Paper Review 4

Mask R-CNN

0. Abstract

Mask R-CNN은 object instance segmentation을 위해 기존의 Faster R-CNN에 mask branch를 더한 모델이다. 이는 이미지의 객체를 탐지함과 동시에 각 instance에 대해 높은 품질의 segmentation mask를 생성한다. Mask R-CNN은 훈련이 간단하고 Faster R-CNN에 약간의 overhead만을 더해 5fps의 속도를 가진다. 또힌 human pose estimate와 같은 다른 task에도 일반화하기 쉽다는 장점을 가진다.

1. Introduction

객체 탐지 분야에서는 Fast/Faster R-CNN, FCN, semantic sementation의 3가지가 주를 이루었다. 본 논문은 위 방식들을 발전시켜 instance segmentation를 위한 비교적 유용한 프레임 워크를 개발하는 것이 목표이다.

Mask R-CNN은 instance segmantation을 위한 모델로서 detection과 segmentation을 동시에 처리해야 하는데, 이를 위해 object detection과 semantic segmentation을 결합하였다.

이러한 Mask R-CNN의 구조는 아래와 같다.

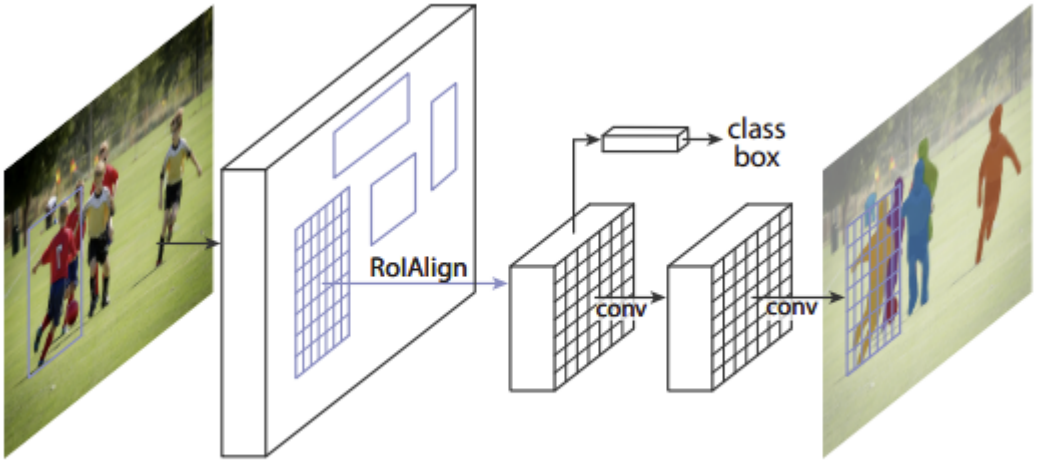


Figure 1. The **Mask R-CNN** framework for instance segmentation.

Mask R-CNN은 Faster R-CNN을 확장하여, classification과 bounding box regression에 평행하게 각 RoI(Region of Interest)에 segmentation mask 를 예측하는

branch를 추가한 구조를 가진다. 이를 Mask branch라고 하며, Mask branch는 각 RoI에 작은 크기의 FCN(Fully Convolutional Network)이 추가된 형태이다. Mask branch는 classification과 bounding box regression branch와 독립적이고, mask prediction과 class prediction을 분리했기 때문에 연산이 빠르다.

Mask R-CNN은 RoIAlign layer를 추가하여 Faster R-CNN의 문제점인 오정렬(misalignment)을 막았으며, 이는 정확한 분할 위치를 보존하여 정확도를 높였다.

2. Related Work

- R-CNN: Faster R-CNN, RPN(Region Proposal Network)
- Instance Segmentation: Semantic Segmentation

3. Mask R-CNN

Faster R-CNN

Faster R-CNN은 2가지 단계로 구성되어 있다.

1. RPN: 객체 bounding box를 제안한다.
2. Fast R-CNN: RoIPooling을 적용하여 1단계에서 얻은 각 box별로 분류와 회귀를 수행한다.

두 단계는 동일한 feature map을 공유한다.

Mask R-CNN

Mask R-CNN도 Faster R-CNN과 같은 2단계의 구조를 가지고 있다. 1단계인 RPN은 동일하고, 2단계에서는 기존 class 와 box offset 예측과 함께 각각 RoI에 대해 binary mask를 출력한다.

각 RoI에 대한 multi-task loss는 다음과 같다.

$$L = L_{cls} + L_{box} + L_{mask}$$

Lcls와 Lbox는 Fast R-CNN과 동일하고, 픽셀별로 sigmoid를 적용하여 Lmask 는 average binary cross-entropy loss로 정의한다. 이는 기존에 픽셀 별로 softmax와 multinomial cross-entropy-loss를 사용한 방법과 다르며 더 좋은 성능을 보여준다.

Mask Representation

Mask는 객체의 spatial layout을 인코딩한다. 따라서 기존 모델이 class label나 box offset과 같이 벡터로 무너지는 것과 달리 mask의 공간 구조를 추출하는 것은 합성곱을 통한 pixel-to-pixel 일치로 자연스럽게 가능하다.

Mask 예측을 위해 FCN을 사용한 mask branch layer들은 m x m 개의 object spatial layout을 유지한다. 이는 fc layer를 사용하는 것보다 매개변수의 개수도 더 작고 정확도도 높다.

RoIAlign

Mask R-CNN은 Faster R-CNN이 RoI Pooling을 사용한 것과 달리 RoIAlign을 사용하였다. RoIPooling에서는 feature map을 계산하기 위해 RoI가 quantize된다. 이 quantization은 오정렬을 야기하고, 이것은 classification에서는 괜찮지만pixel 별로 정확성이 필요한 segmentation에서는 문제가 발생한다.

RoIAilgn의 구조는 다음과 같다.

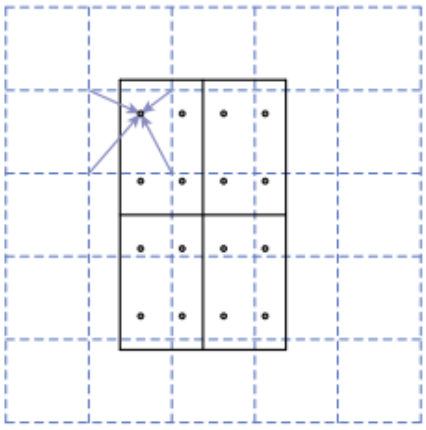
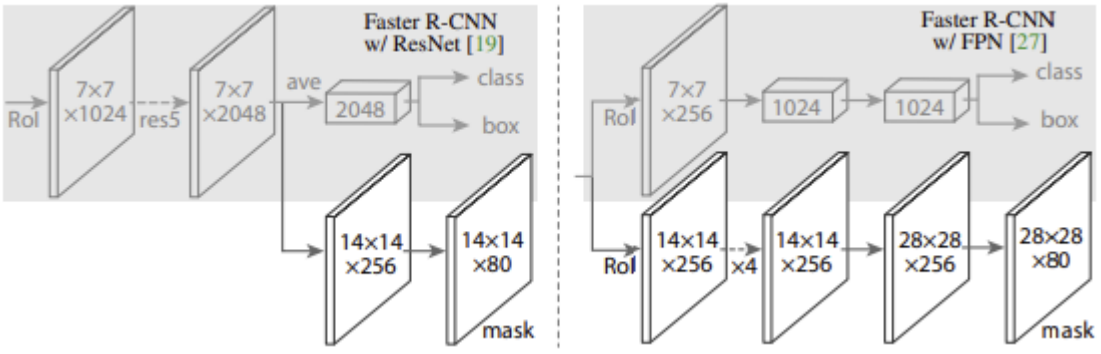


Figure 3. **RoIAlign:** The dashed grid represents a feature map, the solid lines an RoI (with 2×2 bins in this example), and the dots the 4 sampling points in each bin. RoIAlign computes the value of each sampling point by bilinear interpolation from the nearby grid points on the feature map. No quantization is performed on any coordinates involved in the RoI, its bins, or the sampling points.

위 그림에서 박스 한칸이 1px이라고 할때, 박스 안 각 점의 좌표는 정수가 아닌 실수이다. 그러나 이미지 데이터는 정수 좌표 값만을 가지므로 RoIAlign에서는 bilinear interpoliation 방법으로 점의 값을 구하는데, 이 과정에서 quantization은 일어나지 않는다.

Network Architecture

아래 그림은 각각 BackBone이 ResNet인 구조와 FPN인 Mask R-CNN 구조이다.



BackBone은 이미지의 특징을 추출하기 위해 사용되는데, 50/101 layers의 ResNet/ResNeXt과 FPN(Feature Pyramid Network)를 사용하였다. Head는 bounding box 인식과 mask 예측을 위해 사용되는데, Faster R-CNN의 Head에 Mask branch를 추가한 구조이다.

4. Experiments: Instance Segmentation

COCO dataset을 사용하였으며, AP를 통해 정확도를 측정하였다.

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.

위 실험에서는 sigmoid가 softmax보다 좋은 성능을 보였으며, RoIAlign이 RoIPool이나 RoIWarp에 비해 좋은 성능을 나타내는 것을 알 수 있다. 또한 Mask Branch에 MLP를 사용하는 것보다 FCN을 사용하는 것이 더 좋은 성능을 보였다.

	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}	AP^{bb}_{S}	AP^{bb}_{M}	AP^{bb}_{L}
Faster R-CNN+++ [19]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [27]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [21]	Inception-ResNet-v2 [41]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [39]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3	19.8	40.2	48.8
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7	20.1	41.1	50.2
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**	**22.1**	**43.2**	51.2

Backbone이 ResNeXt-101-FPN Mask R-CNN가 가장 좋은 성능을 기록하였다.

5. Mask R-CNN for Human Pose Estimation

Figure 7. Keypoint detection results on COCO `test` using Mask R-CNN (ResNet-50-FPN), with person segmentation masks predicted from the same model. This model has a keypoint AP of 63.1 and runs at 5 fps.

Mask R-CNN은 Human Pose Estimation에도 적용이 가능하다. 각 keypoint의 위치를 one-hot binary mask로 모델링하여 k개의 키포인트 유형에 대해 예측을 할 수 있다.

이 실험은 ResNet-50-FPN Backbone을 사용하여 AP 평가를 진행하였고, Mask R-CNN은 COCO 2016 우승자보다 높은 62.7 APkp를 기록하였다. RoIAlign은 RoIPool에 비해 4.4APkp를 향상시켰으며 추후 다른 instance segmentation에도 유용하게 활용될 것으로 기대된다.