Mask R-CNN

ICCV 2017, Venice, Italy

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick
Facebook AI Research (FAIR)
Visual Perception

Object Detection ✓
Semantic Segmentation ✓
Instance Segmentation ?
A Challenging Problem...

# entries on COCO

Object Det.  31
Instance Seg.  5

# entries on Cityscapes

Semantic Seg.  58
Instance Seg.  11

*on the leaderboards
Object Detection

- Fast/Faster R-CNN
  - Meta-algorithm
  - Good speed
  - Good accuracy
  - Intuitive
  - Easy to use

Semantic Segmentation

- Fully Convolutional Net (FCN)
  - Meta-algorithm
  - Good speed
  - Good accuracy
  - Intuitive
  - Easy to use


Figure credit: Long et al
Instance Segmentation

**Goals** of Mask R-CNN
- Meta-algorithm
- Good speed
- Good accuracy
- Intuitive
- Easy to use
Instance Segmentation Methods

R-CNN driven

FCN driven

Person 1
Person 2
Person 3
Person 4
Person 5

[Hariharan et al, ECCV’14], [Hariharan et al, CVPR’15], [Dai et al, CVPR’15], [Dai et al, CVPR’16], ...

[Li et al, CVPR’17],
[Arnab & Torr, CVPR’17], ...

[Liang et al, arXiv’15], [Kirillov et al, CVPR’17],
[Bai & Urtasun, CVPR’17], ...
What is Mask R-CNN

- Mask R-CNN = **Faster R-CNN** with **FCN** on RoIs
What is Mask R-CNN: Parallel Heads

- Easy, fast to implement and use

(slow) R-CNN

Fast/er R-CNN

Mask R-CNN
What is Mask R-CNN: RoIAlign

• No quantization

Variable size RoI

Bilinear interpolation

Fixed dimensional RoI output

Feat. map
vs. RoIPool

- was not for segmentation
- breaks pixel-to-pixel alignment
What is Mask R-CNN: FCN Mask Head

- Pixel-to-pixel aligned
What is Mask R-CNN: FCN Mask Head

- Pixel-to-pixel aligned

RoI

28x28 FCN prediction

resized soft prediction

final mask
Implementation

• Mask R-CNN is a **meta-algorithm**

• Compatible with other improvements

• We used:
  • ResNet/ResNeXt [Xie et al, CVPR’17]
  • Feature Pyramid Net [Lin et al, CVPR’17]
Results
Instance Segmentation Results on COCO

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP</th>
<th>AP$_{50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNC [7]</td>
<td>ResNet-101-C4</td>
<td>24.6</td>
<td>44.3</td>
</tr>
<tr>
<td><strong>FCIS+++ [20] +OHEM</strong></td>
<td>ResNet-101-C5-dilated</td>
<td><strong>33.6</strong></td>
<td><strong>54.5</strong></td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-C4</td>
<td>33.1</td>
<td>54.9</td>
</tr>
<tr>
<td><strong>Mask R-CNN</strong></td>
<td>ResNet-101-FPN</td>
<td><strong>35.7</strong></td>
<td><strong>58.0</strong></td>
</tr>
</tbody>
</table>

- without bells and whistles, **2 AP better** than 2016 winner
- **200ms / img**
### Instance Segmentation Results on COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>AP</th>
<th>AP$_{50}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNC [7]</td>
<td>ResNet-101-C4</td>
<td>24.6</td>
<td>44.3</td>
</tr>
<tr>
<td>FCIS+++ [20] +OHEM</td>
<td>ResNet-101-C5-dilated</td>
<td>33.6</td>
<td>54.5</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-C4</td>
<td>33.1</td>
<td>54.9</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-FPN</td>
<td>35.7</td>
<td>58.0</td>
</tr>
<tr>
<td><strong>Mask R-CNN</strong></td>
<td>ResNeXt-101-FPN</td>
<td><strong>37.1</strong></td>
<td><strong>60.0</strong></td>
</tr>
</tbody>
</table>

- Better features: ResNeXt [Xie et al, CVPR’17]
### Object Detection Results on COCO

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>$\text{AP}^{\text{bb}}$</th>
<th>$\text{AP}^{\text{bb}_{50}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN+++ [15]</td>
<td>ResNet-101-C4</td>
<td>34.9</td>
<td>55.7</td>
</tr>
<tr>
<td>Faster R-CNN w FPN [22]</td>
<td>ResNet-101-FPN</td>
<td>36.2</td>
<td>59.1</td>
</tr>
<tr>
<td>Faster R-CNN by G-RMI [17]</td>
<td>Inception-ResNet-v2 [32]</td>
<td>34.7</td>
<td>55.5</td>
</tr>
<tr>
<td>Faster R-CNN w TDM [31]</td>
<td>Inception-ResNet-v2-TDM</td>
<td>36.8</td>
<td>57.7</td>
</tr>
<tr>
<td>Faster R-CNN, RoIAlign</td>
<td>ResNet-101-FPN</td>
<td>37.3</td>
<td>59.6</td>
</tr>
</tbody>
</table>

**bbox improved by:**
- RoIAlign
### Object Detection Results on COCO

<table>
<thead>
<tr>
<th>Method</th>
<th>Backbone</th>
<th>AP&lt;sub&gt;bb&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;bb&lt;/sub&gt; 50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN+++ [15]</td>
<td>ResNet-101-C4</td>
<td>34.9</td>
<td>55.7</td>
</tr>
<tr>
<td>Faster R-CNN w FPN [22]</td>
<td>ResNet-101-FPN</td>
<td>36.2</td>
<td>59.1</td>
</tr>
<tr>
<td>Faster R-CNN by G-RMI [17]</td>
<td>Inception-ResNet-v2 [32]</td>
<td>34.7</td>
<td>55.5</td>
</tr>
<tr>
<td>Faster R-CNN w TDM [31]</td>
<td>Inception-ResNet-v2-TDM</td>
<td>36.8</td>
<td>57.7</td>
</tr>
<tr>
<td>Faster R-CNN, RoIAlign</td>
<td>ResNet-101-FPN</td>
<td>37.3</td>
<td>59.6</td>
</tr>
<tr>
<td><strong>Mask R-CNN</strong></td>
<td>ResNet-101-FPN</td>
<td>38.2</td>
<td>60.3</td>
</tr>
</tbody>
</table>

bbox improved by:
- RoIAlign
- Multi-task training w/ mask
Object Detection Results on COCO

<table>
<thead>
<tr>
<th>Model</th>
<th>Backbone</th>
<th>AP&lt;sub&gt;bb&lt;/sub&gt;</th>
<th>AP&lt;sub&gt;bb50&lt;/sub&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN+++ [15]</td>
<td>ResNet-101-C4</td>
<td>34.9</td>
<td>55.7</td>
</tr>
<tr>
<td>Faster R-CNN w FPN [22]</td>
<td>ResNet-101-FPN</td>
<td>36.2</td>
<td>59.1</td>
</tr>
<tr>
<td>Faster R-CNN by G-RMI [17]</td>
<td>Inception-ResNet-v2 [32]</td>
<td>34.7</td>
<td>55.5</td>
</tr>
<tr>
<td>Faster R-CNN w TDM [31]</td>
<td>Inception-ResNet-v2-TDM</td>
<td>36.8</td>
<td>57.7</td>
</tr>
<tr>
<td>Faster R-CNN, RoIAlign</td>
<td>ResNet-101-FPN</td>
<td>37.3</td>
<td>59.6</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-FPN</td>
<td>38.2</td>
<td>60.3</td>
</tr>
<tr>
<td><strong>Mask R-CNN</strong></td>
<td>ResNeXt-101-FPN</td>
<td><strong>39.8</strong></td>
<td><strong>62.3</strong></td>
</tr>
</tbody>
</table>

bbox improved by:
- RoIAlign
- Multi-task training w/ mask
COCO Competition 2017

- Mask R-CNN is used by leading teams

- Our Mask R-CNN achieves a *single-model* result of
  - 47.9 bbox AP, 42.6 mask AP

- More in our talk in COCO workshop (10/29, Sunday)
Examples
surrounded by same-category objects

Mask R-CNN results on COCO
Mask R-CNN results on COCO

disconnected objects
Mask R-CNN results on COCO

small objects
Failure: detection/segmentation

Mask R-CNN results on COCO
Failure: recognition

Mask R-CNN results on COCO

not a kite
For Human Keypoint Detection

- keypoint = 1-hot mask
- human pose = 17 masks

- One framework for
  - bbox
  - mask
  - keypoint
Conclusion

• Mask R-CNN
  ✓ Meta-algorithm
  ✓ Good speed
  ✓ Good accuracy
  ✓ Intuitive
  ✓ Easy to use

Code will be open-sourced as Facebook AI Research’s Detectron platform