Object detection
Faster R-CNN
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
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
Other details - ROI Align

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

Detecting small objects

- Small objects get low resolution features
Feature pyramid networks

Standard detection

Detection on image pyramid

Detection using multiple layers

Detection using feature pyramid layers

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>

Deformable conv

- Convolution uses the same kernel size always
- Want to capture more or less object region depending on scale
  - Or more generally properties of pixel
- Idea: *learn what pixels to combine using convolution*

Figure 2: Illustration of $3 \times 3$ deformable convolution.
Instance segmentation
Till now

horse, person

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

\[
\text{segment} = \frac{\text{intersection}}{\text{union}} = \frac{\bigcap}{\bigcup}
\]
Evaluation Protocol

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

![Average Precision (AP) graph]
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
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*
Detect then segment

• How should we segment a detected object?
• We have already computed features using ROI Pooling
• Idea: use features to predict mask!
  • Can either use a simple linear layer
  • Or can use convolution
  • Issue: can be very coarse
Skip connections with RoI pooling

• Finer-grained segmentation: tap into earlier layers

Skip connections for finer-grained details

Mask R-CNN

• With deeper networks and ROI Align, skip connections not needed (?)

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