Recent developments in object detection

PASCAL VOC

Before deep convnets

Using deep convnets
Beyond sliding windows: Region proposals

- **Advantages:**
  - Cuts down on number of regions detector must evaluate
  - Allows detector to use more powerful features and classifiers
  - Uses low-level *perceptual organization* cues
  - Proposal mechanism can be category-independent
  - Proposal mechanism can be trained
Selective search

Use segmentation

Selective search: Basic idea

- Use hierarchical segmentation: start with small superpixels and merge based on diverse cues

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013
Evaluation of region proposals

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders,
Selective Search for Object Recognition, IJCV 2013
Selective search detection pipeline

- Feature extraction: color SIFT, codebook of size 4K, spatial pyramid with four levels = 360K dimensions

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013
Another proposal method: EdgeBoxes

- Box score: number of edges in the box minus number of edges that overlap the box boundary
- Uses a trained edge detector
- Uses efficient data structures for fast evaluation
- Gets 75% recall with 800 boxes (vs. 1400 for Selective Search), is 40 times faster

R-CNN: Region proposals + CNN features

Source: R. Girshick

R. Girshick, J. Donahue, T. Darrell, and J. Malik, 
R-CNN details

- **Regions**: ~2000 Selective Search proposals
- **Network**: AlexNet *pre-trained* on ImageNet (1000 classes), *fine-tuned* on PASCAL (21 classes)
- **Final detector**: warp proposal regions, extract fc7 network activations (4096 dimensions), classify with linear SVM
- **Bounding box regression** to refine box locations
- **Performance**: mAP of 53.7% on PASCAL 2010 (vs. 35.1% for Selective Search and 33.4% for DPM).

R-CNN pros and cons

• **Pros**
  • Accurate!
  • Any deep architecture can immediately be “plugged in”

• **Cons**
  • Ad hoc training objectives
    • Fine-tune network with softmax classifier (log loss)
    • Train post-hoc linear SVMs (hinge loss)
    • Train post-hoc bounding-box regressions (least squares)
  • Training is slow (84h), takes a lot of disk space
    • 2000 convnet passes per image
  • Inference (detection) is slow (47s / image with VGG16)
Fast R-CNN

Forward whole image through ConvNet

“conv5” feature map of image

“RoI Pooling” layer

Fully-connected layers

Bounding-box regressors

Linear

Linear + softmax

Softmax classifier

Region proposals

FCs

Source: R. Girshick

R. Girshick, Fast R-CNN, ICCV 2015
Fast R-CNN training

Log loss + smooth L1 loss

Multi-task loss

Linear + softmax

Linear

FCs

ConvNet

Trainable

Source: R. Girshick

R. Girshick, Fast R-CNN, ICCV 2015
# Fast R-CNN results

<table>
<thead>
<tr>
<th></th>
<th>Fast R-CNN</th>
<th>R-CNN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train time (h)</td>
<td>9.5</td>
<td>84</td>
</tr>
<tr>
<td>- Speedup</td>
<td>8.8x</td>
<td>1x</td>
</tr>
<tr>
<td>Test time / image</td>
<td>0.32s</td>
<td>47.0s</td>
</tr>
<tr>
<td>Test speedup</td>
<td>146x</td>
<td>1x</td>
</tr>
<tr>
<td>mAP</td>
<td>66.9%</td>
<td>66.0%</td>
</tr>
</tbody>
</table>

Timings exclude object proposal time, which is equal for all methods. All methods use VGG16 from Simonyan and Zisserman.

Source: R. Girshick
Faster R-CNN

S. Ren, K. He, R. Girshick, and J. Sun,
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015
Region proposal network

- Slide a small window over the conv5 layer
  - Predict object/no object
  - Regress bounding box coordinates
  - Box regression is with reference to anchors (3 scales x 3 aspect ratios)
Faster R-CNN results

detection mAP on PASCAL VOC 2007, with VGG-16 pre-trained on ImageNet

<table>
<thead>
<tr>
<th>system</th>
<th>time</th>
<th>07 data</th>
<th>07+12 data</th>
</tr>
</thead>
<tbody>
<tr>
<td>R-CNN</td>
<td>~50s</td>
<td>66.0</td>
<td>-</td>
</tr>
<tr>
<td>Fast R-CNN</td>
<td>~2s</td>
<td>66.9</td>
<td>70.0</td>
</tr>
<tr>
<td>Faster R-CNN</td>
<td>198ms</td>
<td>69.9</td>
<td>73.2</td>
</tr>
</tbody>
</table>
Object detection progress

Before deep convnets

Using deep convnets

Faster R-CNN

Fast R-CNN

R-CNNv1
Next trends

- **New datasets: MSCOCO**
  - 80 categories instead of PASCAL’s 20
  - Current best mAP: 37%

What is Microsoft COCO?

Microsoft COCO is a new image recognition, segmentation, and captioning dataset. Microsoft COCO has several features:

- Object segmentation
- Recognition in Context
- Multiple objects per image
- More than 300,000 images
- More than 2 Million instances
- 80 object categories
- 5 captions per image

http://mscoco.org/home/
Next trends

- Fully convolutional detection networks

Next trends

- Networks with context

S. Bell, L. Zitnick, K. Bala, and R. Girshick,
Review: Object detection with CNNs
Review: R-CNN

R. Girshick, J. Donahue, T. Darrell, and J. Malik,
Review: Fast R-CNN

- **ConvNet**: Forward whole image through ConvNet
- **“conv5” feature map of image**: "RoI Pooling" layer
- **Bounding-box regressors**: Linear + softmax
- **Region proposals**: Fully-connected layers
- **Softmax classifier**: Linear

Review: Faster R-CNN

S. Ren, K. He, R. Girshick, and J. Sun,
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015