Mask R-CNN

ICCV 2017, Facebook AI Research (FAIR)

2018. 12. 31.

Jihun **Kim**, Hanyang Univ.

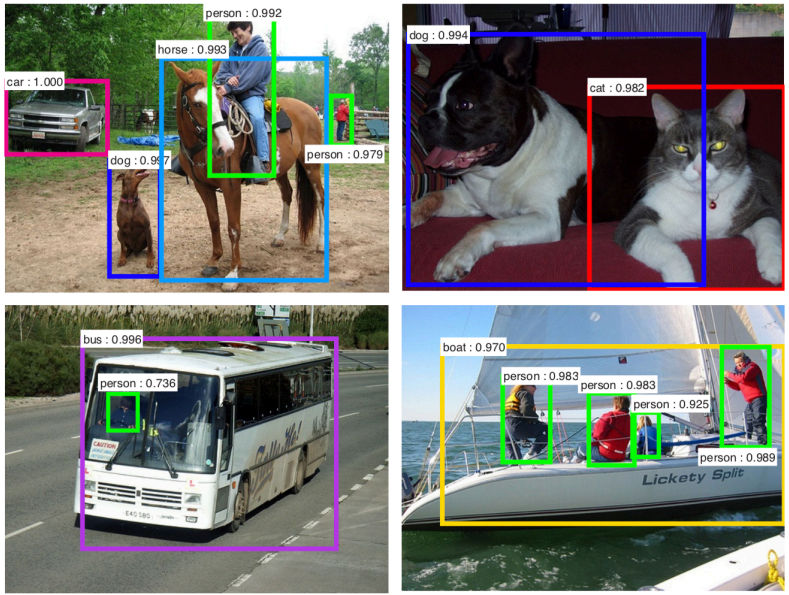
Introduction

He et al. Mask R-CNN. ICCV 2017

Vision Applications

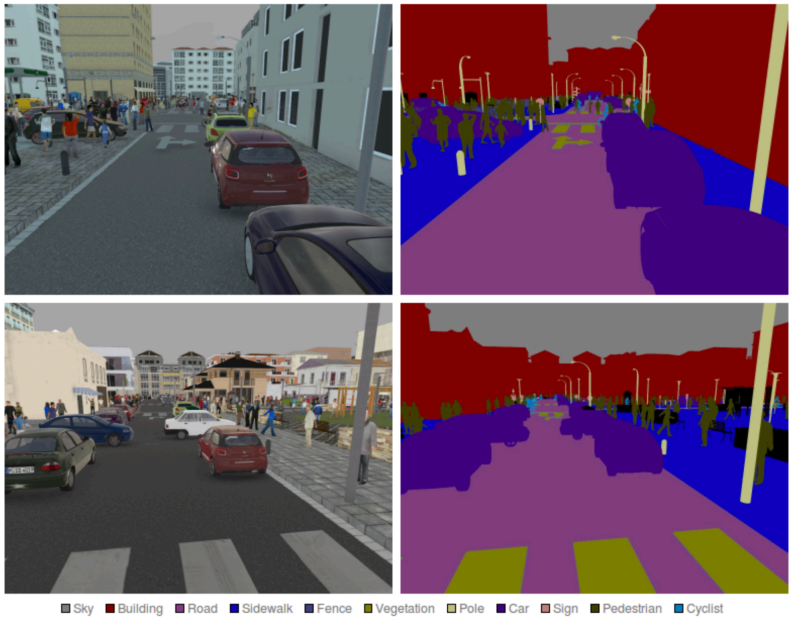
- Visual Classification: Image -> Class label
- Object Detection: Image -> Class label + Localization
- Semantic Segmentation: Image -> Class label + Localization per pixel
- Image Captioning: Image -> Sentences
- Visual Tracking: Image Sequence -> Object localization
- Generative Models: Random vector or Image -> Image
- Structure from Motion: Motion -> Structure

What's different?



Object Detection

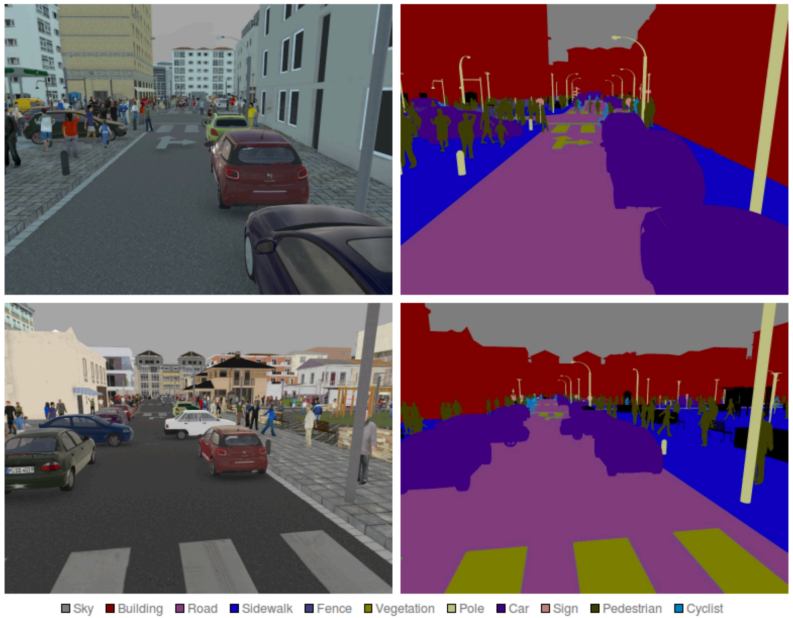
(Ren et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2017.)



Semantic Segmentation

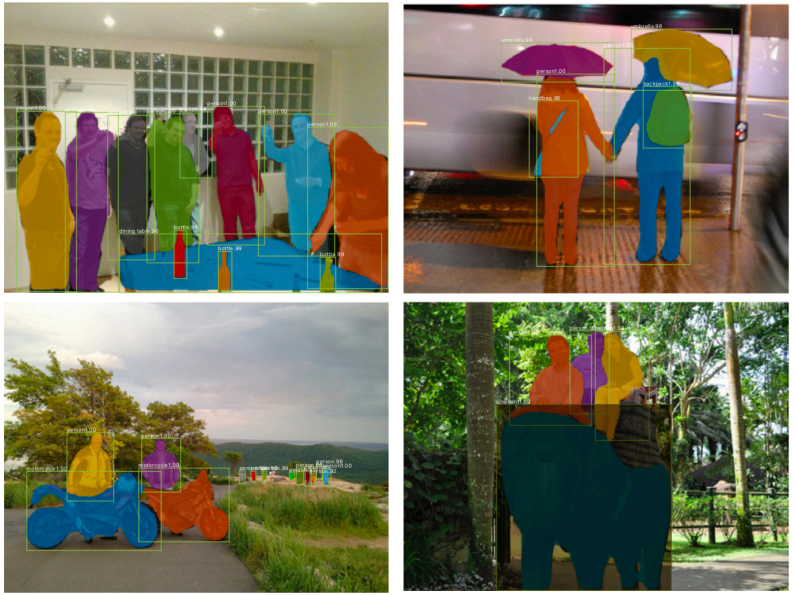
(G. Heinrich. Image Segmentation Using DIGITS 5. NVIDIA Developer Blog. https://devblogs.nvidia.com/image-segmentation-using-digits-5/)

What's different?



Semantic Segmentation

(G. Heinrich. Image Segmentation Using DIGITS 5. NVIDIA Developer Blog. https://devblogs.nvidia.com/image-segmentation-using-digits-5/)



Instance Segmentation

(He et al. Mask R-CNN. ICCV 2017.)

Approach

Instance Segmentation

(He et al. Mask R-CNN. ICCV 2017.)

Difficult!
Because it requires the correct detection of all objects in an image
while also precisely segmenting each instance

1. Segmentation-first strategy
-> Semantic segmentation result (e.g., FCN output)
-> cut pixels of it

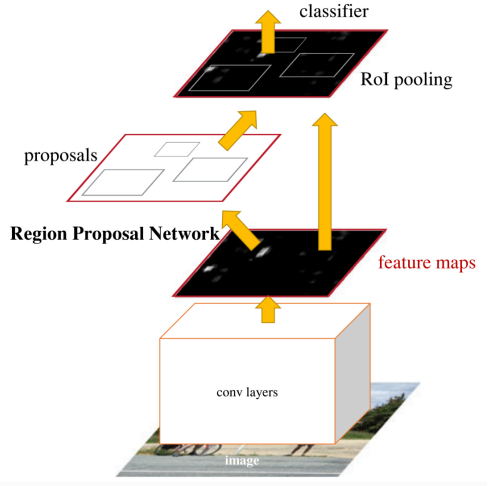
2. Instance-first strategy
-> Object detection + Semantic segmentation?

Structure

He et al. Mask R-CNN. ICCV 2017

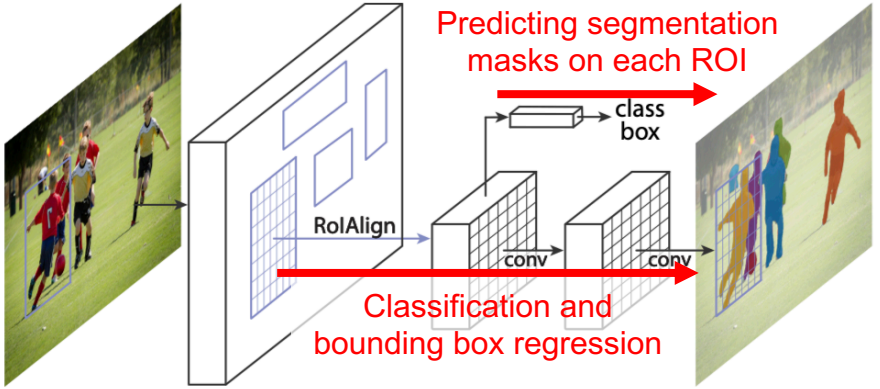
Basic Structure

In principle Mask R-CNN is an intuitive extension of Faster R-CNN.



Faster R-CNN

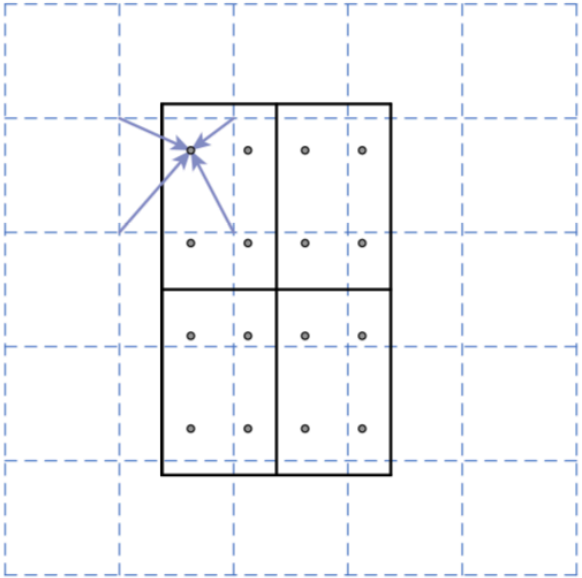
(Ren et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. TPAMI 2017.)



Mask R-CNN

(He et al. Mask R-CNN. ICCV 2017.)

RoIAlign



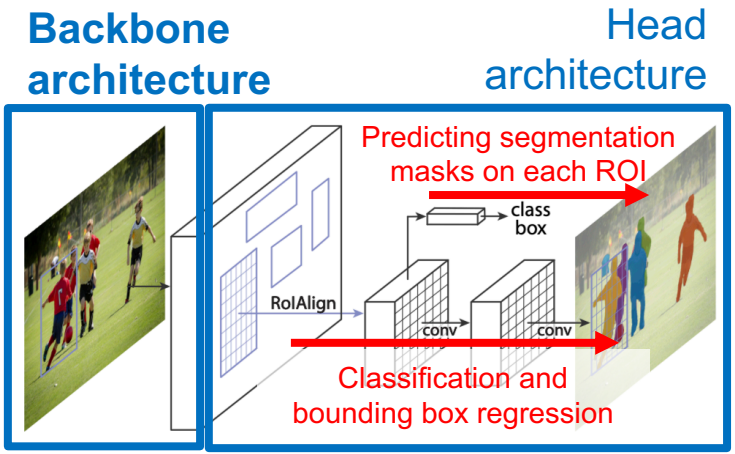
RoIPool is a standard operation for extracting a small feature map (e.g., 7×7) from each RoI.

RoIPool performs quantizations, theses introduce misalignments between the RoI and the extracted features.

Solution -> **RoIAlign!**
No quantization is performed on any coordinates.

RoIAlign has a large impact: it improves mask accuracy by relative 10% to 50%.

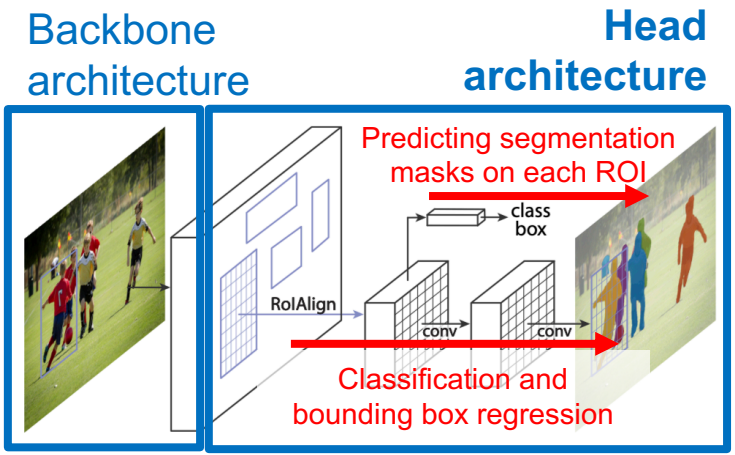
Backbone architecture

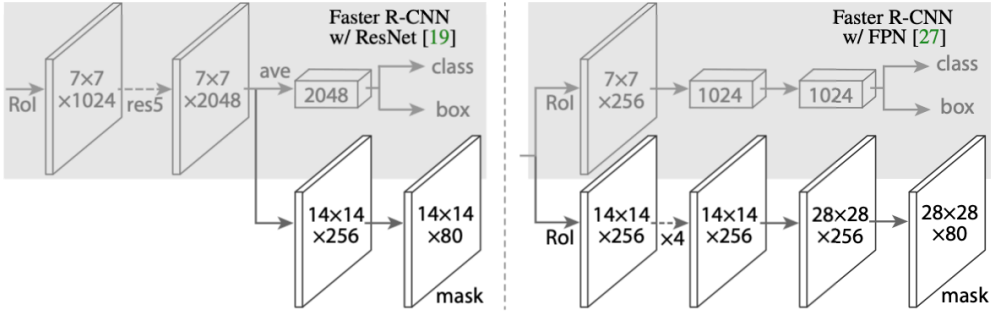


- Evaluated ResNet and ResNeXt networks of depth 50 or 101 layers.
- Also explored Feature Pyriamid Network(FPN).

Using a ResNet-FPN backbone for feature extraction with Mask R- CNN gives excellent gains in both accuracy and speed.

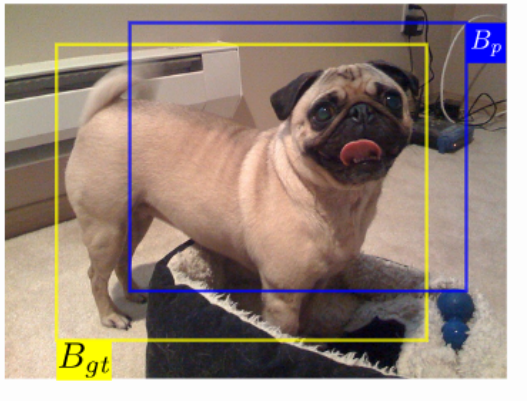
Head architecture



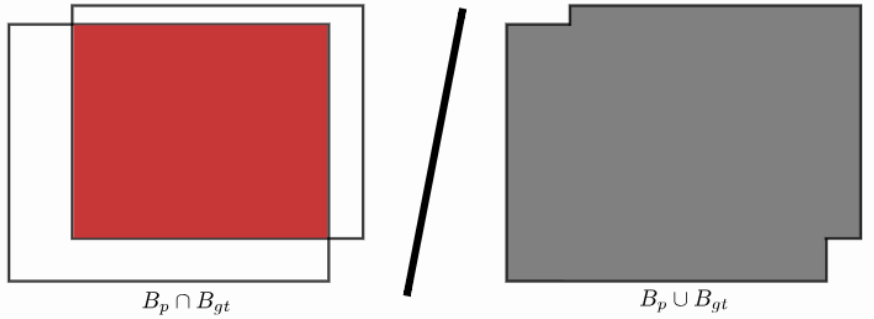


The mask branch can predict K masks per RoI, but we only use the k-th mask, where k is the predicted class by the classification branch.

Training



$$B_p\text{와 }B_{gt}\text{ 의 IOU} = \frac{B_p \cap B_{gt}\text{ 영역 넓이}}{B_p \cup B_{gt}\text{ 영역 넓이}}$$



- An RoI is considered positive if it has IoU with a ground-truth box of at least 0.5 and negative otherwise.
- The mask loss L_{mask} is defined only on positive RoIs.

Results

He et al. Mask R-CNN. ICCV 2017

Result

net-depth-features	AP	AP_{50}	AP_{75}
ResNet-50-C4	30.3	51.2	31.5
ResNet-101-C4	32.7	54.2	34.3
ResNet-50-FPN	33.6	55.2	35.3
ResNet-101-FPN	35.4	57.3	37.5
ResNeXt-101-FPN	**36.7**	**59.5**	**38.9**

(a) **Backbone Architecture**: Better backbones bring expected gains: deeper networks do better, FPN outperforms C4 features, and ResNeXt improves on ResNet.

	AP	AP_{50}	AP_{75}
softmax	24.8	44.1	25.1
sigmoid	**30.3**	**51.2**	**31.5**
	+5.5	*+7.1*	*+6.4*

(b) **Multinomial *vs.* Independent Masks** (ResNet-50-C4): *Decoupling* via per-class binary masks (sigmoid) gives large gains over multinomial masks (softmax).

	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**

(c) **RoIAlign** (ResNet-50-C4): Mask results with various RoI layers. Our RoIAlign layer improves AP by ~3 points and AP_{75} by ~5 points. Using proper alignment is the only factor that contributes to the large gap between RoI layers.

	AP	AP_{50}	AP_{75}	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
RoIPool	23.6	46.5	21.6	28.2	52.7	26.9
RoIAlign	**30.9**	**51.8**	**32.1**	**34.0**	**55.3**	**36.4**
	+7.3	*+ 5.3*	*+10.5*	*+5.8*	*+2.6*	*+9.5*

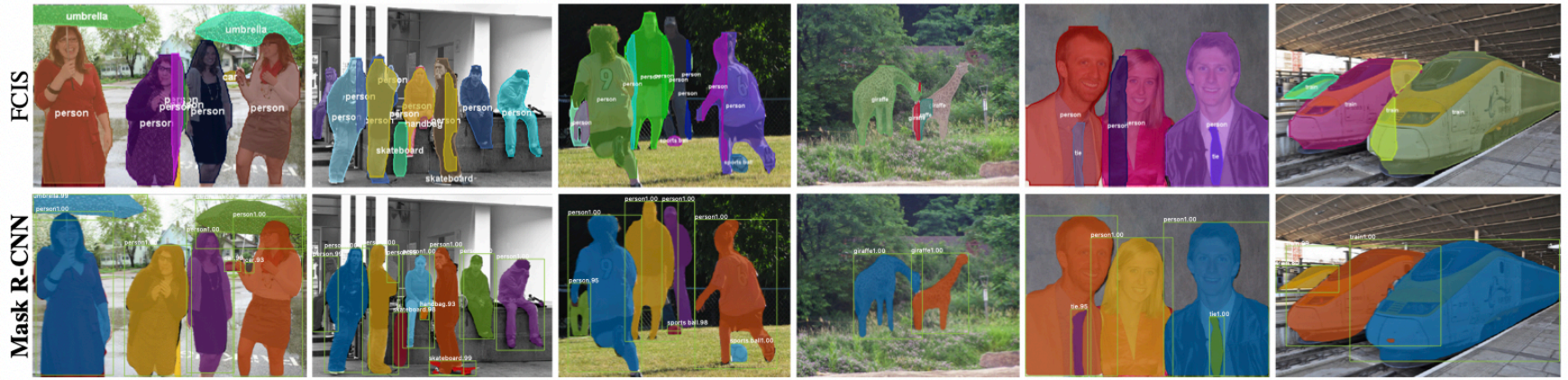
(d) **RoIAlign** (ResNet-50-**C5**, *stride 32*): Mask-level and box-level AP using *large-stride* features. Misalignments are more severe than with stride-16 features (Table 2c), resulting in big accuracy gaps.

	mask branch	AP	AP_{50}	AP_{75}
MLP	fc: 1024→1024→80·28^2	31.5	53.7	32.8
MLP	fc: 1024→1024→1024→80·28^2	31.5	54.0	32.6
FCN	conv: 256→256→256→256→256→80	**33.6**	**55.2**	**35.3**

(e) **Mask Branch** (ResNet-50-FPN): Fully convolutional networks (FCN) *vs.* multi-layer perceptrons (MLP, fully-connected) for mask prediction. FCNs improve results as they take advantage of explicitly encoding spatial layout.

Table 2. **Ablations**. We train on `trainval35k`, test on `minival`, and report *mask* AP unless otherwise noted.

Result



FCIS+++ (top) vs. Mask R-CNN (bottom, ResNet-101-FPN).

Result

	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8	15.5	38.1	52.4
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**	**16.9**	**39.9**	**53.5**

Instance segmentation mask AP on COCO test-dev.

References

K. He, G. Gkioxari, P. Dolla´r, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.