Weakly Supervised Object Recognition with Convolutional Neural Networks

Ivan Laptev

ivan.laptev@inria.fr

WILLOW, INRIA/ENS/CNRS, Paris

Joint work with: Maxime Oquab – Leon Bottou – Josef Sivic
Summary

Training input

Test output

image-level labels:

✔ Person
✔ Chair
✘ Airplane

✔ Reading
✘ Riding bike
✘ Running

More details in http://www.di.ens.fr/willow/research/weakcnn/
Recent Progress: Convolutional Neural Networks

- Success in character recognition [LeCun’88].
- Limited performance on natural images until 2012.

**ILSVRC’12**: 1.2M images, 1K classes

<table>
<thead>
<tr>
<th>Year</th>
<th>Method</th>
<th>Top 5 error</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>SIFT + FVs [7]</td>
<td>26.2%</td>
</tr>
<tr>
<td></td>
<td>1 CNN</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>5 CNNs</td>
<td>16.4%</td>
</tr>
<tr>
<td></td>
<td>1 CNN*</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>7 CNNs*</td>
<td>15.3%</td>
</tr>
</tbody>
</table>

**2014-2015**

- GoogLeNet: 6.6%
- VGG: 6.8%
- BAIDU: 5.3%
- Human: 5.1%
Let’s look at the data

Images of chairs and tables in ImageNet

A typical image with chairs and tables on Flickr.com
How to use CNNs for cluttered scenes?

Use ImageNet pre-trained CNN \( \rightarrow \) Post-train on the new task

[Girshick et al.’14], [Oquab et al.’14], [Sermanet et al.’13], [Donahue et al. ’13], [Zeiler & Fergus ’13] ...
Approach – sliding window training / testing

Annotated input image:
Multi-scale overlapping tiling

- Person : too small
- Sheep : too small
- Person : truncated
- Sheep : too small
- Person : truncated
- Sheep : truncated
- Person : truncated
- Sheep : truncated
- Person : truncated
- Sheep : truncated

2 labels: delete
Approach

1. Design training/test procedure using sliding windows
2. Train adaptation layers to map labels
Results

Pascal VOC

Oquab, Bottou, Laptev and Sivic
CVPR 2014
Results

(a) Representative true positives

(aeroplane)  (aeroplane)

(b) Top ranking false positives

(aeroplane)  (aeroplane)

[Oquab, Bottou, Laptev and Sivic, CVPR 2014]
Results

[Oquab, Bottou, Laptev and Sivic, CVPR 2014]
How to use CNNs for cluttered scenes?

**Problem:** Annotation of bounding boxes is (a): subjective (b): expensive
Motivation: labeling bounding boxes is tedious
Are bounding boxes needed for training CNNs?

Image-level labels: Bicycle, Person
Motivation: image-level labels are plentiful

“Beautiful red leaves in a back street of Freiburg”

[Kuznetsova et al., ACL 2013]

http://www.cs.stonybrook.edu/~pkuznetsova/imgcaption/captions1K.html
Motivation: image-level labels are plentiful

“Public bikes in Warsaw during night”

https://www.flickr.com/photos/jacek_kadaj/8776008002/in/photostream/
Approach: search over object’s location at the \textit{training time}

Oquab, Bottou, Laptev and Sivic CVPR 2015

1. Efficient window sliding to find object location hypothesis
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

See also [Kokkinos et al. ’15, Sermanet et al. ’14, Chaftield et al.’14]
1. Efficient window sliding to find object location

Convolutional feature extraction layers trained on 1512 ImageNet classes (Oquab et al., 2014)

Adaptation layers trained on Pascal VOC.
2. Image-level aggregation using global max-pool

Convolutional feature extraction layers trained on 1512 ImageNet classes (Oquab et al., 2014)
3. Multi-label loss function (to allow for multiple objects in image)

Cost function: Sum of log-loss functions over K classes:

$$\ell( f^*(x), y^- ) = \sum_k \log(1 + e^{-y_k f_k(x)})$$
Training with global max-pooling

Training input: 

- image-level labels:
  - ✓ Airplane
  - × Car
  - × Chair …

Airplane score map

max-pool

Car score map

max-pool

Correct label: increase score
Learn discriminative object parts

Incorrect label: decrease score
Suppress Hard Negatives
Training Motorbikes

Evolution of localization score maps over training epochs
Results for weakly-supervised object recognition in Microsoft COCO dataset
Test results in Microsoft COCO: 80 object classes
Test results in Microsoft COCO: 80 object classes
Test results in Microsoft COCO: 80 object classes
Test results in Microsoft COCO: 80 object classes
Test results in Microsoft COCO: 80 object classes

- baseball bat: 0.82
- person: 1.00
- baseball glove: 0.95
- surfboard: 0.98
- person: 1.00
- kite: 0.93
Test results in Microsoft COCO: 80 object classes
Results for weakly-supervised action recognition in Pascal VOC’12 dataset
Test results in Pascal VOC’12: 10 action classes
Test results in Pascal VOC’12: 10 action classes
Test results in Pascal VOC’12: 10 action classes
Test results in Pascal VOC’12: 10 action classes
Test results in Pascal VOC’12: 10 action classes
Test results in Pascal VOC’12: 10 action classes

Failure cases
## Results PASCAL VOC 2012

Object classification

<table>
<thead>
<tr>
<th>Object-level sup.</th>
<th>plane</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>btl</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. NUS-SCM [43]</td>
<td>97.3</td>
<td>84.2</td>
<td>80.8</td>
<td>85.3</td>
<td>60.8</td>
<td>89.9</td>
<td>86.8</td>
<td>89.3</td>
<td>75.4</td>
<td>77.8</td>
<td>75.1</td>
</tr>
<tr>
<td>B. OQUAB [31]</td>
<td>94.6</td>
<td>82.9</td>
<td>88.2</td>
<td>84.1</td>
<td>60.3</td>
<td>89.0</td>
<td>84.4</td>
<td>90.7</td>
<td>72.1</td>
<td>86.8</td>
<td>69.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Image-level sup.</th>
<th>plane</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>btl</th>
<th>bus</th>
<th>car</th>
<th>cat</th>
<th>chair</th>
<th>cow</th>
<th>table</th>
</tr>
</thead>
<tbody>
<tr>
<td>C. Z&amp;F [51]</td>
<td>96.0</td>
<td>77.1</td>
<td>88.4</td>
<td>85.5</td>
<td>55.8</td>
<td>85.8</td>
<td>78.6</td>
<td>91.2</td>
<td>65.0</td>
<td>74.4</td>
<td>67.7</td>
</tr>
<tr>
<td>D. CHATFIELD [4]</td>
<td>96.8</td>
<td>82.5</td>
<td>91.5</td>
<td>88.1</td>
<td>62.1</td>
<td>88.3</td>
<td>81.9</td>
<td>94.8</td>
<td>70.3</td>
<td>80.2</td>
<td>76.2</td>
</tr>
<tr>
<td>E. NUS-HCP [47]</td>
<td><strong>97.5</strong></td>
<td>84.3</td>
<td><strong>93.0</strong></td>
<td><strong>89.4</strong></td>
<td>62.5</td>
<td>90.2</td>
<td>84.6</td>
<td><strong>94.8</strong></td>
<td>69.7</td>
<td><strong>90.2</strong></td>
<td>74.1</td>
</tr>
<tr>
<td>F. FULL IMAGES</td>
<td>95.3</td>
<td>77.4</td>
<td>85.6</td>
<td>83.1</td>
<td>49.9</td>
<td>86.7</td>
<td>77.7</td>
<td>87.2</td>
<td>67.1</td>
<td>79.4</td>
<td>73.5</td>
</tr>
<tr>
<td>G. WEAK SUP</td>
<td>96.7</td>
<td><strong>88.8</strong></td>
<td>92.0</td>
<td>87.4</td>
<td><strong>64.7</strong></td>
<td><strong>91.1</strong></td>
<td><strong>87.4</strong></td>
<td>94.4</td>
<td>74.9</td>
<td>89.2</td>
<td><strong>76.3</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dog</th>
<th>horse</th>
<th>moto</th>
<th>pers</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>83.0</td>
<td>87.5</td>
<td>90.1</td>
<td>95.0</td>
<td>57.8</td>
<td>79.2</td>
<td><strong>73.4</strong></td>
<td><strong>94.5</strong></td>
<td>80.7</td>
<td>82.2</td>
</tr>
<tr>
<td>92.1</td>
<td>93.4</td>
<td>88.6</td>
<td>96.1</td>
<td>64.3</td>
<td>86.6</td>
<td>62.3</td>
<td><strong>91.1</strong></td>
<td>79.8</td>
<td>82.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>dog</th>
<th>horse</th>
<th>moto</th>
<th>pers</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>87.8</td>
<td>86.0</td>
<td>85.1</td>
<td>90.9</td>
<td>52.2</td>
<td>83.6</td>
<td>61.1</td>
<td><strong>91.8</strong></td>
<td>76.1</td>
<td>79.0</td>
</tr>
<tr>
<td>92.9</td>
<td>90.3</td>
<td>89.3</td>
<td>95.2</td>
<td>57.4</td>
<td>83.6</td>
<td>66.4</td>
<td><strong>93.5</strong></td>
<td>81.9</td>
<td>83.2</td>
</tr>
<tr>
<td>93.4</td>
<td>93.7</td>
<td>88.8</td>
<td>93.2</td>
<td>59.7</td>
<td>90.3</td>
<td>61.8</td>
<td><strong>94.4</strong></td>
<td>78.0</td>
<td>84.2</td>
</tr>
<tr>
<td>85.3</td>
<td>90.3</td>
<td>85.6</td>
<td>92.7</td>
<td>47.8</td>
<td>81.5</td>
<td>63.4</td>
<td><strong>91.4</strong></td>
<td>74.1</td>
<td>78.7</td>
</tr>
<tr>
<td><strong>93.7</strong></td>
<td><strong>95.2</strong></td>
<td><strong>91.1</strong></td>
<td><strong>97.6</strong></td>
<td><strong>66.2</strong></td>
<td><strong>91.2</strong></td>
<td>70.0</td>
<td><strong>94.5</strong></td>
<td><strong>83.7</strong></td>
<td><strong>86.3</strong></td>
</tr>
</tbody>
</table>

VGG 89.3

[Oquab, Bottou, Laptev and Sivic, CVPR 2015]
Summary

Training input

image-level labels:

- Person
- Chair
- Airplane
- Reading
- Riding bike
- Running

Test output

More details in http://www.di.ens.fr/willow/research/weakcnn/
What’s next?

a dog **sitting beside** a red **fire hydrant** in a dog park.

a dog **holding** a **skateboard** trotting down a street.