Deep Learning for Instance-level Object Understanding

CVPR 2017 Tutorial on Deep Learning for Objects and Scenes

Ross Girshick
in collaboration with
Kaiming He, Georgia Gkioxari, Tsung-Yi Lin (Cornell Tech),
Bharath Hariharan, Piotr Dollár

Facebook AI Research (FAIR)
... Now you’re Experts in Deep Representations

- This talk: object detection and instance-level object understanding
Object Detection

Image classification
(what? I don’t care where)

Object detection
(what + where?)
Object Detection circa 2007

Felzenszwalb, Ramanan, McAllester. A Discriminatively Trained, Multiscale, Deformable Part Model. CVPR 2008 (DPM v1)
Object Detection Today

He, Gkioxari, Dollár, Girshick. Mask R-CNN. In ICCV 2017
Instance-level Object Understanding Today

He, Gkioxari, Dollár, Girshick. Mask R-CNN. In ICCV 2017
Instance-level Object Understanding Today

He, Gkioxari, Dollár, Girshick. Mask R-CNN. In ICCV 2017
Instance-level Object Understanding Today

Gkioxari, Girshick, Dollár, He. Detecting and Recognizing Human-Object Interactions. arXiv 2017
Outline

Mask R-CNN [25 min]
- Task
- Model
- Training
- Inference

Deep Learning Methods for Object Detection [10 min brief survey]
- Landscape of methods
- Speed/accuracy tradeoffs
- How we got here (2007 vs. 2017)
- Some open directions in instance-level understanding
Task: Instance Segmentation

Object detection

Person 1
Person 2
Person 3
Person 4
Person 5

Image from PASCAL VOC
Task: Instance Segmentation

- Object detection
- Semantic segmentation

- Person pixels
  - Things become stuff
  - A faceless mass of person pixels

Image from PASCAL VOC
Task: Instance Segmentation

Image from PASCAL VOC

Object detection

Semantic segmentation

Person pixels

- Things become stuff
- A faceless mass of person pixels

Instance segmentation

Best of both tasks
Mask R-CNN: Model Overview

R-CNN-style detection system
1. Backbone architecture
2. Feature Pyramid Network (FPN)
3. Region Proposal Network (RPN)
4. Region of interest feature alignment (RoIAlign)
5. Multi-task network head
   - Box classifier
   - Box regressor
   - Mask predictor
   - Keypoint predictor

Modular composition of many recent ideas
0. **R-CNN-style Approach to Object Detection**

Input image

Object / **region proposals (RoIs)**
(external or internal to network)

Deep Learning-based **region classifier**

**General formula for Region-based CNN models**

Detection = Region classification + box regression + ...


Uijlings, van de Sande, Gevers, Smeulders. Selective Search for Object Recognition. IJCV 2013
1. Backbone ConvNet

Use any standard ConvNet as a “backbone architecture”

- AlexNet, VGG, ResNet, Inception, Inception-ResNet, ResNeXt (poster 67 Tuesday morning), ...

- Use “same” padding everywhere (preserves integer scales)

- Prefer fully convolutional networks (ignoring cls head; see next slide)

- Pre-train on ImageNet classification (or similar)
1. Prepare for Detection! (Minor Surgery)

Replace test-time batch normalization with fixed affine transform

- \( \text{BN}_{\text{test}}(x) = \gamma(x - \mu)/\sigma + \beta \), treat \( \gamma, \beta, \mu, \sigma \) as constants

Remove pre-trained classification head

- Backbone is now fully convolutional (takes any input size)
2. Scale Invariant Detection

Popular strategies for detecting objects over large scale changes:

- SLOW! (Viola & Jones, HOG, DPM, multiscale Fast R-CNN, ...)
- Do nothing; give up; Fast (Fast R-CNN, YOLO, ...)
- Fast! (≈ SSD, ...)

(a) Featurized image pyramid
(b) Single feature map
(c) Pyramidal feature hierarchy
2. Scale Invariant Detection

Let's examine this one

(a) Featurized image pyramid
SOW!!
(Viola & Jones, HOG, DPM, multiscale Fast R-CNN, ...)

(b) Single feature map
Do nothing; give up; Fast!
(Fast R-CNN, YOLO, ...)

(c) Pyramidal feature hierarchy
Fast!
(≈ SSD, ...)

Let's examine this one
Compromise Feature Quality, but Fast ("Free")

The "native" in-network feature pyramid poses an inherent tradeoff
2. ... + Feature Pyramid Network (FPN)

[Optional, but recommended]

Lin et al. Feature Pyramid Networks for Object Detection. In CVPR 2017. Poster 99 Saturday morning. See also: Shrivastava’s TDM
No Compromise on Feature Quality, still Fast

Lin et al. Feature Pyramid Networks for Object Detection. In CVPR 2017. Poster 99 Saturday morning. See also: Shrivastava’s TDM
3. ... + Region Proposal Network (RPN)

Proposals = sliding window object/not-object classifier + box regression inside the same network

4. ... + RoIAlign Transform (on each Proposal)

Smoothly normalize features and predictions into coordinate frame free of scale and aspect ratio
4. ... + RoIAlign Transform (on each Proposal)

Key: No coordinate quantization (cf. RoIPool in Fast R-CNN, etc.)

FPN feature level

Grid of bilinear interpolation points

Proposal from RPN

256-d

RoIAlign transform

Feature value is average of interpolated values on grid

(Fixed dimensional representation)
Compare to RoIPool, RoIWarp, and others

Quantization breaks pixel-to-pixel alignment

Original RoI

RoIPool coordinate quantization

Quantized RoI
5. ... + Task-specific Heads (on each Proposal)

Task specific heads for ...
- Bounding box detection
- Object classification
- Instance mask prediction
- Human keypoint prediction

Standard FPN-based Fast/er R-CNN head

RolAlign transformed features
5. ... + Task-specific Heads (on each Proposal)

Task specific heads for ...
- Object classification
- Bounding box regression
- Instance mask prediction
- Human keypoint prediction

Per-proposal FCN predicts instance masks
Mask R-CNN: Training

Not enough time for details (sorry!)

Same as “image centric” Fast/er R-CNN training

- Use precomputed proposals for faster experimentation
- Use joint / end-to-end training for sharing features

But with training targets for masks
Example Mask Training Targets

Image with training proposal | 28x28 mask target | Image with training proposal | 28x28 mask target
Mask R-CNN: Inference

1. Perform Faster R-CNN inference
   - Generate proposals (RPN)
   - Score the proposals
   - Regress from proposals to refined detection boxes
   - Apply NMS and take the top \( K \) (= 100, e.g.)

2. Run RoIAlign and mask head on top-\( K \) refined, post-NMS boxes
   - Fast (only compute masks for top-\( K \) detections)
   - Improves accuracy (uses \textit{refined} detection boxes, not proposals)
Mask Prediction

Validation image with box detection shown in red

28x28 soft prediction from Mask R-CNN (enlarged)

Soft prediction resampled to image coordinates (bilinear and bicubic interpolation work equally well)

Final prediction (threshold at 0.5)
Mask Prediction

Validation image with box detection shown in red
Mask Prediction

Validation image with box detection shown in red
Add keypoint head (28x28x17)

Predict one “mask” for each keypoint

Softmax over spatial locations (encodes one keypoint per mask “prior”)

17 keypoint “mask” predictions shown as heatmaps with OKS scores from argmax positions

(Not shown: Head architecture is slightly different for keypoints)
Deep Learning for Object Detection: A Bewildering Word Salad

R-CNN OverFeat DetectorNet
DeepMultibox SPP-net Fast R-CNN
MR-CNN SSD YOLO YOLOv2
G-CNN AttractioNet Mask R-CNN
R-FCN RPN FPN Faster R-CNN ...

and many more words
Let’s Organize the Landscape

A random landscape scene on Mt. Baker, just because I like mountains.

Photo credit: Ross Girshick
Common to all Methods

Start by modifying a classification network

Since R-CNN, this network is pre-trained, typically using ImageNet (cf. DetectorNet)

DetectorNet: Szegedy et al. Deep Neural Networks for Object Detection. NIPS 2013
Highest Information Gain Split: “Stage” Count

More than one stage
- DetectorNet (Szegedy et al.)
- R-CNN (Girshick et al.)
- SPP-net (He et al.)
- Fast R-CNN (Girshick)
- Faster R-CNN (Ren et al.)
- R-FCN (Dai et al.)
- Mask R-CNN (He et al.)

One stage
- OverFeat (Sermanet et al.)
- YOLO, YOLOv2 (Redmon et al.)
- SSD (Wei et al.)
- RetinaNet (Lin et al.) [Poster at WICV on Wed.]
Stages

Detection Output space: \( N = H \times W \) pixel image has \( O(N^2) \) windows

Output space is HUGE, even for small images

Classic approaches to dealing with this issue ...

- Sliding window with subsampled aspect ratios, translation, and scale (reduces window count to \( \approx 100,000 \))
- Multiple stages of cascaded classification (helps deal with extreme foreground-vs-background class imbalance)
More than one “stage” (≈ proposal based; but doesn’t require proposals)
More than one “stage” (≈ proposal based; but doesn’t require proposals)

Cascade-like reduction in output space
More than one “stage” (≈ proposal based; but doesn’t require proposals)

Classification of reduced output space elements

Cascade-like reduction in output space

Input image

Object / region proposals

Deep Learning region classifier

Region classification, box regression
More than one “stage” (≈ proposal based; but doesn’t require proposals)

Cascade-like reduction in output space

Classification of reduced output space elements

One stage

More than one “stage” ($\approx$ proposal based; but doesn’t require proposals)

Cascade-like reduction in output space

Direct classification of all output space elements

Classification of reduced output space elements


“You only look once”
“Single shot”
Thinking about Multi-stage Detectors

Networks on Convolutional Feature Maps

Shifting computation between the “head” and the “trunk”
Speed / Accuracy Tradeoffs

Speed is mainly a function of:

- Input image resolution
- Network complexity
- Number of proposals (if applicable)

Optimizing AP on PASCAL VOC has lead to confusion about speed / accuracy tradeoffs

COCO provides a clear picture of tradeoffs
Huang et al. Speed/Accuracy Tradeoffs for Modern Convolutional Object Detectors. CVPR 2017

YOLOv2
(different impl.; not 100% comparable)
Past vs. Present

How did we get here?

PASCAL VOC 2007 **Object Detection** mAP (%)

- **HOG, DPM v1**: 2007 - 21 (shallow)
- **HOG, DPM v5**: 2012 - 34 (shallow)
- **AlexNet (R-CNN)**: 2013 - 58 (8 layers)
- **VGG (R-CNN)**: 2014 - 66 (16 layers)
- **ResNet (Faster R-CNN)***: 2015 - 86 (101 layers)

*Already a 2x AP improvement!*

*[^other improvements & more data


Slide credit: Kaiming He
False in 2012 and earlier, but True since 2013

State-of-the-art detectors

- Improve with more data
- Improve with increased model capacity
- Improve from transfer learning
- Immediately benefit from image classification research
- Share a common modeling framework with speech and NLP
Conclusions. Thank you! Questions?

*Object detection has come a very long way in a short time!*

We’re moving from box detection to *instance-level understanding*

**Major challenges remain:**

- Long-tailed examples-per-category distribution (implies *low-shot learning*)
- Open vocabulary recognition (implies *benchmarking challenges*)
- Low accuracy with very heavy occlusion / clutter (esp. for keypoints; “*reasoning*”)  
- Bottleneck may still be the raw engine of recognition (improve *backbone, data*)
In-network region proposals from RPN

Shared region-wise subnetwork

CNN applied to entire image

Class/box

RoIPool op
In-network region proposals from RPN

Position-sensitive RoIPool op

CNN applied to entire image