From image classification to object detection

**Image classification**

**Object detection**
Real-time Multi-Person 2D Pose Estimation Using Part Affinity Fields

Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh
Carnegie Mellon University
What are the challenges of object detection?

• Images may contain more than one class, multiple instances from the same class
• Bounding box localization
• Evaluation
Outline

• Task definition and evaluation
• Conceptual approaches to detection
• Zoo of deep detection approaches
  • R-CNN
  • Fast R-CNN
  • Faster R-CNN
  • Yolo
  • SSD
Object detection evaluation

- At test time, predict bounding boxes, class labels, and confidence scores
- For each detection, determine whether it is a true or false positive
  - PASCAL criterion: $\frac{\text{Area}(\text{GT} \cap \text{Det})}{\text{Area}(\text{GT} \cup \text{Det})} > 0.5$
  - For multiple detections of the same ground truth box, only one considered a true positive
Object detection evaluation

- At test time, predict bounding boxes, class labels, and confidence scores
- For each detection, determine whether it is a true or false positive
- For each class, plot Recall-Precision curve and compute Average Precision (area under the curve)
- Take mean of AP over classes to get mAP

**Precision:**
true positive detections / total detections

**Recall:**
true positive detections / total positive test instances
Precision and Recall
Object detection evaluation

- At test time, predict bounding boxes, class labels, and confidence scores
- For each detection, determine whether it is a true or false positive
- For each class, plot Recall-Precision curve and compute Average Precision (area under the curve)
- Take mean of AP over classes to get mAP

Precision:
true positive detections / total detections

Recall:
true positive detections / total positive test instances
PASCAL VOC Challenge (2005-2012)

- 20 challenge classes:
  - *Person*
  - *Animals*: bird, cat, cow, dog, horse, sheep
  - *Vehicles*: aeroplane, bicycle, boat, bus, car, motorbike, train
  - *Indoor*: bottle, chair, dining table, potted plant, sofa, tv/monitor

- Dataset size (by 2012): 11.5K training/validation images, 27K bounding boxes, 7K segmentations

http://host.robots.ox.ac.uk/pascal/VOC/
Progress on PASCAL detection

Before CNNs

After CNNs

PASCAL VOC
Newer benchmark: COCO

What is COCO?

COCO is a large-scale object detection, segmentation, and captioning dataset. COCO has several features:

- Object segmentation
- Recognition in context
- Superpixel stuff segmentation
- 330K images (>200K labeled)
- 1.5 million object instances
- 80 object categories
- 91 stuff categories
- 5 captions per image
- 250,000 people with keypoints

http://cocodataset.org/#home
COCO dataset: Tasks

- image classification
- object detection
- semantic segmentation
- instance segmentation

- Also: keypoint prediction, captioning, question answering…
COCO detection metrics

**Average Precision (AP):**
- \( \text{AP} \) % AP at IoU=.50:.05:.95 (primary challenge metric)
- \( \text{AP}_{\text{IoU}=0.50} \) % AP at IoU=.50 (PASCAL VOC metric)
- \( \text{AP}_{\text{IoU}=0.75} \) % AP at IoU=.75 (strict metric)

**AP Across Scales:**
- \( \text{AP}_{\text{small}} \) % AP for small objects: area < 32²
- \( \text{AP}_{\text{medium}} \) % AP for medium objects: 32² < area < 96²
- \( \text{AP}_{\text{large}} \) % AP for large objects: area > 96²

**Average Recall (AR):**
- \( \text{AR}_{\text{max}=1} \) % AR given 1 detection per image
- \( \text{AR}_{\text{max}=10} \) % AR given 10 detections per image
- \( \text{AR}_{\text{max}=100} \) % AR given 100 detections per image

**AR Across Scales:**
- \( \text{AR}_{\text{small}} \) % AR for small objects: area < 32²
- \( \text{AR}_{\text{medium}} \) % AR for medium objects: 32² < area < 96²
- \( \text{AR}_{\text{large}} \) % AR for large objects: area > 96²

- Leaderboard: [http://cocodataset.org/#detection-leaderboard](http://cocodataset.org/#detection-leaderboard)
- Official COCO challenges no longer include detection
  - Emphasis has shifted to instance segmentation and dense semantic segmentation
Conceptual approach: Sliding window detection

- Slide a window across the image and evaluate a detection model at each location
  - Thousands of windows to evaluate: efficiency and low false positive rates are essential
  - Difficult to extend to a large range of scales, aspect ratios
Conceptual approach: Proposal-driven detection

- Generate and evaluate a few hundred region proposals
  - Proposal mechanism can take advantage of low-level perceptual organization cues
  - Proposal mechanism can be category-specific or category-independent, hand-crafted or trained
  - Classifier can be slower but more powerful
Selective search for detection

- 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
Selective search for detection

Evaluation of region proposals

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013
Classical detector with selective search

- 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 (incl. integral images) 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-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 Deformable Part Models)
R-CNN pros and cons

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

• **Cons**
  • Not a single end-to-end system
    • 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 CNN passes per image
  • Inference (detection) is slow (47s / image with VGG16)
Fast R-CNN

Fast R-CNN

Source: R. Girshick

R. Girshick, Fast R-CNN, ICCV 2015
RoI pooling

- “Crop and resample” a fixed-size feature representing a region of interest out of the outputs of the last conv layer
- Use nearest-neighbor interpolation of coordinates, max pooling

Source: R. Girshick, K. He
RoI pooling illustration
Prediction

- For each RoI, network predicts probabilities for C+1 classes (class 0 is background) and four bounding box offsets for C classes

Fast R-CNN training

Log loss + smooth L1 loss

Linear + softmax

Linear

FCs

Multi-task loss

Trainable

ConvNet

Source: R. Girshick

R. Girshick, Fast R-CNN, ICCV 2015
<table>
<thead>
<tr>
<th>Class</th>
<th>Softmax Loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cat</td>
<td>0.1</td>
</tr>
<tr>
<td>Dog</td>
<td>0.5</td>
</tr>
<tr>
<td>Car</td>
<td>0.2</td>
</tr>
<tr>
<td>Plane</td>
<td>0.1</td>
</tr>
<tr>
<td>Chair</td>
<td>0.1</td>
</tr>
</tbody>
</table>

If Dog is the answer, 
- \( \log(0.5) = 0.30 \) is the loss

\[
L(p, u, t^u, v) = L_{cls}(p, u) + \lambda[u \geq 1]L_{loc}(t^u, v),
\]

in which \( L_{cls}(p, u) = -\log p_u \) is log loss for true class \( u \).

The second task loss, \( L_{loc} \), is defined over a tuple of true bounding-box regression targets for class \( u \), \( v = (v_x, v_y, v_w, v_h) \), and a predicted tuple \( t^u = (t^u_x, t^u_y, t^u_w, t^u_h) \), again for class \( u \). The Iverson bracket indicator function \([u \geq 1]\) evaluates to 1 when \( u \geq 1 \) and 0 otherwise. By convention the catch-all background class is labeled \( u = 0 \). For background RoIs there is no notion of a ground-truth...
Softmax Loss

If Car is the answer - \( \log(0.2) = 0.7 \) is the loss

\[
L(p, u, t^u, v) = L_{\text{cls}}(p, u) + \lambda [u \geq 1] L_{\text{loc}}(t^u, v),
\]

in which \( L_{\text{cls}}(p, u) = -\log p_u \) is log loss for true class \( u \).

The second task loss, \( L_{\text{loc}} \), is defined over a tuple of true bounding-box regression targets for class \( u \), \( t^u = (v_x, v_y, v_w, v_h) \), and a predicted tuple \( t'^u = (t'^x, t'^y, t'^w, t'^h) \), again for class \( u \). The Iverson bracket indicator function \([u \geq 1]\) evaluates to 1 when \( u \geq 1 \) and 0 otherwise. By convention the catch-all background class is labeled \( u = 0 \). For background RoIs there is no notion of a ground-truth
Bounding box regression

Ground truth box

Target offset to predict*

Loss

Predicted offset

Region proposal (a.k.a default box, prior, reference, anchor)

Predicted box

*Typically in transformed, normalized coordinates
Multi-task loss

- Regression loss: *smooth L1 loss* on top of log space offsets relative to proposal

\[
\text{smooth}_{L_1}(x) = \begin{cases} 
0.5x^2 & \text{if } |x| < 1 \\
|x| - 0.5 & \text{otherwise}
\end{cases}
\]
ROI pooling: Backpropagation

- Similar to max pooling, has to take into account overlap of pooling regions.
ROI pooling: Backpropagation

- Similar to max pooling, has to take into account overlap of pooling regions

\[ \frac{\partial e}{\partial x_i} = \sum_r \sum_j \frac{\partial e}{\partial z_{rj}} \frac{\partial z_{rj}}{\partial x_i} = \sum_r \sum_j \mathbb{I}[i = i^*(r,j)] \frac{\partial e}{\partial z_{rj}} \]

Over regions, RoI indices

1 if “pooled” input; 0 o/w

Source: Ross Girshick
Mini-batch sampling

• Sample a few images (e.g., 2)
• Sample many regions from each image (64)

Source: R. Girshick, K. He
Fast R-CNN training

Log loss + smooth L1 loss

Linear + softmax

Linear

FCs

ConvNet

Multi-task loss

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.

( vs. 53.7% for AlexNet)

Source: R. Girshick
Faster R-CNN

Region proposal network (RPN)

- Slide a small window (3x3) over the conv5 layer
  - Predict object/no object
  - Regress bounding box coordinates with reference to anchors (3 scales x 3 aspect ratios)
One network, four losses

Source: R. Girshick, K. He
Figure 3: **Left:** Region Proposal Network (RPN). **Right:** Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.
Figure 3: **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.
## Faster R-CNN results

<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>

detection mAP on PASCAL VOC 2007, with VGG-16 pre-trained on ImageNet
Object detection progress
Streamlined detection architectures

- The Faster R-CNN pipeline separates proposal generation and region classification:

  - Is it possible do detection in one shot?

  - Conv feature map of the entire image
    - RPN
    - Region Proposals
    - RoI pooling
    - RoI features
    - Classification + Regression
    - Detections

  - Conv feature map of the entire image
    - Classification + Regression
    - Detections
What is strange in faster rcnn?
YOLO

- Divide the image into a coarse grid and directly predict class label and a few candidate boxes for each grid cell

YOLO

1. Take conv feature maps at 7x7 resolution
2. Add two FC layers to predict, at each location, a score for each class and 2 bboxes w/ confidences
   • For PASCAL, output is 7x7x30 (30 = 20 + 2*(4+1))

YOLO

- Objective function:

\[
\begin{align*}
\lambda_{\text{coord}} & \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left[ (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
& + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left[ (\sqrt{w_i} - \sqrt{\hat{w}_i})^2 + (\sqrt{h_i} - \sqrt{\hat{h}_i})^2 \right] \\
& + \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left( C_i - \hat{C}_i \right)^2 \\
& + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{noobj}} \left( C_i - \hat{C}_i \right)^2 \\
& + \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left( p_i(c) - \hat{p}_i(c) \right)^2
\end{align*}
\]
YOLO

- **Objective function:**

\[
\lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left[ (x_i - \hat{x}_i)^2 + (y_i - \hat{y}_i)^2 \right] \\
+ \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left[ \left( \sqrt{w_i} - \sqrt{\hat{w}_i} \right)^2 + \left( \sqrt{h_i} - \sqrt{\hat{h}_i} \right)^2 \right] \\
+ \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{obj}} \left( C_i - \hat{C}_i \right)^2 \\
+ \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathbb{1}_{i,j}^{\text{noobj}} \left( C_i - \hat{C}_i \right)^2 \\
+ \sum_{i=0}^{S^2} \mathbb{1}_{i}^{\text{obj}} \sum_{c \in \text{classes}} \left( p_i(c) - \hat{p}_i(c) \right)^2
\]

1. Cell \( i \) contains object, predictor \( j \) is responsible for it.
2. Small deviations matter less for larger boxes than for smaller boxes.
3. Confidence for object.
4. Confidence for no object.
5. Down-weight loss from boxes that don’t contain objects.
6. Class probability.
YOLO: Results

- Each grid cell predicts only two boxes and can only have one class – this limits the number of nearby objects that can be predicted.
- Localization accuracy suffers compared to Fast(er) R-CNN due to coarser features, errors on small boxes.
- 7x speedup over Faster R-CNN (45-155 FPS vs. 7-18 FPS).

Performance on PASCAL 2007
• Similarly to YOLO, predict bounding boxes directly from conv maps
• Unlike YOLO, do not use FC layers and predict different size boxes from conv maps at different resolutions
• Similarly to RPN, use anchors

SSD: Results (PASCAL 2007)

- More accurate and faster than YOLO and Faster R-CNN

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP</th>
<th>FPS</th>
<th>batch size</th>
<th># Boxes</th>
<th>Input resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN (VGG16)</td>
<td>73.2</td>
<td>7</td>
<td>1</td>
<td>~6000</td>
<td>~1000 × 600</td>
</tr>
<tr>
<td>Fast YOLO</td>
<td>52.7</td>
<td>155</td>
<td>1</td>
<td>98</td>
<td>448 × 448</td>
</tr>
<tr>
<td>YOLO (VGG16)</td>
<td>66.4</td>
<td>21</td>
<td>1</td>
<td>98</td>
<td>448 × 448</td>
</tr>
<tr>
<td>SSD300</td>
<td>74.3</td>
<td>46</td>
<td>1</td>
<td>8732</td>
<td>300 × 300</td>
</tr>
<tr>
<td>SSD512</td>
<td>76.8</td>
<td>19</td>
<td>1</td>
<td>24564</td>
<td>512 × 512</td>
</tr>
<tr>
<td>SSD300</td>
<td>74.3</td>
<td>59</td>
<td>8</td>
<td>8732</td>
<td>300 × 300</td>
</tr>
<tr>
<td>SSD512</td>
<td>76.8</td>
<td>22</td>
<td>8</td>
<td>24564</td>
<td>512 × 512</td>
</tr>
</tbody>
</table>
YOLO v2

- Remove FC layer, do convolutional prediction with anchor boxes instead
- Increase resolution of input images and conv feature maps
- Improve accuracy using batch normalization and other tricks

VOC 2007 results

YouTube demo

J. Redmon and A. Farhadi, YOLO9000: Better, Faster, Stronger, CVPR 2017
Multi-resolution prediction

- SSD predicts boxes of different size from different conv maps, but each level of resolution has its own predictors and higher-level context does not get propagated back to lower-level feature maps.
- Can we have a more elegant multi-resolution prediction architecture?
Feature pyramid networks

- Improve predictive power of lower-level feature maps by adding contextual information from higher-level feature maps.
- Predict different sizes of bounding boxes from different levels of the pyramid (but share parameters of predictors).

Feature pyramid networks or U-Net
RetinaNet

- Combine feature pyramid network with focal loss to reduce the standard cross-entropy loss for well-classified examples

RetinaNet

- Combine feature pyramid network with *focal loss* to reduce the standard cross-entropy loss for well-classified examples

\[
\begin{align*}
\text{CE}(p_t) &= -\log(p_t) \\
\text{FL}(p_t) &= -(1 - p_t)^\gamma \log(p_t)
\end{align*}
\]

RetinaNet: Results

Deconvolutional SSD

- Improve performance of SSD by increasing resolution through learned “deconvolutional” layers

YOLO v3

YOLOv3: An Incremental Improvement

Joseph Redmon, Ali Farhadi
University of Washington

Abstract

We present some updates to YOLO! We made a bunch of little design changes to make it better. We also trained this new network that’s pretty swell. It’s a little bigger than last time but more accurate. It’s still fast though, don’t worry. At $320 \times 320$ YOLOv3 runs in 22 ms at 28.2 mAP, as accurate as SSD but three times faster. When we look at the old .5 IOU mAP detection metric YOLOv3 is quite good. It achieves 57.9 AP$_{50}$ in 51 ms on a Titan X, compared to 57.5 AP$_{50}$ in 198 ms by RetinaNet, similar performance but 3.8× faster. As always, all the code is online at https://pjreddie.com/yolo/.

1. Introduction

RetinaNet: Results

Summary: Object detection with CNNs

- **R-CNN**: region proposals + CNN on cropped, resampled regions
- **Fast R-CNN**: region proposals + RoI pooling on top of a conv feature map
- **Faster R-CNN**: RPN + RoI pooling
- **Next generation of detectors**
  - Direct prediction of BB offsets, class scores on top of conv feature maps
  - Get better context by combining feature maps at multiple resolutions
Review: R-CNN

Review: Fast R-CNN

Review: Faster R-CNN