-shot learning

Image/video description

Interactive recognition

“Find a more formal shoe”

Zebras have stripes and four legs...

A lone cow grazes in a green pasture.

What color is the beak?

The language of visual attributes

• **Attributes as operators**
  Attributes: adjectives that *modify* objects: nouns

• **Attributes for comparisons**
  Relative differences that people first describe

• **Attributes for visual styles**
  Semantic topic models for data-driven styles
Attributes and objects are fundamentally different.
Attribute and Object Representations

Yet status quo treats attributes and objects the same...

As latent vector encodings

e.g., Wang CVPR16, Liu CVPR15, Singh ECCV16, Lu CVPR17, Su ECCV16, …
Attribute vs. Object Representations

**object**
- Prototypical “car” instance

**attribute**
- Prototypical “sliced” instance

?
Challenges for the status quo approach

Object-agnostic attribute representation

Has to capture interactions with every object
Challenges for the status quo approach

Object-agnostic attribute representation

Old car vs. Old man

Has to capture attributes’ distinct manifestations
Our idea – Attributes as operators

Attributes are *operators* that transform object encodings

[Nagarajan & Grauman, ECCV 2018]
Our idea – Attributes as operators

Objects are vectors
Attributes are operators

Composition is:
an attribute operator
transforming an object vector

[Nagarajan & Grauman, ECCV 2018]
Linguistically inspired regularizers

Antonym-consistency:

“Unripe should *undo* the effect of ripe”

[Nagarajan & Grauman, ECCV 2018]
Linguistically inspired regularizers

Attribute commutation:
Attribute effects should stack.

[Nagarajan & Grauman, ECCV 2018]
Learning attribute operators

[Nagarajan & Grauman, ECCV 2018]
Learning attribute operators

Triplet loss to learn embedding space

[Nagarajan & Grauman, ECCV 2018]
Learning attribute operators

Triplet loss [plus linguistic regularizers] to learn embedding space

Initialize with GloVe word embeddings [Pennington et al. EMNLP 2014]
Learning attribute operators

Allows unseen compositions

Input image: $x$

$f(x)$

$g(p)$

True pair: $p$ (sliced banana)

$g(p')$

Negative pair: $p'$

Unseen objects

new composition (diced grape)

[Unseen objects]

[Attribute operators]

[bannana, pear, orange, apple, tomato, fig]

[Nagarajan & Grauman, ECCV 2018]
Evaluation

UT-Zappos 50k
(Yu & Grauman, CVPR 14)

16 attributes x 12 objects

MIT States
(Isola et al., CVPR 15)

115 attributes x 245 objects
Evaluating our composition model

Train time
- Sliced carrot
- Unripe orange
- Diced onion
- Sliced apple

Test time
- Sliced orange
- Diced carrot
Evaluating our composition model

**Train time**
- Diced onion
- Sliced carrot
- Sliced apple

**Test time**
- Unripe orange
- Sliced orange
- Diced carrot

Combination never seen during training
Results – Attribute+object composition recognition

MIT States: 6% increase in open world (3% h-mean)

UT-Zap: 14% increase in open world (12% h-mean)

<table>
<thead>
<tr>
<th>Method</th>
<th>MIT-States</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>closed</td>
</tr>
<tr>
<td><strong>Chance</strong></td>
<td>0.1</td>
</tr>
<tr>
<td>VisProd (SVM)</td>
<td>11.1</td>
</tr>
<tr>
<td>VisProd (NN)</td>
<td>13.9</td>
</tr>
<tr>
<td>AnalogousAttr#</td>
<td>1.4</td>
</tr>
<tr>
<td>RedWine*</td>
<td>12.5</td>
</tr>
<tr>
<td>LabelEmbed</td>
<td>13.4</td>
</tr>
<tr>
<td>LabelEmbed+</td>
<td>14.8</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td>12.0</td>
</tr>
</tbody>
</table>

*Misra et al. CVPR 2017
#Chen & Grauman CVPR 2014

[Nagarajan & Grauman, ECCV 2018]
Results - Retrieving unseen (unseen) compositions

Rusty Lock

query Nearest Images in ImageNet
The language of visual attributes

- **Attributes as operators**
  Attributes: adjectives that *modify* objects: nouns

- **Attributes for comparisons**
  Relative differences that people first describe

- **Attributes for visual styles**
  Semantic topic models for data-driven styles
Relative attributes

Parikh & Grauman, ICCV 2011
Singh & Lee, ECCV 2016
Relative attributes

Learn a ranking function per attribute

Parikh & Grauman, ICCV 2011
Singh & Lee, ECCV 2016
Relative attributes

Compare images by an attribute’s “strength”

- **bright**
  - ![Images of shoes]

- **smiling**
  - ![Images of people]

- **natural**
  - ![Images of landscapes]

[Parikh & Grauman, ICCV 2011]
Challenge #1: fine-grained comparisons

Which is more sporty?

Sparsity of supervision problem:
1. Label availability: lots of possible pairs.
2. Image availability: subtleties hard to curate.
Idea: Semantic jitter

Overcome sparsity of available fine-grained image pairs with attribute-conditioned image generation

Our idea: Semantic jitter vs. Status quo: Low-level jitter

Yu & Grauman, ICCV 2017
Semantic jitter for attribute learning

Train rankers with both real and synthetic image pairs, test on real fine-grained pairs.

Ranking functions trained with deep spatial transformer ranking networks [Singh & Lee 2016] or Local RankSVM [Yu & Grauman 2014]
Challenge #2: Which attributes matter?

Left shoe is _____ than right shoe:

- Less colorful
- Less comfortable
- More rugged
- More shiny
- Less feminine
- More stylish
- More formal
Idea: Prominent relative attributes

Infer which comparisons are perceptually salient

Left shoe is _____ than right shoe:

Less colorful
Less comfortable
More rugged
More shiny
Less feminine
More stylish
More formal

Prominent Differences

More formal
More shiny
Less comfortable
Less feminine
Less colorful
More rugged
More stylish

Chen & Grauman, CVPR 2018
Approach: What causes prominence?

- Large difference in attribute strength:

- Unusual and uncommon attribute occurrences:

- Absence of other noticeable differences:

**In general:** Interactions between all the relative attributes in an image pair cause prominent differences.

*Chen & Grauman, CVPR 2018*
Approach: Predicting prominent differences

input: $y_{uv} = (x_u, x_v)$

Relative Attribute Rankers

$\phi(y_{uv})$

Prominence Multiclass Classifier

Prominent Difference: Visible Teeth $A_{uv}$

Chen & Grauman, CVPR 2018
Results: Prominent differences

(a) colorful (>)
sporty, comfortable

(b) sporty (>)
colorful, comfortable

(c) tall (<)
colorful, sporty

(d) shiny (>)
feminine, colorful

(e) rugged (<)
tall, feminine

(f) feminine (>)
comfortable, shiny

(j) masculine (>)
smiling, visible teeth

(k) bald head (<)
dark hair, visible teeth

(l) dark hair (<)
mouth open, smiling

(Top 3 prominent differences for each pair)
Results: Prominent differences

**Rank-SVM**

Accuracy vs. # Top prominent as ground truth

**Rank-CNN**

Accuracy vs. # Top prominent as ground truth

- Ours (Relative SVM Features)
- Widest Relative Difference (SVM, Parikh and Grauman)
- Binary Dominance (Turakhia and Parikh)
- Single Image Prominence
- Prior

- Ours (Relative CNN Features)
- Widest Relative Difference (CNN, Singh and Lee)
- Binary Dominance (Turakhia and Parikh)
- Single Image Prominence
- Prior
Prominent differences: impact on visual search

Query: “white high-heeled shoes”

Feedback: “shinier than these”

Feedback: “less formal than these”

Leverage prominence to better focus search results

Chen & Grauman, CVPR 2018
Prominent differences: impact on visual search

Faster retrieval of user’s target image without using any additional user feedback.

Leverage prominence to better focus search results

Chen & Grauman, CVPR 2018
From items to styles
The language of visual attributes

• **Attributes as operators**
  Attributes: adjectives that *modify* objects: nouns

• **Attributes for comparisons**
  Relative differences that people first describe

• **Attributes for visual styles**
  Semantic topic models for data-driven styles
How to represent visual style?

Challenges:
- Same “look” manifests in different garments
- Emerges organically and evolves over time
- Soft boundaries
Idea: Discovering visual styles

Unsupervised learning of a style-coherent embedding with a polylingual topic model

An outfit is a mixture of (latent) styles.  
A style is a distribution over attributes.


Hsiao & Grauman, ICCV 2017
Example discovered styles (dresses)

Styles we automatically discover in the Amazon dataset [McAuley et al. 2015]
Example discovered styles (full outfit)

Styles automatically discovered in the **HipsterWars** dataset [Kiapour et al]
Bohemian Hipster

Our embedding naturally facilitates browsing for mixes of user-selected styles

Mixing styles

Hsiao & Grauman, ICCV 2017
Creating a “capsule” wardrobe

**Goal:** Select minimal set of pieces that mix and match well to create many viable outfits

Pose as *subset selection* problem

set of garments $= \arg \max$ **compatibility** + **versatility**

*Hsiao & Grauman, CVPR 2018*
Creating a “capsule” wardrobe
Discover user’s style preferences from album

Hsiao & Grauman, CVPR 2018
Visual trend forecasting
We predict the future popularity of each style

Amazon dataset [McAuley et al. SIGIR 2015]
Visual trend forecasting

What kind of fabric, texture, color will be popular next year?
VizWiz: Answer blind people’s visual questions

[Gurari et al. CVPR 2018] Spotlight/Poster Wednesday

- Goal-oriented visual questions
- Conversational language
- Assistive technology
Summary: the language of visual attributes

New ideas for attributes as operators, comparisons, style basis
Applications for visual search and fashion image analysis
Papers/code

• **Attributes as Operators**. T. Nagarajan and K. Grauman. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, Sept 2018. [pdf] [supp] [code]


• **Compare and Contrast: Learning Prominent Visual Differences**. S. Chen and K. Grauman. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 2018. [pdf] [supp] [project page]


