Neural Object Detection

Hyun Soo Park
Image Classification vs. Object Detection

Image classification

Object detection
Input image
OBJECT DETECTION PIPELINE

Input image

Localization

Sliding window
**Object Detection Pipeline**

- **Input image**
- **Localization**
  - Sliding window
- **Feature extraction**
  - HOG/SIFT/BoW
**Object Detection Pipeline**

- **Input image**
- **Localization**
  - Sliding window
- **Feature extraction**
  - HOG/SIFT/BoW
- **Classification**
  - SVM

**Limitations**
- Slow
- ~ 1M of evaluations
- Shallow
- n-classifiers
**R-CNN (GIRSHICK ET AL.)**

- **Input image**
- **Localization**
  - Sliding window
- **Feature extraction**
  - CNN
- **Classification**
  - SVM

**Limitations**
- Slow
  - ~ 1M of evaluations

**Classification**
- Dog
- Not dog
- n-classifiers
R-CNN (Girshick et al.)

Input image | Localization | Feature extraction | Classification
---|---|---|---
Region proposal | CNN | SVM

Limitations
- 2000 evaluations
- Deep
- n-classifiers
**R-CNN (Girshick et al.)**

- **Input image**
- **Localization**
  - **Region proposal**
- **Feature extraction**
  - **CNN**
- **Classification**
  - **SVM**

**Limitations**
- 2000 evaluations
- Deep
- n-classifiers
**Objectness (Selective Search ~ Uijlings et al.)**

- Merging regions from over-segmentation
- Objectness classification via BoW on merged regions
OBJECTNESS (SELECTIVE SEARCH~UIJLINGS ET AL.)
**R-CNN (Girshick et al.)**

- **Input image**
- **Localization**
  - Region proposal
- **Feature extraction**
  - CNN
- **Classification**
  - SVM

**Limitations**
- 2000 evaluations

**Deep**

**Dog**

**Not dog**
**Domain Adaptation (Image Class. → Object Det.)**

1000 image classes (~15M)

4096x1000
DOMAIN ADAPTATION (IMAGE CLASS. $\rightarrow$ OBJECT DET.)

1000 image classes ($\sim$15M)

4096x21

$\frac{\partial L}{\partial w_n}$

w/ small learning rate

20 object classes ($\sim$20K)
Region-CNN

Region proposal
REGION-CNN

Region proposal

GT
**TRAINING DATA**

Region proposal

GT
**REGION-CNN**

Region proposal

GT

GT

Region proposal

+ if
intersection of union (IoU) > 0.5
REGION-CNN

Region proposal

GT

Region proposal

(x, y)

W

H

Detection offset

GT

Region proposal

+ if

intersection of union (IoU) > 0.5
REGION-CNN

Region proposal

\[(x, y)\]

\[W\]

\[H\]

GT

Recall

Pool5

AlexNet

FC

GT

Region proposal

+ if intersection of union (IoU) > 0.5
**REGION-CNN**

Region proposal

Region classification person?

Bounding box regression (x, y, W, H)

Relative offset:

\[
    x = \frac{P_x - G_x}{P_w}, \quad y = \frac{P_y - G_y}{P_H}, \quad W = \log\left(\frac{G_w}{P_w}\right), \quad H = \log\left(\frac{G_H}{P_H}\right)
\]

\[(P_x, P_y, P_w, P_H) \quad (G_x, G_y, G_w, G_H)\]
**R-CNN (Girshick et al.)**

<table>
<thead>
<tr>
<th>Input image</th>
<th>Localization</th>
<th>Feature extraction</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Region proposal</td>
<td>CNN</td>
<td>SVM</td>
</tr>
<tr>
<td>Limitations</td>
<td>2000 evaluations</td>
<td>Deep</td>
<td>n-classifiers</td>
</tr>
</tbody>
</table>
Classification

Max margin SVM classifier

\[ x \cdot w + b > 0 \quad \text{Positive D.} \]

\[ x \cdot w + b < 0 \quad \text{Negative D.} \]
<table>
<thead>
<tr>
<th>VOC 2007 test</th>
<th>aero</th>
<th>bike</th>
<th>bird</th>
<th>boat</th>
<th>bottle</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>mbike</th>
<th>person</th>
<th>plant</th>
<th>sheep</th>
<th>sofa</th>
<th>train</th>
<th>tv</th>
<th>mAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN pool(_5)</td>
<td>51.8</td>
<td>60.2</td>
<td>36.4</td>
<td>27.8</td>
<td>23.2</td>
<td>52.8</td>
<td>60.6</td>
<td>49.2</td>
<td>18.3</td>
<td>47.8</td>
<td>44.3</td>
<td>40.8</td>
<td>56.6</td>
<td>58.7</td>
<td>42.4</td>
<td>23.4</td>
<td>46.1</td>
<td>36.7</td>
<td>51.3</td>
<td>55.7</td>
<td>44.2</td>
</tr>
<tr>
<td>R-CNN fc(_6)</td>
<td>59.3</td>
<td>61.8</td>
<td>43.1</td>
<td>34.0</td>
<td>25.1</td>
<td>53.1</td>
<td>60.6</td>
<td>52.8</td>
<td>21.7</td>
<td>47.8</td>
<td>42.7</td>
<td>47.8</td>
<td>52.5</td>
<td>58.5</td>
<td>44.6</td>
<td>25.6</td>
<td>48.3</td>
<td>34.0</td>
<td>53.1</td>
<td>58.0</td>
<td>46.2</td>
</tr>
<tr>
<td>R-CNN fc(_7)</td>
<td>57.6</td>
<td>57.9</td>
<td>38.5</td>
<td>31.8</td>
<td>23.7</td>
<td>51.2</td>
<td>58.9</td>
<td>51.4</td>
<td>20.0</td>
<td>50.5</td>
<td>40.9</td>
<td>46.0</td>
<td>51.6</td>
<td>55.9</td>
<td>43.3</td>
<td>23.3</td>
<td>48.1</td>
<td>35.3</td>
<td>51.0</td>
<td>57.4</td>
<td>44.7</td>
</tr>
<tr>
<td>R-CNN FT pool(_5)</td>
<td>58.2</td>
<td>63.3</td>
<td>37.9</td>
<td>27.6</td>
<td>26.1</td>
<td>54.1</td>
<td>66.9</td>
<td>51.4</td>
<td>26.7</td>
<td>55.5</td>
<td>43.4</td>
<td>43.1</td>
<td>57.7</td>
<td>59.0</td>
<td>45.8</td>
<td>28.1</td>
<td>50.8</td>
<td>40.6</td>
<td>53.1</td>
<td>56.4</td>
<td>47.3</td>
</tr>
<tr>
<td>R-CNN FT fc(_6)</td>
<td>63.5</td>
<td>66.0</td>
<td>47.9</td>
<td>37.7</td>
<td>29.9</td>
<td>62.5</td>
<td>70.2</td>
<td>60.2</td>
<td>32.0</td>
<td>57.9</td>
<td>47.0</td>
<td>53.5</td>
<td>60.1</td>
<td>64.2</td>
<td>52.2</td>
<td>31.3</td>
<td>55.0</td>
<td>50.0</td>
<td>57.7</td>
<td>63.0</td>
<td>53.1</td>
</tr>
<tr>
<td>R-CNN FT fc(_7)</td>
<td>64.2</td>
<td>69.7</td>
<td>50.0</td>
<td>41.9</td>
<td>32.0</td>
<td>62.6</td>
<td>71.0</td>
<td>60.7</td>
<td>32.7</td>
<td>58.5</td>
<td>46.5</td>
<td>56.1</td>
<td>60.6</td>
<td>66.8</td>
<td>54.2</td>
<td>31.5</td>
<td>52.8</td>
<td>48.9</td>
<td>57.9</td>
<td>64.7</td>
<td>54.2</td>
</tr>
<tr>
<td>R-CNN FT fc(_7) BB</td>
<td>68.1</td>
<td>72.8</td>
<td>56.8</td>
<td>43.0</td>
<td>36.8</td>
<td>66.3</td>
<td>74.2</td>
<td>67.6</td>
<td>34.4</td>
<td>63.5</td>
<td>54.5</td>
<td>61.2</td>
<td>69.1</td>
<td>68.6</td>
<td>58.7</td>
<td>33.4</td>
<td>62.9</td>
<td>51.1</td>
<td>62.5</td>
<td>64.8</td>
<td>58.5</td>
</tr>
</tbody>
</table>

**Table 2: Detection average precision (%) on VOC 2007 test.** Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval. Row 7 includes a simple bounding-box regression (BB) stage that reduces localization errors (Section C). Rows 8-10 present DPM methods as a strong baseline. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG.
Figure 4: Top regions for six $\text{pool}_4$ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).
Object Detection Pipeline

Input image

Localization
Region proposal

Feature extraction
CNN

Classification
SVM

Limitations
2000 evaluations
Deep
n-classifiers

Too slow in testing time
Post-hoc optimization

Dog
Not dog
Redundant Feature Map
**Feature Map Recycling**

Ground truth region

$P_{human} = (x, y, W, H)$

Input: image/ROI

Region classification
person?

Bounding box regression
(x,y,W,H)

Conv. layers

FC

Rol pooling
**Back-propagation**

Ground truth region

\[ P_{\text{human}} = (x, y, W, H) \]
<table>
<thead>
<tr>
<th></th>
<th>Fast R-CNN</th>
<th></th>
<th>R-CNN</th>
<th></th>
<th>SPPnet †L</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>M</td>
<td>L</td>
<td>S</td>
<td>M</td>
</tr>
<tr>
<td>train time (h)</td>
<td>1.2</td>
<td>2.0</td>
<td>9.5</td>
<td>22</td>
<td>28</td>
</tr>
<tr>
<td>train speedup</td>
<td>18.3×</td>
<td>14.0×</td>
<td>8.8×</td>
<td>1×</td>
<td>1×</td>
</tr>
<tr>
<td>test rate (s/im)</td>
<td>0.10</td>
<td>0.15</td>
<td>0.32</td>
<td>9.8</td>
<td>12.1</td>
</tr>
<tr>
<td>▶ with SVD</td>
<td>0.06</td>
<td>0.08</td>
<td>0.22</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>test speedup</td>
<td>98×</td>
<td>80×</td>
<td>146×</td>
<td>1×</td>
<td>1×</td>
</tr>
<tr>
<td>▶ with SVD</td>
<td>169×</td>
<td>150×</td>
<td>213×</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>VOC07 mAP</td>
<td>57.1</td>
<td>59.2</td>
<td>66.9</td>
<td>58.5</td>
<td>60.2</td>
</tr>
<tr>
<td>▶ with SVD</td>
<td>56.5</td>
<td>58.7</td>
<td>66.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>method</td>
<td>train set</td>
<td>aero</td>
<td>bike</td>
<td>bird</td>
<td>boat</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>BabyLearning</td>
<td>Prop.</td>
<td>78.0</td>
<td>74.2</td>
<td>61.3</td>
<td>45.7</td>
</tr>
<tr>
<td>NUS_NIN_c2000</td>
<td>Unk.</td>
<td>80.2</td>
<td>73.8</td>
<td>61.9</td>
<td>43.7</td>
</tr>
<tr>
<td>R-CNN BB [10]</td>
<td>12</td>
<td>79.6</td>
<td>72.7</td>
<td>61.9</td>
<td>41.2</td>
</tr>
<tr>
<td>FRCN [ours]</td>
<td>12</td>
<td>80.3</td>
<td>74.7</td>
<td>66.9</td>
<td>46.9</td>
</tr>
<tr>
<td>FRCN [ours]</td>
<td>07++12</td>
<td><strong>82.3</strong></td>
<td><strong>78.4</strong></td>
<td><strong>70.8</strong></td>
<td><strong>52.3</strong></td>
</tr>
</tbody>
</table>

Table 3. **VOC 2012 test** detection average precision (%). BabyLearning and NUS_NIN_c2000 use networks based on [17]. All other methods use VGG16. Training set key: see Table 2. **Unk.**: unknown.
<table>
<thead>
<tr>
<th>method</th>
<th>classifier</th>
<th>S</th>
<th>M</th>
<th>L</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN [9, 10]</td>
<td>SVM</td>
<td>58.5</td>
<td>60.2</td>
<td>66.0</td>
</tr>
<tr>
<td>FRCN [ours]</td>
<td>SVM</td>
<td>56.3</td>
<td>58.7</td>
<td>66.8</td>
</tr>
<tr>
<td>FRCN [ours]</td>
<td>softmax</td>
<td>57.1</td>
<td>59.2</td>
<td>66.9</td>
</tr>
</tbody>
</table>
FASTER RCNN (REN ET AL.)

Input image

Localization
Region proposal

Feature extraction
CNN

Classification
SVM

Dog
Not dog

One network
Region Proposal Network

Anchor location
Region Proposal Network

Proposed regions
3 scales
Region Proposal Network

Proposed regions
3 scales
3 aspect ratio

1 anchor location

9 proposals per anchor
Region Proposal Network

Proposed regions
3 scales
3 aspect ratio

9 proposals per anchor

Region classification
Object? or not?

Bounding box regression
(x, y, W, H)

Anchor location

FC
Faster R-CNN

Object proposal per pixel

Region classification object?

Bounding box regression (x,y,W,H)

Conv. layers

FC

Input: image
Faster RCNN

Input: image

Conv. layers

Object proposal per pixel

Region classification object?

Bounding box regression (x, y, W, H)

FC
FASTER RCNN

Object proposal per pixel

| Region classification
| object? |

| Bounding box regression |
| (x,y,W,H) |

FC

Conv. layers

Rol pooling

FC

Object classification
person?

Input: image
Table 2: Detection results on **PASCAL VOC 2007 test set**. The detector is Fast R-CNN and VGG-16. Training data: “07”: VOC 2007 trainval, “07+12”: union set of VOC 2007 trainval and VOC 2012 trainval. For RPN, the train-time proposals for Fast R-CNN are 2k. †: this was reported in [5]; using the repository provided by this paper, this number is higher (68.0±0.3 in six runs).

<table>
<thead>
<tr>
<th>method</th>
<th># proposals</th>
<th>data</th>
<th>mAP (%)</th>
<th>time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>2k</td>
<td>07</td>
<td>66.9†</td>
<td>1830</td>
</tr>
<tr>
<td>SS</td>
<td>2k</td>
<td>07+12</td>
<td>70.0</td>
<td>1830</td>
</tr>
<tr>
<td>RPN+VGG, unshared</td>
<td>300</td>
<td>07</td>
<td>68.5</td>
<td>342</td>
</tr>
<tr>
<td>RPN+VGG, shared</td>
<td>300</td>
<td>07</td>
<td>69.9</td>
<td><strong>198</strong></td>
</tr>
<tr>
<td>RPN+VGG, shared</td>
<td>300</td>
<td>07+12</td>
<td><strong>73.2</strong></td>
<td><strong>198</strong></td>
</tr>
</tbody>
</table>
**Object Detection Pipeline**

- **Input image**
- **Localization**
  - Region proposal
  - Per grid classification
- **Feature extraction**
  - CNN
- **Classification**
  - SVM

Diagram showing a relationship between dogs and not dogs.
YOLO (You Only Look Once, Redmon et al.)

S × S grid on input

Bounding boxes + confidence

Class probability map

Final detections
<table>
<thead>
<tr>
<th>Real-Time Detectors</th>
<th>Train</th>
<th>mAP</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>100Hz DPM [31]</td>
<td>2007</td>
<td>16.0</td>
<td>100</td>
</tr>
<tr>
<td>30Hz DPM [31]</td>
<td>2007</td>
<td>26.1</td>
<td>30</td>
</tr>
<tr>
<td>Fast YOLO</td>
<td>2007+2012</td>
<td>52.7</td>
<td><strong>155</strong></td>
</tr>
<tr>
<td>YOLO</td>
<td>2007+2012</td>
<td><strong>63.4</strong></td>
<td>45</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Less Than Real-Time</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Fastest DPM [38]</td>
<td>2007</td>
<td>30.4</td>
<td>15</td>
</tr>
<tr>
<td>R-CNN Minus R [20]</td>
<td>2007</td>
<td>53.5</td>
<td>6</td>
</tr>
<tr>
<td>Fast R-CNN [14]</td>
<td>2007+2012</td>
<td>70.0</td>
<td>0.5</td>
</tr>
<tr>
<td>Faster R-CNN VGG-16[28]</td>
<td>2007+2012</td>
<td>73.2</td>
<td>7</td>
</tr>
<tr>
<td>YOLO VGG-16</td>
<td>2007+2012</td>
<td>66.4</td>
<td>21</td>
</tr>
</tbody>
</table>