Plan for today

- Questions?
- Sign-up for paper
- Research overview
Success in modern visual recognition research

- Image classification
- Semantic segmentation
- Pose recognition
- Object detection

... and many more
Ingredients for success today

1. Fast computation (GPUs)

2. Large labeled image dataset

3. Network architecture

Which ingredient will be the bottleneck for tomorrow’s success?
Ingredients for success today

1. Fast computation (GPUs)

2. Large **labeled** image dataset

   Requires expensive, **detailed** human supervision

3. Network architecture
Take semantic segmentation as an example...

70,000+ annotation hours for 328K images but only 80 object categories (MS COCO)

Requires pixel-level semantic labels
Ingredients for success today

1. Fast computation (GPUs)

2. Large *weakly-labeled* image dataset

3. Network architecture
Fully-supervised vs. Weakly-supervised

**Fully-supervised:**
*Pixel-level labels*

**Weakly-supervised:**
*Image-level labels*

{Person, umbrella, hat, ...}
Overview: Weakly-supervised visual recognition

1. Learning object detectors with tags

2. Learning visual attributes with pairwise comparisons

3. Learning visual grounding with captions
Weakly-supervised object detection

Weakly-supervised object detection

Annotators

Mine discriminative patches

Weakly-labeled training images

Weakly-supervised object detection

Annotators

Mine discriminative patches

Weakly-labeled training images

Weakly-supervised object detection

- Supervision is provided at the \textit{image-level} \rightarrow scalable!

Weakly-supervised object detection

• Supervision is provided at the image-level → scalable!
• Due to intra-class appearance variations, occlusion, clutter, mined regions correspond to object-part or include background

Our idea: Track and transfer

Weakly-labeled videos tagged with “car”

Weakly-labeled training images tagged with “car”

Transferring tracked object boundaries

Mined positive image region
Transferring tracked object boundaries

• For each video frame, search over scales and locations

[Unsupervised video object proposals, Xiao and Lee, CVPR 2016]
Transferring tracked object boundaries

- For each video frame, search over scales and locations

Mined positive image region
Transferring tracked object boundaries

- For each video frame, search over scales and locations

Mined positive image region
Transferring tracked object boundaries

- For each video frame, search over scales and locations

Mined positive image region
Transferring tracked object boundaries

For each video frame, search over scales and locations.
Transferring tracked object boundaries

• For each video frame, search over scales and locations
Discovered pseudo ground-truth boxes

• Handles atypical pose, partial occlusion, and clutter
Weakly-supervised object detection accuracy

<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aero</th>
<th>bird</th>
<th>boat</th>
<th>car</th>
<th>cat</th>
<th>cow</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>train</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Song et al., 2014 [33]</td>
<td>27.6</td>
<td>19.7</td>
<td>9.1</td>
<td>39.1</td>
<td>33.6</td>
<td>20.9</td>
<td>27.7</td>
<td>29.4</td>
<td>39.2</td>
<td>35.6</td>
<td>28.2</td>
</tr>
<tr>
<td>Song et al., 2014 [34]</td>
<td>36.3</td>
<td>23.3</td>
<td>12.3</td>
<td>46.6</td>
<td>25.4</td>
<td>23.5</td>
<td>23.5</td>
<td>27.9</td>
<td>40.9</td>
<td>37.7</td>
<td>29.7</td>
</tr>
<tr>
<td>Bilen et al., 2014 [1]</td>
<td>42.2</td>
<td>23.1</td>
<td>9.2</td>
<td>45.1</td>
<td>24.9</td>
<td>24.0</td>
<td>18.6</td>
<td>31.6</td>
<td>43.6</td>
<td>35.9</td>
<td>29.8</td>
</tr>
<tr>
<td>Wang et al., 2014 [37]</td>
<td>48.9</td>
<td>26.1</td>
<td>11.3</td>
<td>40.9</td>
<td>34.7</td>
<td><strong>34.7</strong></td>
<td>34.4</td>
<td>35.4</td>
<td><strong>52.7</strong></td>
<td>34.8</td>
<td>35.4</td>
</tr>
<tr>
<td>Cinbis et al., 2015 [4]</td>
<td>39.3</td>
<td>28.8</td>
<td><strong>20.4</strong></td>
<td>47.9</td>
<td>22.1</td>
<td>33.5</td>
<td>29.2</td>
<td>38.5</td>
<td>47.9</td>
<td>41.0</td>
<td>34.9</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>53.9</strong></td>
<td><strong>37.7</strong></td>
<td><strong>13.7</strong></td>
<td><strong>56.6</strong></td>
<td><strong>51.3</strong></td>
<td>24.0</td>
<td><strong>38.5</strong></td>
<td>47.9</td>
<td>47.0</td>
<td><strong>48.4</strong></td>
<td><strong>41.9</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>VOC 2010 test</th>
<th>aero</th>
<th>bird</th>
<th>boat</th>
<th>car</th>
<th>cat</th>
<th>cow</th>
<th>dog</th>
<th>horse</th>
<th>mbike</th>
<th>train</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cinbis et al., 2015 [4]</td>
<td>44.6</td>
<td>25.5</td>
<td><strong>14.1</strong></td>
<td>36.3</td>
<td>23.2</td>
<td>26.1</td>
<td>29.2</td>
<td>36.0</td>
<td><strong>54.3</strong></td>
<td>31.2</td>
<td>32.1</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td><strong>53.5</strong></td>
<td><strong>37.5</strong></td>
<td><strong>8.0</strong></td>
<td><strong>44.2</strong></td>
<td><strong>49.4</strong></td>
<td><strong>33.7</strong></td>
<td><strong>43.8</strong></td>
<td><strong>42.5</strong></td>
<td><strong>47.6</strong></td>
<td><strong>40.6</strong></td>
<td><strong>40.1</strong></td>
</tr>
</tbody>
</table>

- Significantly outperformed existing weakly-supervised detection methods (at the time)
Overview: Weakly-supervised visual recognition

1. Learning object detectors with tags

2. Learning visual attributes with pairwise comparisons

3. Learning visual grounding with captions
Goal of weakly-supervised attribute localization

Attribute: Smile

- Localizing the relevant region improves attribute modeling
- In the relative setting, we are provided with pairwise comparisons
Previous approaches

• Require pre-trained part detectors or crowdsourcing [Kumar et al. ICCV 2009, Bourdev et al. ICCV 2011, Duan et al. CVPR 2012, Zhang et al. CVPR 2014]

• “Pipeline” where features, localizer, and classifier are trained separately and sequentially; suboptimal and slow [Xiao and Lee, ICCV 2015]

• **Our idea:** jointly learn features, localizer, and classifier *end-to-end* in a **deep learning** framework

• We focus on the *relative attribute* setting
Overview of our end-to-end approach

- **Goal:** Given pairs of ordered training images, simultaneously localize attribute in each image and learn a ranker

Attribute: Smile

I₁ → Siamese Network (S₁) → V₁

I₂ → Siamese Network (S₂) → V₂

Loss Function

[Singh and Lee, “End-to-End Localization and Ranking for Relative Attributes”, ECCV 2016]
Our *end-to-end* approach

- Ranker network takes the localized region to produce a ranking score
- Combine the global image for global context
Progression of localized region over training epochs

* Heatmap: distribution of localized region across entire training dataset
Testing

- Localize the relevant attribute region
- Produce a ranking score for the test image
Results: Discovered regions and ranking on LFW-10 Faces

- Our network discovers relevant attribute regions
- Leads to accurate rankings
Results: Discovered regions and ranking UT-Zap50K Shoes
Results: Image pair ranking accuracy

- % of test image pairs whose predicted relative attribute ranking is correct
- State-of-the-art results on LFW-10, UT-Zap50K, OSR, Shoe-with-Attribute

<table>
<thead>
<tr>
<th>Method</th>
<th>BH</th>
<th>DH</th>
<th>EO</th>
<th>GL</th>
<th>ML</th>
<th>MO</th>
<th>S</th>
<th>VT</th>
<th>VF</th>
<th>Y</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parikh &amp; Grauman [3] + CNN</td>
<td>78.10</td>
<td>83.09</td>
<td>71.43</td>
<td>68.73</td>
<td>95.40</td>
<td>65.77</td>
<td>63.84</td>
<td>66.46</td>
<td>81.25</td>
<td>72.07</td>
<td>74.61</td>
</tr>
<tr>
<td>Sandeep et al. [13]</td>
<td>82.04</td>
<td>80.56</td>
<td>83.52</td>
<td>68.98</td>
<td>90.94</td>
<td>82.04</td>
<td>85.01</td>
<td>82.63</td>
<td>83.52</td>
<td>71.36</td>
<td>81.06</td>
</tr>
<tr>
<td>Xiao &amp; Lee [14]</td>
<td>83.21</td>
<td>88.13</td>
<td>82.71</td>
<td>72.76</td>
<td>93.68</td>
<td>88.26</td>
<td>86.16</td>
<td>86.46</td>
<td>90.23</td>
<td>75.05</td>
<td>84.66</td>
</tr>
<tr>
<td>Ours</td>
<td>83.94</td>
<td>92.58</td>
<td>90.23</td>
<td>71.21</td>
<td>96.55</td>
<td>91.28</td>
<td>84.75</td>
<td>89.85</td>
<td>87.89</td>
<td>80.81</td>
<td>86.91</td>
</tr>
</tbody>
</table>

- % of test image pairs whose predicted relative attribute ranking is correct
- State-of-the-art results on LFW-10, UT-Zap50K, OSR, Shoe-with-Attribute
Overview: Weakly-supervised visual recognition

1. Learning object detectors with tags

2. Learning visual attributes with pairwise comparisons

3. Learning visual grounding with captions
Tag-based weakly supervised localization

Localizing objects

image + tag

Bounding box detections

Deselaers et al. 2010, Pandey & Lazebnik 2011, Song et al. 2014, Singh et al. 2016, ...
Tag-based weakly supervised localization

Semantic segmentation

Figure 1. Our approach takes labels in the form of which classes are present in the scene during training, and learns a segmentation model, even though no annotations at the pixel-wise are available.

Xu et al. 2014, Pathak et al. 2015, Pinheiro & Collobert 2015, ...
Tags vs. natural language
Tags vs. natural language

Tags: "cat", "hand", "donut"
Tags vs. natural language

Tags: "cat", "hand", "donut"

Tags convey very limited amount of information about images
Tags vs. natural language

Caption: "A grey cat staring at a hand with a donut."
Tags vs. natural language

Caption: "A grey cat staring at a hand with a donut."

Much more informative and natural to describe images with caption
Our goal: Leverage structure in natural language for weakly supervised visual grounding

A grey cat staring at a hand with a donut.

Our goal: Leverage structure in natural language for weakly supervised visual grounding

A grey cat staring at a hand with a donut.

Our goal: Leverage structure in natural language for weakly supervised visual grounding

A grey cat staring at a hand with a donut.

Visual grounding of free-from language with only image-level supervision

Our goal: Leverage structure in natural language for weakly supervised visual grounding

A grey cat staring at a hand with a donut.

• Exploit *linguistic structure* in caption
• Parse caption into syntactic parse tree [Socher et al. 2013]
Key idea: Transfer linguistic structure to visual domain

A grey cat staring at a hand with a donut.

(1) Implies sibling exclusivity:
Image regions for sibling nodes should be exclusive
**Key idea:** Transfer linguistic structure to visual domain

A grey cat staring at a hand with a donut.

(2) Implies *parent-child inclusivity*:
Image region of parent node is *union* of children's regions
Four submodules: (1) visual encoder, (2) language encoder, (3) semantic embedding module, and (4) loss functions.
Phrase grounding on Visual Genome

*Left:* our model

*Right:* baseline trained without structural constraints (Discriminative loss only)

- the train is on the bridge
- a person driving a boat
- a jockey riding a horse
Multiple queries on Visual Genome

clock tower is tall
buildings by street
snow covered park
bench is black and white
the walls are dark purple
yellow and orange cat

Our groundings generated for different phrases of the same image
Multiple queries on MS COCO

Our groundings generated for different tags of the same image
## Phrase grounding on Visual Genome

<table>
<thead>
<tr>
<th></th>
<th>Random</th>
<th>Disc-only</th>
<th>Token</th>
<th>PC</th>
<th>SIB</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>0.115</td>
<td>0.230</td>
<td>0.222</td>
<td>0.236</td>
<td>0.231</td>
<td><strong>0.244</strong></td>
</tr>
</tbody>
</table>

Table 1. Localization accuracy as measured by the “pointing game” [52] on Visual Genome. Our model outperforms all baselines, including variants of our method that lack one or more loss terms.
Animal keypoint detection

Input

Output
Motivation

- Veterinary research has found facial expressions of pain for horses, sheep, and cows to be a good indicator of pain.

Motivation

• Automatic pain detection can have a huge impact on animal welfare
  – $0.5 billion spent annually in the US to treat lameness in horses

• Human detection of animal pain is challenging
  – Humans need to be trained to detect pain
  – Livestock may hide expressions of pain near humans
Key idea

• Leverage existing *human* training data for animal facial keypoint detection (with a small amount of animal training data)
• Warp animal faces to resemble human shape

Maheen Rashid, Xiuye Gu, Yong Jae Lee. Interspecies Knowledge Transfer. CVPR 2017.
Approach

- Warping network warps animal to have human-like shape
- Keypoint network predicts facial keypoints on warped animal face
Experiments

• Horse Facial Keypoint Dataset
  – 3531 training, 186 testing images
  – Left eye, right eye, nose center, left mouth corner, right mouth corner

• Sheep Facial Keypoint Dataset [Yang et al. 2016]
  – 432 training, 99 testing images
  – Same 5 keypoints as above
Qualitative results

<table>
<thead>
<tr>
<th>Ground truth</th>
<th>Our warp</th>
<th>Our pred</th>
<th>No warp</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13.png" alt="Image" /></td>
<td><img src="image14.png" alt="Image" /></td>
<td><img src="image15.png" alt="Image" /></td>
<td><img src="image16.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image17.png" alt="Image" /></td>
<td><img src="image18.png" alt="Image" /></td>
<td><img src="image19.png" alt="Image" /></td>
<td><img src="image20.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Qualitative results

- By leveraging existing human training data, our CNN model performs much better.
Quantitative results

- Triplet Interpolated Features [Yang et al. 2016]
Conclusions

• Weakly-supervised learning enables *scalable* visual recognition
  - Object detection, attribute recognition, visual grounding

• Still, challenges remain in dealing with
  - Noise in data
  - Limited training data
Coming up

• Sign-up for paper
• Next class: CNN/PyTorch/Torch tutorial