Mask R-CNN: A Perspective on Equivariance

ICCV 2017 Tutorial, Venice, Italy

Kaiming He
in collaboration with: Georgia Gkioxari, Piotr Dollár, and Ross Girshick
Facebook AI Research (FAIR)
Introduction
Visual Perception Problems

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

# entries on COCO leaderboard

Object Det. 31  
Instance Seg. 5

# entries on Cityscapes leaderboard

Semantic Seg. 58  
Instance Seg. 11
Object Detection

• Fast/Faster R-CNN
  ✓ Good speed
  ✓ Good accuracy
  ✓ Intuitive
  ✓ Easy to use

Semantic Segmentation

• Fully Convolutional Net (FCN)
  ✓ Good speed
  ✓ Good accuracy
  ✓ Intuitive
  ✓ Easy to use

Instance Segmentation

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

R-CNN driven

FCN driven

(proposals)

(R-CNN driven)

(FCN driven)

Person 1
Person 2
Person 3
Person 4
Person 5

Person 1
Person 2
Person 3
Person 4
Person 5
Instance Segmentation Methods

RCNN-driven

- SDS [Hariharan et al, ECCV’14]
- HyperCol [Hariharan et al, CVPR’15]
- CFM [Dai et al, CVPR’15]
- MNC [Dai et al, CVPR’16]

FCN-driven

- PFN [Liang et al, arXiv’15]
- InstanceCut [Kirillov et al, CVPR’17]
- Watershed [Bai & Urtasun, CVPR’17]
- FCIS [Li et al, CVPR’17]
- DIN [Arnab & Torr, CVPR’17]
Mask R-CNN

• Mask R-CNN = **Faster R-CNN** with **FCN** on RoIs
Parallel Heads

• Easy, fast to implement and train

(slow) R-CNN

Fast/er R-CNN

Mask R-CNN
A Perspective on Equivariance
How can we draw the *Apple* logo?
How can we draw the Apple logo?

figure source: http://mymodernmet.com/famous-company-logos-memory/
How can we draw the *Apple* logo?

ground truth

figure source: http://mymodernmet.com/famous-company-logos-memory/
What is given?

<table>
<thead>
<tr>
<th>Memory + Blank Paper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ground truth seen</td>
</tr>
<tr>
<td>Ground truth reference on paper</td>
</tr>
</tbody>
</table>

What can be drawn?

<table>
<thead>
<tr>
<th>Apple, with a bite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Apple, with a bite on the right, a leaf on top</td>
</tr>
<tr>
<td>THE apple logo, pixel-to-pixel aligned</td>
</tr>
</tbody>
</table>
Invariance vs. Equivariance

see also “What is wrong with convolutional neural nets?”, Geoffrey Hinton, 2017
Invariance vs. Equivariance

- **Equivariance**: changes in input lead to corresponding changes in output

- *Classification* desires *invariant* representations: output a label

- *Instance Seg.* desires *equivariant* representations:
  - Translated object => translated mask
  - Scaled object => scaled mask
  - *Big and small* objects are equally important (due to AP metric)
    - unlike semantic seg. (counting pixels)
Invariance vs. Equivariance

• Convolutions are translation-\textit{equivariant}

• \textit{Fully-ConvNet} (FCN) is translation-\textit{equivariant}

• ConvNet becomes translation-\textit{invariant} due to fully-connected or global pool layers
Equivariance in Mask R-CNN

1. Fully-Conv Features:
equivariant to global (image) translation
Equivariance in Mask R-CNN

2. Fully-Conv on RoI: equivariant to translation within RoI
Fully-Conv on RoI

Translation of object in RoI => Same translation of mask in RoI
- Equivariant to small translation of RoIs
- More robust to RoI’s localization imperfection
Equivariance in Mask R-CNN

3. RoIAlign:
3a. maintain translation-equivariance before/after RoI
RoIAlign

FAQs: how to sample grid points within a cell?
- 4 regular points in 2x2 sub-cells
- other implementation could work

(conv feat. map)

Grid points of bilinear interpolation

RoIAlign output

(Variable size RoI)

(Fixed dimensional representation)
RoIAlign vs. RoIPool

- RoIPool *breaks* pixel-to-pixel translation-equivariance

See also "What is wrong with convolutional neural nets?", Geoffrey Hinton, 2017
Equivariance in Mask R-CNN

3. RoIAlign:
   3b. Scale-equivariant (and aspect-ratio-equivariant)
RoIAlign: Scale-Equivariance

- RoIAlign creates *scale-invariant* representations
- RoIAlign + "output pasted back" provides *scale-equivariance*
More about Scale-Equivariance: FPN

- RoIAlign is scale-invariant if on raw pixels:
  - = (slow) R-CNN: crops and warps RoIs
- RoIAlign is scale-invariant if on scale-invariant feature maps

- Feature Pyramid Network (FPN) [Lin et al. CVPR’17] creates approx. scale-invariant features
Equivariance in Mask R-CNN: Summary

• Translation-equivariant
  • FCN features
  • FCN mask head
  • RoIAlign (pixel-to-pixel behavior)

• Scale-equivariant (and aspect-ratio-equivariant)
  • RoIAlign (warping and normalization behavior) + paste-back
  • FPN features
Instance Seg: When we don’t want equivariance?

- A pixel $x$ could have a different label w.r.t. different RoIs
  - zero-padding in RoI boundary breaks equivariance
  - outside objects are suppressed
  - only equivariant to small changes of RoIs (which is desired)
Mask R-CNN results on COCO

object surrounded by same-category objects
Result Analysis
### Ablation: RoIPool vs. RoIAlign

**Baseline:** ResNet-50-Conv5 backbone, \textit{stride=32}

<table>
<thead>
<tr>
<th></th>
<th>mask AP</th>
<th></th>
<th></th>
<th>box AP</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP</td>
<td>$AP_{50}$</td>
<td>$AP_{75}$</td>
<td>$AP^{bb}$</td>
<td>$AP^{bb}_{50}$</td>
<td>$AP^{bb}_{75}$</td>
</tr>
<tr>
<td><strong>RoIPool</strong></td>
<td>23.6</td>
<td>46.5</td>
<td>21.6</td>
<td>28.2</td>
<td>52.7</td>
<td>26.9</td>
</tr>
<tr>
<td><strong>RoIAlign</strong></td>
<td>30.9</td>
<td>51.8</td>
<td>32.1</td>
<td>34.0</td>
<td>55.3</td>
<td>36.4</td>
</tr>
<tr>
<td></td>
<td>+7.3</td>
<td>+5.3</td>
<td>+10.5</td>
<td>+5.8</td>
<td>+2.6</td>
<td>+9.5</td>
</tr>
</tbody>
</table>

- huge gain at high IoU, in case of big stride (32)
Ablation: RoIPool vs. RoIAlign

baseline: ResNet-50-Conv5 backbone, stride=32

<table>
<thead>
<tr>
<th>Mask AP</th>
<th>Box AP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AP</td>
</tr>
<tr>
<td>RoIPool</td>
<td>23.6</td>
</tr>
<tr>
<td>RoIAlign</td>
<td>30.9</td>
</tr>
</tbody>
</table>

- nice box AP without dilation/upsampling
Ablation: Multinomial vs. Binary Masks

baseline: ResNet-50-Conv4 backbone, stride=16

<table>
<thead>
<tr>
<th></th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>softmax</td>
<td>24.8</td>
<td>44.1</td>
<td>25.1</td>
</tr>
<tr>
<td>sigmoid</td>
<td><strong>30.3</strong></td>
<td><strong>51.2</strong></td>
<td><strong>31.5</strong></td>
</tr>
</tbody>
</table>

• cls head: did recognition
  • mask head: no need to recognize again

"apple"
Ablation: MLP vs. FCN mask

**baseline: ResNet-50-FPN backbone**

<table>
<thead>
<tr>
<th>mask branch</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP</td>
<td>31.5</td>
<td>53.7</td>
<td>32.8</td>
</tr>
<tr>
<td>MLP</td>
<td>31.5</td>
<td>54.0</td>
<td>32.6</td>
</tr>
<tr>
<td>FCN</td>
<td><strong>33.6</strong></td>
<td><strong>55.2</strong></td>
<td><strong>35.3</strong></td>
</tr>
</tbody>
</table>

- **MLP**: lose “place-coded” info, too abstract
- **FCN**: translation-equivariant

• +2.1 point
# Instance Segmentation Results on COCO

<table>
<thead>
<tr>
<th></th>
<th>backbone</th>
<th>AP</th>
<th>AP$_{50}$</th>
<th>AP$_{75}$</th>
<th>AP$_S$</th>
<th>AP$_M$</th>
<th>AP$_L$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNC [7]</td>
<td>ResNet-101-C4</td>
<td>24.6</td>
<td>44.3</td>
<td>24.8</td>
<td>4.7</td>
<td>25.9</td>
<td>43.6</td>
</tr>
<tr>
<td>FCIS [20] +OHEM</td>
<td>ResNet-101-C5-dilated</td>
<td>29.2</td>
<td>49.5</td>
<td>-</td>
<td>7.1</td>
<td>31.3</td>
<td>50.0</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>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-C4</td>
<td>33.1</td>
<td>54.9</td>
<td>34.8</td>
<td>12.1</td>
<td>35.6</td>
<td>51.1</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>
<td><strong>37.8</strong></td>
<td>15.5</td>
<td>38.1</td>
<td>52.4</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNeXt-101-FPN</td>
<td><strong>37.1</strong></td>
<td><strong>60.0</strong></td>
<td><strong>39.4</strong></td>
<td><strong>16.9</strong></td>
<td><strong>39.9</strong></td>
<td><strong>53.5</strong></td>
</tr>
</tbody>
</table>

- 2 AP better than SOTA w/ R101, without bells and whistles
- 200ms / img
## Instance Segmentation Results on COCO

<table>
<thead>
<tr>
<th></th>
<th>backbone</th>
<th>AP</th>
<th>AP\textsubscript{50}</th>
<th>AP\textsubscript{75}</th>
<th>AP\textsubscript{S}</th>
<th>AP\textsubscript{M}</th>
<th>AP\textsubscript{L}</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNC [7]</td>
<td>ResNet-101-C4</td>
<td>24.6</td>
<td>44.3</td>
<td>24.8</td>
<td>4.7</td>
<td>25.9</td>
<td>43.6</td>
</tr>
<tr>
<td>FCIS [20] +OHEM</td>
<td>ResNet-101-C5-dilated</td>
<td>29.2</td>
<td>49.5</td>
<td>-</td>
<td>7.1</td>
<td>31.3</td>
<td>50.0</td>
</tr>
<tr>
<td>FCIS++ [20] +OHEM</td>
<td>ResNet-101-C5-dilated</td>
<td>33.6</td>
<td>54.5</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-C4</td>
<td>33.1</td>
<td>54.9</td>
<td>34.8</td>
<td>12.1</td>
<td>35.6</td>
<td>51.1</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-FPN</td>
<td>35.7</td>
<td>58.0</td>
<td>37.8</td>
<td>15.5</td>
<td>38.1</td>
<td>52.4</td>
</tr>
<tr>
<td><strong>Mask R-CNN</strong></td>
<td><strong>ResNeXt-101-FPN</strong></td>
<td><strong>37.1</strong></td>
<td><strong>60.0</strong></td>
<td><strong>39.4</strong></td>
<td><strong>16.9</strong></td>
<td><strong>39.9</strong></td>
<td><strong>53.5</strong></td>
</tr>
</tbody>
</table>

- benefit from better features (ResNeXt [Xie et al. CVPR’17])
## Object Detection Results on COCO

<table>
<thead>
<tr>
<th></th>
<th>backbone</th>
<th>$\text{AP}_{\text{bb}}$</th>
<th>$\text{AP}_{\text{bb}}^{50}$</th>
<th>$\text{AP}_{\text{bb}}^{75}$</th>
<th>$\text{AP}_{\text{bb}}^{S}$</th>
<th>$\text{AP}_{\text{bb}}^{M}$</th>
<th>$\text{AP}_{\text{bb}}^{L}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Faster R-CNN+++ [15]</td>
<td>ResNet-101-C4</td>
<td>34.9</td>
<td>55.7</td>
<td>37.4</td>
<td>15.6</td>
<td>38.7</td>
<td>50.9</td>
</tr>
<tr>
<td>Faster R-CNN w FPN [22]</td>
<td>ResNet-101-FPN</td>
<td>36.2</td>
<td>59.1</td>
<td>39.0</td>
<td>18.2</td>
<td>39.0</td>
<td>48.2</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>
<td>36.7</td>
<td>13.5</td>
<td>38.1</td>
<td>52.0</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>
<td>39.2</td>
<td>16.2</td>
<td>39.8</td>
<td>52.1</td>
</tr>
<tr>
<td>Faster R-CNN, RoIAlign</td>
<td>ResNet-101-FPN</td>
<td>37.3</td>
<td>59.6</td>
<td>40.3</td>
<td>19.8</td>
<td>40.2</td>
<td>48.8</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-FPN</td>
<td>38.2</td>
<td>60.3</td>
<td>41.7</td>
<td>20.1</td>
<td>41.1</td>
<td>50.2</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNeXt-101-FPN</td>
<td><strong>39.8</strong></td>
<td><strong>62.3</strong></td>
<td><strong>43.4</strong></td>
<td><strong>22.1</strong></td>
<td><strong>43.2</strong></td>
<td><strong>51.2</strong></td>
</tr>
</tbody>
</table>

bbox detection improved by:
- RoIAlign
Object Detection Results on COCO

<table>
<thead>
<tr>
<th>Faster R-CNN+++ [15]</th>
<th>ResNet-101-C4</th>
<th>AP^{bb}</th>
<th>AP^{bb}_{50}</th>
<th>AP^{bb}_{75}</th>
<th>AP^{bb}_S</th>
<th>AP^{bb}_M</th>
<th>AP^{bb}_L</th>
</tr>
</thead>
<tbody>
<tr>
<td>34.9</td>
<td>55.7</td>
<td>37.4</td>
<td>15.6</td>
<td>38.7</td>
<td>50.9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Faster R-CNN w FPN [22]</td>
<td>ResNet-101-FPN</td>
<td>36.2</td>
<td>59.1</td>
<td>39.0</td>
<td>18.2</td>
<td>39.0</td>
<td>48.2</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>
<td>36.7</td>
<td>13.5</td>
<td>38.1</td>
<td>52.0</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>
<td>39.2</td>
<td>16.2</td>
<td>39.8</td>
<td>52.1</td>
</tr>
<tr>
<td>Faster R-CNN, RoIAlign</td>
<td>ResNet-101-FPN</td>
<td>37.3</td>
<td>59.6</td>
<td>40.3</td>
<td>19.8</td>
<td>40.2</td>
<td>48.8</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNet-101-FPN</td>
<td>38.2</td>
<td>60.3</td>
<td>41.7</td>
<td>20.1</td>
<td>41.1</td>
<td>50.2</td>
</tr>
<tr>
<td>Mask R-CNN</td>
<td>ResNeXt-101-FPN</td>
<td>39.8</td>
<td>62.3</td>
<td>43.4</td>
<td>22.1</td>
<td>43.2</td>
<td>51.2</td>
</tr>
</tbody>
</table>

bbox detection improved by:
- RoIAlign
- Multi-task training w/ mask
Mask R-CNN results on COCO
Mask R-CNN results on COCO

small objects
Mask R-CNN results on CityScapes
Failure case: detection/segmentation

Mask R-CNN results on COCO

missing

missing, false mask
Failure case: recognition

Mask R-CNN results on COCO

not a kite
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)
Validation image with box detection shown in red

28x28 soft prediction

Resized Soft prediction

Final mask
Mask R-CNN: for Human Keypoint Detection

• 1 keypoint = 1-hot “mask”
• Human pose = 17 masks

• Softmax over spatial locations
  • e.g. $56^2$-way softmax on 56x56

• Desire the same equivariances
  • translation, scale, aspect ratio
Conclusion

Mask R-CNN
✓ Good speed
✓ Good accuracy
✓ Intuitive
✓ Easy to use
✓ Equivariance matters

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

More about Mask R-CNN in this ICCV
• ICCV oral presentation, 10/26, 9am
• COCO workshop talk, 10/29, 9am