Object detection
The Task
Step 1: Object proposals

- Use segmentation to produce ~5K candidates

*Selective Search for Object Recognition*

J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders

In International Journal of Computer Vision 2013.
R-CNN: Regions with CNN features

Input image

Extract region proposals (~2k / image)

Compute CNN features

Classify regions (linear SVM)

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

R. Girshick, J. Donahue, T. Darrell, J. Malik

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

Slide credit: Ross Girshick
Faster R-CNN
Faster R-CNN

• At each location, consider boxes of many different sizes and aspect ratios
Faster R-CNN

• At each location, consider boxes of many different sizes and aspect ratios
Faster R-CNN

- At each location, consider boxes of many different sizes and aspect ratios
Faster R-CNN

• s scales * a aspect ratios = sa anchor boxes
• Use convolutional layer on top of filter map to produce sa scores
• Pick top few boxes as proposals
## Faster R-CNN

<table>
<thead>
<tr>
<th>Method</th>
<th>mean AP (PASCAL VOC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast R-CNN</td>
<td>65.7</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>67.0</td>
</tr>
</tbody>
</table>
Impact of Feature Extractors

<table>
<thead>
<tr>
<th>ConvNet</th>
<th>mean AP (PASCAL VOC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>VGG</td>
<td>70.4</td>
</tr>
<tr>
<td>ResNet 101</td>
<td>73.8</td>
</tr>
</tbody>
</table>
## Impact of Additional Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Training data</th>
<th>mean AP (PASCAL VOC 2012 Test)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fast R-CNN</td>
<td>VOC 12 Train (10K)</td>
<td>65.7</td>
</tr>
<tr>
<td>Fast R-CNN</td>
<td>VOC07 Trainval + VOC 12 Train</td>
<td>68.4</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>VOC 12 Train (10K)</td>
<td>67.0</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>VOC07 Trainval + VOC 12 Train</td>
<td>70.4</td>
</tr>
</tbody>
</table>
The R-CNN family of detectors

![Graph showing mean AP for different R-CNN models.]

- R-CNN/10K/VGG
- Fast R-CNN/10K/VGG
- Fast R-CNN/20K/VGG
- Faster R-CNN/20K/VGG
- Faster R-CNN/20K/ResNet101
SSD (Single Shot Detector)

• Why go through separate proposals?
• Directly produce class-specific scores at each location for every scale and aspect ratio
  • $s$ scales * $a$ aspects * $c$ classes = sac scores per location

## SSD

<table>
<thead>
<tr>
<th>Method</th>
<th>mean AP (PASCAL VOC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN</td>
<td>70.4</td>
</tr>
<tr>
<td>SSD</td>
<td>74.9</td>
</tr>
</tbody>
</table>
A comprehensive evaluation

Speed and accuracy trade-offs for modern convolutional object detectors

Alireza Fathi, Anoop Korattikara, Chen Sun, Ian Fischer, Jonathan Huang, Kevin Murphy, Menglong Zhu, Sergio Guadarrama, Vivek Rathod, Yang Song, Zbigniew Wojna

CVPR 2017
Detecting small objects

- Small objects get low resolution features
Feature pyramid networks

### Feature pyramid networks

<table>
<thead>
<tr>
<th>Faster R-CNN</th>
<th>proposals</th>
<th>feature</th>
<th>head</th>
<th>lateral?</th>
<th>top-down?</th>
<th>AP@0.5</th>
<th>AP</th>
<th>APs</th>
<th>APm</th>
<th>APl</th>
</tr>
</thead>
<tbody>
<tr>
<td>(*) baseline from He et al. [16]</td>
<td>RPN, (C_4)</td>
<td>(C_4)</td>
<td>conv5</td>
<td></td>
<td></td>
<td>47.3</td>
<td>26.3</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>(a) baseline on conv4</td>
<td>RPN, (C_4)</td>
<td>(C_4)</td>
<td>conv5</td>
<td></td>
<td></td>
<td>53.1</td>
<td>31.6</td>
<td>13.2</td>
<td>35.6</td>
<td>47.1</td>
</tr>
<tr>
<td>(b) baseline on conv5</td>
<td>RPN, (C_5)</td>
<td>(C_5)</td>
<td>2fc</td>
<td></td>
<td></td>
<td>51.7</td>
<td>28.0</td>
<td>9.6</td>
<td>31.9</td>
<td>43.1</td>
</tr>
<tr>
<td>(c) FPN</td>
<td>RPN, ({P_k})</td>
<td>({P_k})</td>
<td>2fc</td>
<td>✓</td>
<td>✓</td>
<td>56.9</td>
<td>33.9</td>
<td>17.8</td>
<td>37.7</td>
<td>45.8</td>
</tr>
</tbody>
</table>

Other details - Non-max suppression
Other details - Non-max suppression

• Go down the list of detections starting from highest scoring
• Eliminate any detection that overlaps highly with a higher scoring detection
• Separate, heuristic step
A comprehensive evaluation

Speed and accuracy trade-offs for modern convolutional object detectors
Alireza Fathi, Anoop Korattikara, Chen Sun, Ian Fischer, Jonathan Huang, Kevin Murphy, Menglong Zhu, Sergio Guadarrama, Vivek Rathod, Yang Song, Zbigniew Wojna
CVPR 2017
Challenges with RoI pooling

• Still scales with the number of bounding boxes
RoI pooling (Faster R-CNN) vs convolution (SSD)

• Why does Faster R-CNN tend to be more accurate?
• RoI pooling takes information from the precise box
• Convolution takes information from just the kernel window
R-FCN

• Can we get benefits of RoI pooling in linear time?

Other details - ROI Align

• Snapping box to grid introduces quantization artifacts
• Instead, use bilinear interpolation

Instance segmentation
Till now

Object Detection

Image Classification

Semantic Segmentation
Fine-grained Localization

Instance Segmentation
Evaluation Protocol

• Sort predicted instances by confidence
• Match prediction to closest annotation based on segment overlap
  • If segment overlap > threshold, correct
Evaluation Protocol

Labels = [ ✔ ✔ ✗ ✔ ✔ ✗ ✔ .... ]
Scores = [ 0.90 0.87 0.82 0.78 0.70 0.69 0.60 .... ]
Evaluation protocol

![Average Precision (AP)](Chart)
The COCO Challenge

mscoco.org
Two strategies

• Segment then classify
  • Use bottom-up techniques to come up with *segment* proposals
  • Classify segment proposals with convnets
  • Segmentation is category agnostic
  • Modification: use convnets to produce segmentation proposals

• Detect then segment
  • Use standard object detection to produce boxes
  • Segment boxes
  • Segmentation is *category specific*
Box proposals

• Use segmentation to produce ~5K candidates

Selective Search for Object Recognition
J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, A. W. M. Smeulders
In International Journal of Computer Vision 2013.
Segment proposals

R-CNN for instance segmentation

Simultaneous Detection and Segmentation Bharath Hariharan, Pablo Arbelaez, Ross Girshick, Jitendra Malik. In ECCV 2014
Fast R-CNN for instance segmentation

Fast R-CNN for instance segmentation

Two strategies

• Segment then classify
  • Use bottom-up techniques to come up with *segment* proposals
  • Classify segment proposals with convnets
  • Segmentation is category agnostic
  • Modification: use convnets to produce segmentation proposals

• Detect then segment
  • Use standard object detection to produce boxes
  • Segment boxes
  • Segmentation is *category specific*
Predicting segmentation for bounding boxes

• Given a detection, predict segmentation
Predicting segmentation for bounding boxes
Mask R-CNN

The resolution issue: Skip connections
Skip connections with RoI pooling

The resolution issue: Skip connections

Fully convolutional or not?

• Convolution is *location invariant*
• Semantic segmentation is *location invariant*
• Figure-ground segmentation of a detected bounding box is *not location invariant*
Fully convolutional or not?

- Convolution is cheaper
- Local connections without parameter sharing?
  - Too many parameters?
  - Predicting 50x50 segmentation: 2500 times more parameters
- Predict lower resolution, then upsample

Fully convolutional:

Non-fully convolutional:
Instance fully-convolutional networks

Final results - what works?

• First detect, then segment
• Big problem for instance segmentation is object detection
• Mask R-CNN (Faster R-CNN + convolution on RoI Pooled feature to get masks) is good starting point
Human pose estimation
The task

• Mark joint locations for person
• Nose
• Right/left shoulder
• Right/left elbow
• Right/left hip
• ...

![Image of two people with joint locations marked with colored lines]
Two versions of task

• Assume people have been detected
• Rough bounding box given
• Key info available:
  • scale
  • only 1 location per joint
• Pros: disentangles detection and pose estimation
• Cons: unrealistic

• Tabula rasa without detections
• Challenge: no idea of scale or number
• Possible opportunity: use keypoint estimates to improve detections
• Pros: realistic
• Cons: conflates detection and pose estimation
Pose estimation given detection
Evaluation metric - given detection

• Evaluate every keypoint separately
• For each person, check if keypoint is correct
• Compute fraction of people for which keypoint is correct: PCK (Probability of Correct Keypoint)
Evaluation metric - given detection

\[ \frac{d}{h} < \alpha ? \]
R-CNN: Regions with CNN features

Input image

Extract region proposals (~2k / image)

Compute CNN features

Classify regions (linear SVM)

Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation

R. Girshick, J. Donahue, T. Darrell, J. Malik

IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014

Slide credit: Ross Girshick
Bounding-box regression

\[
\begin{align*}
\text{original} & \quad \text{predicted} \\
\Delta h \times h + h & \quad \Delta h \times h + h \\
\Delta w \times w + w & \\
(x, y) & \\
\end{align*}
\]

Slide credit: Ross Girshick
Strategy 1: Regression

Strategy 1: Regression

• Assumes global object features has enough information for accurate localization
  • Localization info missing due to subsampling?
• Solution: Refinement!

Strategy 1: Regression

• Multimodal distributions?

Minimizer of

\[
\mathbb{E}(\|X - x_{\text{pred}}\|^2)
\]
Strategy 2: Heatmaps
Strategy 2: Heatmaps

• Still have the resolution issue
• Same solutions
  • Dilation?
  • Multiple layers?
  • Multiple image scales?
• abc