Learning and transferring mid-level image representations using convolutional neural networks

Maxime Oquab,
Léon Bottou, Ivan Laptev, Josef Sivic
Image classification (easy)

Is there a car?

Source: Pascal VOC dataset
Image classification (harder)

Is there a boat?

Source: Pascal VOC dataset
Image classification (v.hard)

Is there a person?

Source: Pascal VOC dataset
Image classification (v.hard)

Source : Pascal VOC dataset
Pascal VOC vs. ImageNet classification

Pascal VOC:
- complex scenes
- 20 object classes
- 10k images

ImageNet:
- object-centric
- 1000 object classes
- 1.2M images
Image classification

- Traditional methods: HOG, SIFT, FV, SVMs, DPM, k-Means, GMM...
  [Csurka et al.'04], [Lowe'04], [Sivic & Zisserman'03],
  [Perronin et al.'10], [Lazebnik et al.'06], [Zhang et al. '07],
  [Boureau et al.'10], [Singh et al.'12], [Juneja et al.'13],
  [Chatfield et al. '11], [van Gemert et al. '08], [Wang et al. '10],
  [Zhou et al. '10], [Dong et al. '13], [Feifei et al. '05],
  [Shotton et al. '05], [Moosmann et al.'05], [Grauman & Darrell '05]
  [Harzallah et al. '09], [...]

- Convolutional neural networks
  ImageNet challenge
  [Krizhevsky et al. 2012]
Brief history of CNNs

- Hubel & Wiesel 1959: *Receptive fields of single neurons in the cat’s striate cortex*
- Fukushima 1980: *Neocognition*
- Rumelhart et al. 1986: *Learning representations by back-propagating errors*

- **LeCun et al. 1989**: *Backpropagation applied to handwritten zip code recognition.*
- LeCun et al. 1998: *Efficient Backprop*
- LeCun et al. 1998: *Gradient-based learning applied to document recognition*
- Hinton & Salakhutdinov, 2006: *Reducing the Dimensionality of Data with Neural Networks*

- Zeiler & Fergus, 2013: *Visualizing and understanding neural networks*
- Sermanet et al. 2013: *Overfeat*
- Donahue et al. 2013: *Decaf*
- Girshick et al. 2014: *Rich feature hierarchies for accurate object detection and semantic segmentation*
- Razavian et al. 2014: *CNN features off-the-shelf, an astounding baseline for recognition*
- Chatfield et al. 2014: *Return of the devil in the details*
Neural Networks

Differentiable operations:
weights trained by gradient descent.
8–layer NN
[Krizhevsky et al.]

60 million parameters:
- ImageNet (1.2M images) : OK
- Pascal VOC (10k images) : ?
Pascal VOC : different task

Car examples from Pascal VOC

Typical car examples from ImageNet
Pascal VOC: different task

Car examples from Pascal VOC

Typical car examples from ImageNet
Solution: multi-scale patch tiling

- Goal: obtain a dataset that looks like ImageNet.

Typical Pascal VOC car example ... in disguise

Small-scale tiling

Large-scale tiling

Typical car examples from ImageNet
Solution: multi-scale patch tiling

- Around 500 tiles per image.
- Multiple scales and positions.
- Label depending on overlap.
First attempt

- Train CNN on Pascal VOC patches:
  - Result: 70.9% mAP.
  - We observe overfitting.
  - State of the art: 82.2% mAP (NUS–PSL).

- How to benefit from the power of neural networks?

We propose transfer learning.
Transfer learning

ImageNet

Source task

Layers L1–L7 → L8

ImageNet network

Source task labels

- African elephant
- Wall clock
- Green snake
- Yorkshire terrier
Transfer learning

Source task

Source task labels
- African elephant
- Wall clock
- Green snake
- Yorkshire terrier

Target task labels
- Chair
- Background
- Person
- TV/monitor

ImageNet

Pascal VOC

Layers L1-L7

Sliding patches

Target task
Transfer learning

ImageNet

Source task

Layers L1-L7

L8

Source task labels

African elephant
Wall clock
Green snake
Yorkshire terrier

Pascal VOC

Sliding patches

Target task

Layers L1-L7

La

Lb

Target task labels

Chair
Background
Person
TV/monitor
Transfer learning

Source task

Layers L1–L7

Transfer parameters

Layers L1–L7

Target task

Source task labels
- African elephant
- Wall clock
- Green snake
- Yorkshire terrier

Target task labels
- Chair
- Background
- Person
- TV/monitor

ImageNet

Pascal VOC

Sliding patches
Second attempt (with pre-training)

- After pre-training on the ILSVRC-2012 dataset, we obtain 78.7% mean AP (no pre-train : 70.9%).
- Significantly better but can we improve more?
- Observe large boosts for dog and bird classes.
- Well-represented groups in ILSVRC-2012.

<table>
<thead>
<tr>
<th></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>
<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>NUS-PSL [48]</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>
<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>73.4</td>
<td>94.5</td>
<td>80.7</td>
<td>82.2</td>
</tr>
<tr>
<td>No pretrain</td>
<td>85.2</td>
<td>75.0</td>
<td>69.4</td>
<td>66.2</td>
<td>48.8</td>
<td>82.1</td>
<td>79.5</td>
<td>79.8</td>
<td>62.4</td>
<td>61.9</td>
<td>49.8</td>
<td>75.9</td>
<td>71.4</td>
<td>82.7</td>
<td>93.1</td>
<td>59.1</td>
<td>69.7</td>
<td>49.3</td>
<td>80.0</td>
<td>76.7</td>
<td>70.9</td>
</tr>
<tr>
<td>Pre-1000C</td>
<td>93.5</td>
<td>78.4</td>
<td>87.7</td>
<td>80.9</td>
<td>57.3</td>
<td>85.0</td>
<td>81.6</td>
<td>89.4</td>
<td>66.9</td>
<td>73.8</td>
<td>62.0</td>
<td>89.5</td>
<td>83.2</td>
<td>87.6</td>
<td>95.8</td>
<td>61.4</td>
<td>79.0</td>
<td>54.3</td>
<td>88.0</td>
<td>78.3</td>
<td>78.7</td>
</tr>
</tbody>
</table>

+18 %  +14 %
Pre-training data

- Inspect 22k classes of the ImageNet tree:
  - «furniture» subtree contains chairs, dining tables, sofas
  - «hoofed mammal» subtree contains sheep, horses, cows
  - ...

- Add 512 classes to the pre-training,
- Result improves from 78.8% to **82.8%** mAP.
- All scores increase, targeted classes improve more.
Computing scores at test time

- We extract 500 multi-scale patches.
- Image score = sum of all patch scores.
- Pixel score = sum of overlapping patches scores (heat maps)
Qualitative results

Source: Pascal VOC’12 test set
Qualitative results

Source: Pascal VOC’12 test set
Qualitative results

Source: Pascal VOC'12 test set
Qualitative results

Source: Pascal VOC’12 test set
Visualizations (aeroplane)

Source: Pascal VOC’12 test set

First false positive
Visualizations (bicycle)

Source: Pascal VOC’12 test set
Visualizations (bicycle)

Source: Pascal VOC’12 test set

First false positive
Visualizations (sheep)

First false positive

Source: Pascal VOC’12 test set
Visualizations (sheep)

Source: Pascal VOC’12 test set
Quantitative results

Pascal VOC’12 object classification:

<table>
<thead>
<tr>
<th></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>
<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>NUS-PSL [48]</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>
<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>73.4</td>
<td>94.5</td>
<td>80.7</td>
<td>82.2</td>
</tr>
<tr>
<td>NO PRETRAIN</td>
<td>85.2</td>
<td>75.0</td>
<td>69.4</td>
<td>66.2</td>
<td>48.8</td>
<td>82.1</td>
<td>79.5</td>
<td>79.8</td>
<td>62.4</td>
<td>61.9</td>
<td>49.8</td>
<td>75.9</td>
<td>71.4</td>
<td>82.7</td>
<td>93.1</td>
<td>59.1</td>
<td>69.7</td>
<td>49.3</td>
<td>80.0</td>
<td>76.7</td>
<td>70.9</td>
</tr>
<tr>
<td>PRE-1000C</td>
<td>93.5</td>
<td>78.4</td>
<td>87.7</td>
<td>80.9</td>
<td>57.3</td>
<td>85.0</td>
<td>81.6</td>
<td>89.4</td>
<td>66.9</td>
<td>73.8</td>
<td>62.0</td>
<td>89.5</td>
<td>83.2</td>
<td>87.6</td>
<td>95.8</td>
<td>67.4</td>
<td>79.0</td>
<td>54.3</td>
<td>88.0</td>
<td>78.3</td>
<td>78.7</td>
</tr>
<tr>
<td>PRE-1000R</td>
<td>93.2</td>
<td>77.9</td>
<td>83.8</td>
<td>80.0</td>
<td>55.8</td>
<td>82.7</td>
<td>79.0</td>
<td>84.3</td>
<td>66.2</td>
<td>71.7</td>
<td>59.5</td>
<td>83.4</td>
<td>81.4</td>
<td>84.8</td>
<td>95.2</td>
<td>59.8</td>
<td>74.9</td>
<td>52.9</td>
<td>83.8</td>
<td>75.7</td>
<td>76.3</td>
</tr>
<tr>
<td>PRE-1512</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>
<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>91.1</td>
<td>79.8</td>
<td>82.8</td>
</tr>
</tbody>
</table>

State of the art: 82.2
Quantitative results

Pascal VOC’12 object classification:

<table>
<thead>
<tr>
<th></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>
<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>NUS-PSL [48]</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>
<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>94.5</td>
<td>90.7</td>
<td>82.2</td>
<td></td>
</tr>
<tr>
<td>No pretrain</td>
<td>85.2</td>
<td>75.0</td>
<td>69.4</td>
<td>66.2</td>
<td>48.8</td>
<td>82.1</td>
<td>79.5</td>
<td>79.8</td>
<td>62.4</td>
<td>61.9</td>
<td>49.8</td>
<td>75.9</td>
<td>71.4</td>
<td>82.7</td>
<td>93.1</td>
<td>59.1</td>
<td>69.7</td>
<td>49.3</td>
<td>80.0</td>
<td>76.7</td>
<td>82.2</td>
</tr>
<tr>
<td>Pre-1000C</td>
<td>93.5</td>
<td>78.4</td>
<td>87.7</td>
<td>80.9</td>
<td>57.3</td>
<td>85.0</td>
<td>81.6</td>
<td>89.4</td>
<td>66.9</td>
<td>73.8</td>
<td>62.0</td>
<td>89.5</td>
<td>83.2</td>
<td>87.6</td>
<td>95.8</td>
<td>67.4</td>
<td>79.0</td>
<td>54.3</td>
<td>88.0</td>
<td>73.3</td>
<td>78.7</td>
</tr>
<tr>
<td>Pre-1000R</td>
<td>93.2</td>
<td>77.9</td>
<td>83.8</td>
<td>80.0</td>
<td>55.8</td>
<td>82.7</td>
<td>79.0</td>
<td>84.3</td>
<td>66.2</td>
<td>71.7</td>
<td>59.5</td>
<td>83.4</td>
<td>81.4</td>
<td>84.8</td>
<td>95.2</td>
<td>59.8</td>
<td>74.9</td>
<td>52.9</td>
<td>83.8</td>
<td>75.7</td>
<td>76.3</td>
</tr>
<tr>
<td>Pre-1512</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>
<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>91.1</td>
<td>79.8</td>
<td>82.8</td>
</tr>
</tbody>
</table>

State of the art: 82.2
No pre-training baseline: 70.9
Quantitative results

Pascal VOC’12 object classification:

|                | plane | bike | bird  | boat  | btl   | bus  | car  | cat  | chair | cow  | table | dog  | horse | moto  | pers  | plant | sheep | sofa  | train | tv    | mAP  |
|----------------|-------|------|-------|-------|-------|------|------|------|-------|------|-------|------|-------|-------|-------|-------|-------|-------|-------|-------|
| NUS-PCL [48]   | 97.3  | 84.2 | 80.8  | 85.3  | 60.8  | 89.9 | 86.8 | 89.3 | 75.4  | 77.8 | 75.1  | 83.0 | 87.5  | 90.1  | 95.0  | 57.8  | 79.2  | 73.4  | 94.5  | 80.7  | 82.2  |
| NO PRETRAIN    | 85.2  | 75.0 | 69.4  | 66.2  | 48.8  | 82.1 | 79.5 | 79.8 | 62.4  | 61.9 | 49.8  | 75.9 | 71.4  | 82.7  | 93.1  | 59.1  | 69.7  | 49.3  | 80.0  | 76.7  | 70.9  |
| PRE-1000C      | 93.5  | 78.4 | 87.7  | 80.9  | 57.3  | 85.0 | 81.6 | 89.4 | 73.8  | 62.0 | 89.5  | 83.2 | 87.6  | 95.8  | 61.4  | 79.0  | 54.3  | 88.0  | 73.3  | 78.7  | 78.7  |
| PRE-1000R      | 93.2  | 77.9 | 83.8  | 80.0  | 55.8  | 82.7 | 79.0 | 84.3 | 66.2  | 71.7 | 59.5  | 83.4 | 81.4  | 84.8  | 95.2  | 59.8  | 74.9  | 52.9  | 83.8  | 75.7  | 76.3  |
| PRE-1512       | 94.6  | 82.9 | 88.2  | 84.1  | 60.3  | 89.0 | 84.4 | 90.7 | 72.1  | 86.8 | 69.0  | 92.1 | 93.4  | 88.6  | 96.1  | 64.3  | 86.6  | 62.3  | 91.1  | 79.8  | 82.8  |

State of the art: 82.2
No pre-training baseline: 70.9
1000 ILSVRC classes: 78.7
Quantitative results

Pascal VOC’12 object classification:

State of the art: 82.2
No pre-training baseline: 70.9
1000 ILSVRC classes: 78.7
1512 classes (our best): 82.8
Quantitative results

Pascal VOC’12 object classification:

<table>
<thead>
<tr>
<th></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>
<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>NUS-PSL [48]</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>
<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>73.4</td>
<td>94.5</td>
<td>80.7</td>
<td>82.2</td>
</tr>
<tr>
<td>No pretrain</td>
<td>85.2</td>
<td>75.0</td>
<td>69.4</td>
<td>66.2</td>
<td>48.8</td>
<td>82.1</td>
<td>79.5</td>
<td>79.8</td>
<td>62.4</td>
<td>61.9</td>
<td>49.8</td>
<td>75.9</td>
<td>71.4</td>
<td>82.7</td>
<td>93.1</td>
<td>59.1</td>
<td>69.7</td>
<td>49.3</td>
<td>80.0</td>
<td>76.7</td>
<td>70.9</td>
</tr>
<tr>
<td>Pre-1000C</td>
<td>93.5</td>
<td>78.4</td>
<td>87.7</td>
<td>80.9</td>
<td>57.3</td>
<td>85.0</td>
<td>81.6</td>
<td>89.4</td>
<td>66.9</td>
<td>73.8</td>
<td>62.0</td>
<td>89.5</td>
<td>83.2</td>
<td>87.6</td>
<td>95.8</td>
<td>67.4</td>
<td>79.0</td>
<td>54.3</td>
<td>88.0</td>
<td>73.3</td>
<td>78.7</td>
</tr>
<tr>
<td>Pre-1000R</td>
<td>93.2</td>
<td>77.9</td>
<td>83.8</td>
<td>80.0</td>
<td>55.8</td>
<td>82.7</td>
<td>79.0</td>
<td>84.3</td>
<td>66.2</td>
<td>71.7</td>
<td>59.5</td>
<td>83.4</td>
<td>81.4</td>
<td>84.8</td>
<td>95.2</td>
<td>59.8</td>
<td>74.9</td>
<td>52.9</td>
<td>83.8</td>
<td>75.7</td>
<td>76.3</td>
</tr>
<tr>
<td>Pre-1512</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>
<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>91.1</td>
<td>79.8</td>
<td>82.8</td>
</tr>
</tbody>
</table>

State of the art: 82.2
No pre-training baseline: 70.9
1000 ILSVRC classes: 78.7
Random 1000 classes: 76.3
1512 classes (our best): 82.8
Different task: action classification (still images)

- Playing instrument
- Jumping
- Running

Source: Pascal VOC'12 Action classification test set
State-of-the-art 70.2% mAP result
Different task: action classification (still images)

playing instrument

jumping

running

Source: Pascal VOC’12  Action classification test set
State-of-the-art 70.2% mAP result
Qualitative results
(reading)
Qualitative results
(playing instrument)
Qualitative results (phoning)
Take-home messages

● **Transfer learning with CNNs avoids overfitting**
  ● See also: [Girshick et al.’14], [Sermanet et al.’13], [Donahue et al. ’13], [Zeiler & Fergus ’13], [Razavian et al. ’14], [Chatfield et al. ’14]

● **We study the effect of pre-training data:**
  ● More pre-training data => better
  ● Related pre-training data => even better

● **Transfer to action classification.**

  ● Implementation (Torch7 modules) available soon
  ● Includes efficient and flexible GPU training code
This work

Bounding box annotation is expensive. Can we avoid it?

YES WE CAN!
Follow-up work

- Weakly supervised, no bounding boxes required
- 82.8 => 86.3% mean AP on VOC classification
- Appearing on Arxiv soon (check our webpage)

image-level labels only

«dog» heatmap
Weakly supervised object recognition with convolutional neural networks

Maxime Oquab,
Léon Bottou, Ivan Laptev, Josef Sivic

(All following slides stolen from Josef Sivic)
Are bounding boxes needed for training CNNs?

Image-level labels: Bicycle, Person

[Oquab, Bottou, Laptev, Sivic, In submission, 2014]
Motivation: labeling bounding boxes is tedious
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
Let the algorithm localize the object in the image

Example training images with bounding boxes

<table>
<thead>
<tr>
<th>Typical</th>
<th>Cluttered</th>
<th>Cropped</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Typical Training Image" /></td>
<td><img src="image2.png" alt="Cluttered Training Image" /></td>
<td><img src="image3.png" alt="Cropped Training Image" /></td>
</tr>
</tbody>
</table>

The locations of objects learnt by the CNN

NB: Related to multiple instance learning, e.g. [Viola et al.'05] and weakly supervised object localization, e.g. [Pandy and Lazebnik’11], [Prest et al.’12], ...

[Oquab, Bottou, Laptev, Sivic, In submission, 2014]
Approach: search over object’s location

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 [Sermanet et al. ’14] and [Chaftield et al.’14]
Approach: search over object’s location

Note: All FC-layers are now large convolutions

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 [Sermanet et al. ’14] and [Chaftield et al.’14]
Approach: search over object’s location

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 [Sermanet et al. ’14] and [Chaftiel et al.’14]
Search for objects using max-pooling

Correct label: increase score for this class
Incorrect label: decrease score for this class
Search for objects using max-pooling

learn from :

What is the effect of errors?

learn from :

Most discriminative part

at training time

Hardest negative

<=>
Multi-scale training and testing

Figure 3: Weakly supervised training

Figure 4: Multiscale object recognition
Evolution of maps during training

aeroplane - training iteration 0030
Results

<table>
<thead>
<tr>
<th></th>
<th>mAP</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>
</tr>
</thead>
<tbody>
<tr>
<td>A. Zeiler and Fergus [40]</td>
<td>79.0</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>
</tr>
<tr>
<td>B. Oquab et al. [26]</td>
<td>82.8</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>
</tr>
<tr>
<td>C. Chatfield et al. [4]</td>
<td>83.2</td>
<td><strong>96.8</strong></td>
<td>82.5</td>
<td>91.5</td>
<td><strong>88.1</strong></td>
<td>62.1</td>
<td>88.3</td>
<td>81.9</td>
<td><strong>94.8</strong></td>
<td>70.3</td>
<td>80.2</td>
</tr>
<tr>
<td>D. Full Images (Our)</td>
<td>78.7</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>
</tr>
<tr>
<td>E. Strong+Weak (Our)</td>
<td>86.0</td>
<td>96.5</td>
<td>88.3</td>
<td>91.9</td>
<td>87.7</td>
<td>64.0</td>
<td>90.3</td>
<td>86.8</td>
<td>93.7</td>
<td>74.0</td>
<td><strong>89.8</strong></td>
</tr>
<tr>
<td>F. Weak Supervision (Our)</td>
<td>86.3</td>
<td>96.7</td>
<td><strong>88.8</strong></td>
<td><strong>92.0</strong></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><strong>74.9</strong></td>
<td>89.2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>table</th>
<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>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67.7</td>
<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>91.8</td>
<td>76.1</td>
</tr>
<tr>
<td></td>
<td>69.0</td>
<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>91.1</td>
<td>79.8</td>
</tr>
<tr>
<td></td>
<td>76.2</td>
<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>93.5</td>
<td>81.9</td>
</tr>
<tr>
<td></td>
<td>73.5</td>
<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>91.4</td>
<td>74.1</td>
</tr>
<tr>
<td></td>
<td>76.3</td>
<td>93.4</td>
<td>94.9</td>
<td>91.2</td>
<td>97.3</td>
<td>66.0</td>
<td>90.9</td>
<td>69.9</td>
<td>93.9</td>
<td>83.2</td>
</tr>
<tr>
<td></td>
<td><strong>76.3</strong></td>
<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><strong>70.0</strong></td>
<td><strong>94.5</strong></td>
<td><strong>83.7</strong></td>
</tr>
</tbody>
</table>

- Localizing objects by sliding helps
- Full supervision does not improve over weak supervision
- New state-of-the-art on Pascal VOC 2012 object classification
Object localization examples in testing data

(a) Representative true positives

(b) Top ranking false positives

Figure 5: Output probability maps on representative images of several categories from the Pascal VOC 2012 test set. The rightmost column contains the highest-scoring false positive (according to our judgement) for each of these categories. Note that the proposed method provides an approximate localization of the object or its discriminative parts in the image despite being trained only from image-level labels without providing the location of the objects in the training data.

Please see more qualitative localization results for training and test images in the supplementary material.
Are bounding boxes harmful?

Output of the fully supervised CVPR’14 network:

- Why a higher score on the dog’s head?
- Responses are inconsistent with the annotations.
- Maybe we are doing it wrong.
Are bounding boxes harmful?

Bounding boxes are NOT alignment.

Should be treated as **guidance** not supervision (at least for object classification)