Attributes

Tues May 2
Kristen Grauman
UT Austin

A5 end game

• Deadline extended to Friday EXCEPT for extra credit
• Two leaderboards will be posted
  • Tuesday, Friday
  • Extra credit for top 5 performing submissions

Final exam

• Tues May 16, 9-12 noon in GDC 1.304
• Comprehensive
• Closed book
• Two pages of notes allowed
Last time

- Neural networks / multi-layer perceptrons
  - View of neural networks as learning hierarchy of features
- Convolutional neural networks
  - Architecture of network accounts for image structure: local connections, shared weights.
  - “End-to-end” recognition from pixels
  - Together with big (labeled) data and lots of computation → major success on benchmarks, image classification and beyond

Recall Traditional Image Categorization

- Training: Images, Labels, Image Features, Classifier, Trained Classifier
- Testing: Test Image, Image Features, Trained Classifier, Prediction

Recall: Learning a Hierarchy of Feature Extractors

- Each layer of hierarchy extracts features from output of previous layer
- All the way from pixels → classifier
- Layers have the (nearly) same structure
- Train all layers jointly
Recall: Two-layer neural network

Pre-training a representation

Transfer Learning with CNNs

• Improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned.

• Weight initialization for CNN
Understanding and Visualizing CNN

- Find images that maximize some class scores
- Individual neuron activation
- Visualize input pattern using deconvnet

Recall: visualizing what was learned

- What do the learned filters look like?

Typical first layer filters

Individual Neuron Activation
Recall: Learning Feature Hierarchy

**Goal:** Learn **useful higher-level features** from images

Feature representation:
- 3rd layer: "Objects"
- 2nd layer: "Object parts"
- 1st layer: "Edges"
- Pixels

**Input data**

Lee et al., ICML 2009; CACM 2011

Slide: Rob Fergus
Map activation back to the input pixel space

- What input pattern originally caused a given activation in the feature maps?

Layer 1

Layer 2
Layer 3

Layer 4 and 5

Attributes
and learning to rank
and local learning
What are visual attributes?

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

- Material, Appearance, Function/affordance, Parts...
- Adjectives
- Statements about visual concepts

Examples: Binary Attributes

Facial properties

"Smiling Asian Men With Glasses"

Kumar et al. 2008

Examples: Binary Attributes

Object parts and shapes

Farhadi et al. 2009
Examples: Binary Attributes

Animal properties

Lampert et al. 2009

Examples: Binary Attributes

Animal properties

Welinder et al. 2010

Examples: Binary Attributes

Scene properties

Patterson and Hays 2011
Examples: Binary Attributes

Shopping descriptors

Berg et al. 2010

Why attributes?

• Why would a robot need to recognize a scene?

Can I walk around here? Is this walkable?

Why attributes?

• Why would a robot need to recognize an object?

How hard should I grip this? Is it brittle?
Why attributes?

- How do people naturally describe visual concepts?

I want elegant silver sandals with high heels.

Zebras have stripes.

Training attribute classifiers

<table>
<thead>
<tr>
<th>Labeled Images</th>
<th>Features</th>
<th>Classifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>Farhadi et al., CVPR 2009</td>
<td>Kovashka et al, CVPR 2012</td>
<td>Kumar et al, ECCV 2008</td>
</tr>
<tr>
<td>Kumar et al., ECCV 2008</td>
<td>Lampert et al, CVPR 2009</td>
<td>Yu et al, CVPR 2013</td>
</tr>
</tbody>
</table>

Slide credit: Dinesh Jayaraman

Mule
Attributes

A mule…
- Is furry
- Has four legs
- Has a tail

Binary attributes

A mule…
- Is furry
- Has four legs
- Has a tail


Zero-shot Learning

- Seen categories with labeled images
  - Train attribute predictors
- Unseen categories
  - No examples, only description

Relative attributes

A mule…

- Is furry
- Has four legs
- Has a tail

- Legs shorter than horses’
- Tail longer than donkeys’

---

Relative attributes

**Idea:** represent *visual comparisons* between classes, images, and their properties.

[Parikh & Grauman, ICCV 2011]

---

How to teach relative visual concepts?

*How much is the person smiling?*

[Slide credit: Kristen Grauman]
How to teach relative visual concepts?

*How much is the person smiling?*

Slide credit: Kristen Grauman
Learning relative attributes

For each attribute, use ordered image pairs to train a ranking function:

\[ O_{m} \subset \{ \begin{array}{c} \text{Im1} \rightarrow \ldots \rightarrow \text{Im}_m \end{array} \} \]

\[ w_{m}(x_{i}) > w_{m}(x_{j}) \quad \forall (i, j) \in O_{m} \]

[Parikh & Grauman, ICCV 2011; Joachims 2002]

Slide credit: Kristen Grauman

Learning relative attributes

Max-margin learning to rank formulation

\[
\begin{align*}
\min & \left( \frac{1}{2}||w_{m}||^{2} + C \left( \sum_{i} \xi_{i} + \sum_{j} \gamma_{j} \right) \right) \\
\text{s.t.} & \quad w_{m}^{T}(x_{i} - x_{j}) \geq 1 - \xi_{ij} \\
& \quad |w_{m}^{T}(x_{i} - x_{j})| \leq \gamma_{ij} \\
& \quad \xi_{ij} \geq 0; \gamma_{ij} \geq 0
\end{align*}
\]

Joachims, KDD 2002

Slide credit: Devi Parikh

Relating images

Rather than simply label images with their properties,

[Parikh & Grauman, ICCV 2011]

Slide credit: Kristen Grauman

Not bright

Smiling

Not natural
Relating images
Now we can compare images by attribute’s “strength”

- bright
- smiling
- natural

[Parikh & Grauman, ICCV 2011]

Slide credit: Kristen Grauman

Relative zero-shot learning
Predict new classes based on their relationships to existing classes – even without training images.

- Leg length: Horse > Mule
- Tail length: Mule > Donkey

Relative zero-shot learning

Comparative descriptions are more discriminative than categorical definitions.
Attributes for search and recognition

Attributes give human user way to
- Teach novel categories with description
- Communicate search queries
- Give feedback in interactive search
- Assist in interactive recognition

Image search

- Meta-data commonly used, but insufficient

Keyword query: “smiling asian men with glasses”

Why are attributes relevant to image search?

- Human understandable
- Support familiar keyword-based queries
- Composable for different specificities
- Efficiently divide space of images
Attributes are composable

Attributes efficiently divide the space of images

Search applications: finding people
Search applications: finding people

Search **surveillance** feeds for suspects

http://lacrimestoppers.com/wanteds.aspx

Adapted from: Rogerio Feris

Search images from **ad hoc cameras** using semantic descriptions

Adapted from: Rogerio Feris
Search applications: finding people

What actress looks like a young Hillary Clinton?

Face Search with Attributes


Attribute Classifier Accuracies

Binary facial attributes in Columbia Face Database
Typically 80%-90% accuracy
FaceTracer: Searching for faces with attributes

- **Offline:**
  - Apply attribute classifiers to database images
  - Map classifier outputs to probabilities

- **Online:**
  - Convey available attribute names to user
  - Given query attributes, rank database images by confidence (e.g., product of probabilities)

Google: "smiling asian men with glasses"

FaceTracer: "smiling asian men with glasses"
FaceTracer: “older men with mustaches”

Attribute-based person search in video

Example query: Boston bombing scenario

Example query: Boston bombing scenario
Problem with one-shot visual search

- Keywords (even attributes) can be insufficient to capture query in one shot.
- Complete “indicator vector” over attributes need not adequately capture envisioned target.

Interactive visual search

- Iteratively refine the set of retrieved images based on user feedback on results so far
- Potential to communicate more precisely the desired visual content

How is interactive search done today?

- Traditional binary feedback is imprecise
- Coarse communication between user and system

Idea: Search via comparisons

[Kovashka et al., CVPR 2012]

“Like this... but more ornate”

• Whittle away irrelevant images via comparative feedback on properties of results

WhittleSearch: Relative attribute feedback

[Kovashka et al., CVPR 2012]

Query: “white high-heeled shoes”

Feedback: “less formal than these”

Feedback: “shinier than these”

Initial top search results

Refined top search results

WhittleSearch: Relative attribute feedback

[Kovashka et al., CVPR 2012]

Feedback: “similar hair style”

Feedback: “broader nose”

Initial reference images

Refined top search results

Slide credit: Kristen Grauman
"I want something more formal than this."

"I want something less formal than this."

"I want something more shiny than this."

WhittleSearch: Relative attribute feedback

[Kovashka et al., CVPR 2012]

Datasets

Shoes:
14,618 shoe images;
10 attributes:
"pointy", "bright", "high-heeled", "feminine" etc.

OSR:
2,688 scene images;
6 attributes:
"natural", "perspective", "open-air", "close-depth" etc.

PubFig/LFW:
772-2000 face images;
11 attributes:
"masculine", "young", "smiling", "round-face", etc.
WhittleSearch results

![Graphs comparing binary relevance feedback vs. relative attribute feedback]

More rapidly converge on the envisioned visual content.

---

**Problem**: Fine-grained attribute comparisons

![Comparison of coarse vs. fine-grained attributes]

Which is *more comfortable*?

---

**Idea**: Local learning for fine-grained relative attributes

- **Lazy**: Train query-specific model on the fly.
- **Local**: Use only pairs that are similar/relevant to test case.
**Idea:** Local learning for fine-grained relative attributes

![Diagram showing the concept of global vs. local learning for fine-grained attributes.](yu_grauman_cvpr_2014)

Learning attribute-specific metrics

Determine neighbor pairs based on learned distance

- E.g., Information-Theoretic Metric Learning [Davis et al. ICML '07]

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)

![Images of shoes from Zappos50K dataset.](yu_grauman_cvpr_2014)
Results: Fine-grained attributes

Accuracy of comparisons – all attributes

<table>
<thead>
<tr>
<th>Model</th>
<th>Zap50k-1</th>
<th>Zap50k-2</th>
<th>OSR</th>
<th>Pool</th>
<th>Fig</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global [3]</td>
<td>89.57</td>
<td>64.62</td>
<td>88.80</td>
<td>80.56</td>
<td></td>
</tr>
<tr>
<td>RandPair</td>
<td>84.34</td>
<td>57.98</td>
<td>86.93</td>
<td>72.46</td>
<td></td>
</tr>
<tr>
<td>HGS+LocalPair</td>
<td>91.64</td>
<td>64.43</td>
<td>92.37</td>
<td>89.72</td>
<td></td>
</tr>
</tbody>
</table>

Global: Parikh & Grauman, ICCV 2011

Local model succeeds, global model fails:

More sporty than

Local model failure cases:

Just noticeable differences

At what point is the relative strength of an attribute in two images indistinguishable?
Just noticeable differences
Non-trivial: relative attribute space is non-uniform

Approach: Just noticeable differences
We propose Bayesian local learning strategy to predict whether images $x_m, x_n$ are distinguishable:

$$P(D = 1) = \frac{1}{K} |P'|$$

Results: just noticeable differences
Proposed model pinpoints those pairs that are not distinguishable
Visualizing learned JND for an attribute

t-SNE embedding for “pointy”

Example “indistinguishable” predictions

Results: just noticeable differences

Positive impact on WhittleSearch: user can now express “equal” constraints

Yu & Grauman, ICCV 2015
Attributes for search and recognition

Attributes give human user way to
- Teach novel categories with description
- Communicate search queries
- Give feedback in interactive search
- Assist in interactive recognition
Categories of Recognition

Basic-Level

Airplane? Chair? Bottle? ...

Humans
   Easy
Computers
   Some Success

Recognition With Humans in the Loop

Wah et al., Multi-class Recognition and Part Localization with Humans in the Loop, ICCV 2011

- Computers: reduce number of required questions
- Humans: drive up accuracy of vision algorithms

Example Questions: Localize

Wah et al., Multi-class Recognition and Part localization with Humans in the Loop, ICCV 2011
Example Questions: Name attributes

Input Image \( (x) \)

Max Expected Information Gain

Question 1:
Click on the belly

Max Expected Information Gain

Question 2:
Is the bill hooked?

\( p(c \mid x) \)

\( p(c \mid x, u_1) \)

\( p(c \mid x, u_1, u_2) \)

Basic Algorithm

CUB-200-2011 Dataset

11,877 images, 200 bird species

13 part locations

288 binary attributes
Results: Without Computer Vision

- Perfect Users, Field Guide Attributes: 100% accuracy in $8 \cdot \log_2(200)$ questions if users agree with field guides.
- Real Users, Probabilistic User Model: Tolerate ambiguous responses, user error.

Results: With Computer Vision

- Incorporating computer vision reduces average time to identify true species from 109 sec to 37 sec.
- Intelligently selecting questions reduces average time from 69 sec to 37 sec.
Coming up

• Last lecture on Thursday
• Course wrap-up
• Applications and frontiers of computer vision