Date: 30.10.2023 / 12.11.2023

Blog Post#2

Ezgi Gümüştekin & Emir Yorgun

Mask R-CNN is a deep neural network that aims to carry out Instance Segmentation. It can separate different objects in an image or video.

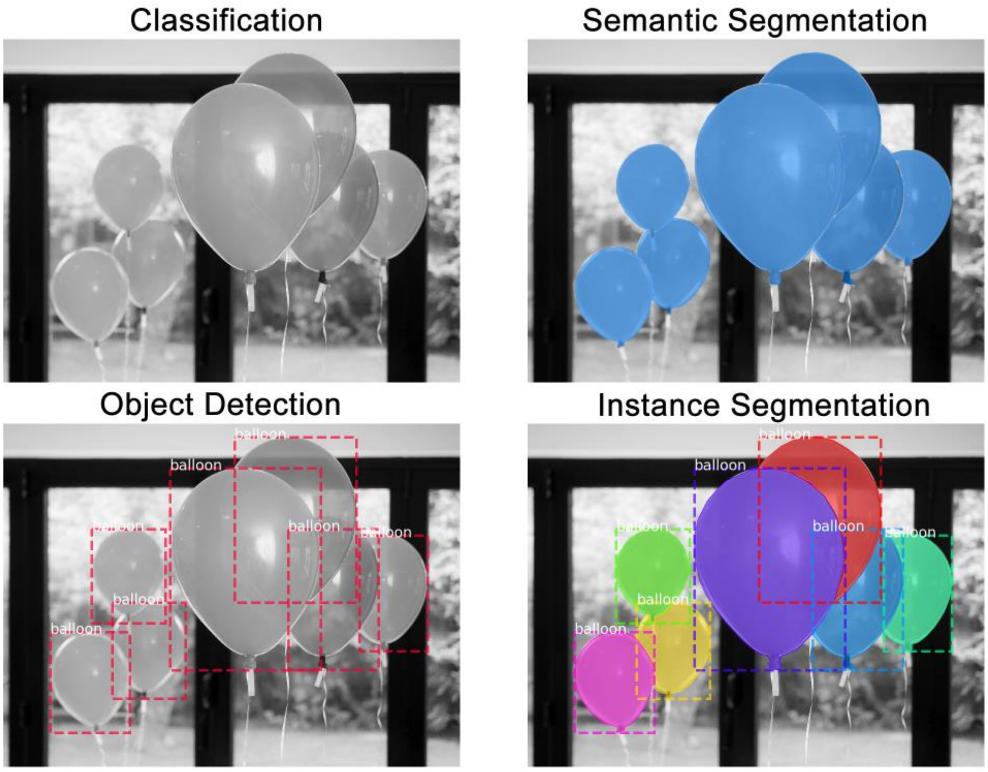


Fig. [1] - Instance Segmentation

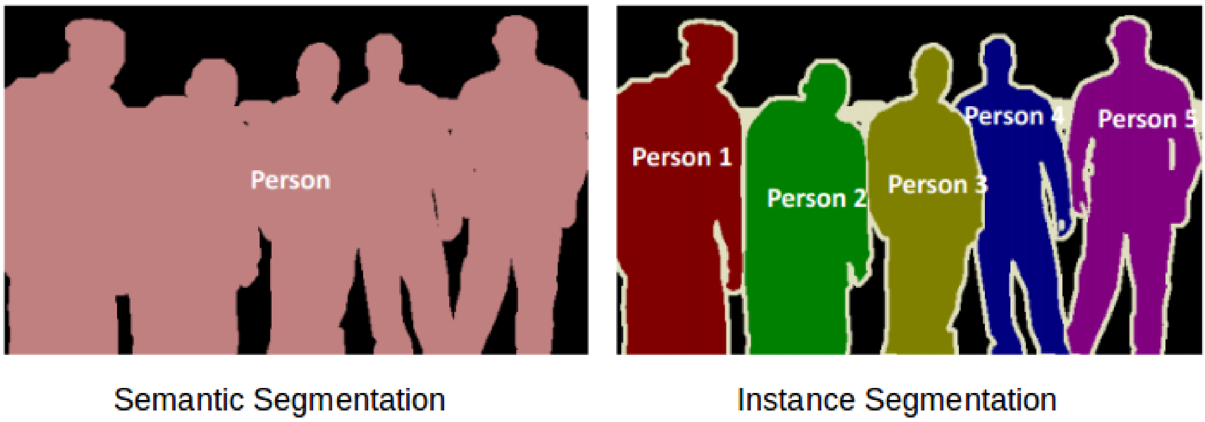
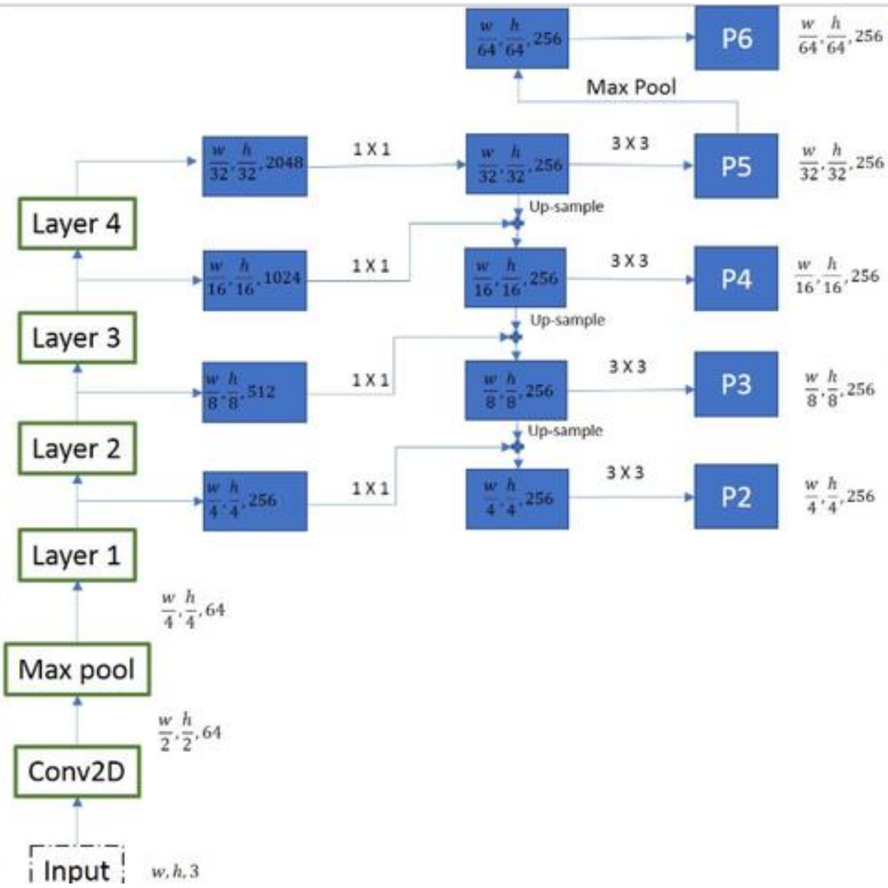


Fig. [2] - Instance Segmentation

It comprises of; Backbone Network, Region Proposal Network, Mask Representation, RoI Align.

This, Backbone Network, multi-layer feature pyramid network generates RoI of different scale which improves the accuracy of previous ResNet architecture.



Input
w, h, 3
Conv2D
$\frac{w}{2}, \frac{h}{2}, 64$
Max pool
$\frac{w}{4}, \frac{h}{4}, 64$
Layer 1
Layer 2
Layer 3
Layer 4
$\frac{w}{4}, \frac{h}{4}, 256$
$\frac{w}{8}, \frac{h}{8}, 512$
$\frac{w}{16}, \frac{h}{16}, 1024$
$\frac{w}{32}, \frac{h}{32}, 2048$
1 X 1
1 X 1
1 X 1
1 X 1
$\frac{w}{64}, \frac{h}{64}, 256$
$\frac{w}{32}, \frac{h}{32}, 256$
$\frac{w}{16}, \frac{h}{16}, 256$
$\frac{w}{8}, \frac{h}{8}, 256$
$\frac{w}{4}, \frac{h}{4}, 256$
Up-sample
Up-sample
Up-sample
Max Pool
3 X 3
3 X 3
3 X 3
3 X 3
P6
P5
P4
P3
P2
$\frac{w}{64}, \frac{h}{64}, 256$
$\frac{w}{32}, \frac{h}{32}, 256$
$\frac{w}{16}, \frac{h}{16}, 256$
$\frac{w}{8}, \frac{h}{8}, 256$
$\frac{w}{4}, \frac{h}{4}, 256$

Fig. 3 - Mask R-CNN Backbone Architecture [4]

Region Proposal Network, all the convolution feature map that is generated by the previous layer is passed through a 3*3 convolution layer. The output of this is then passed into two parallel branches that determine the objectness score and regress the bounding box coordinates.

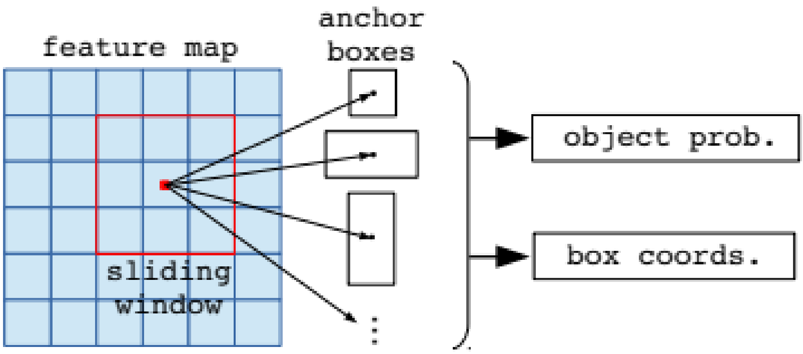
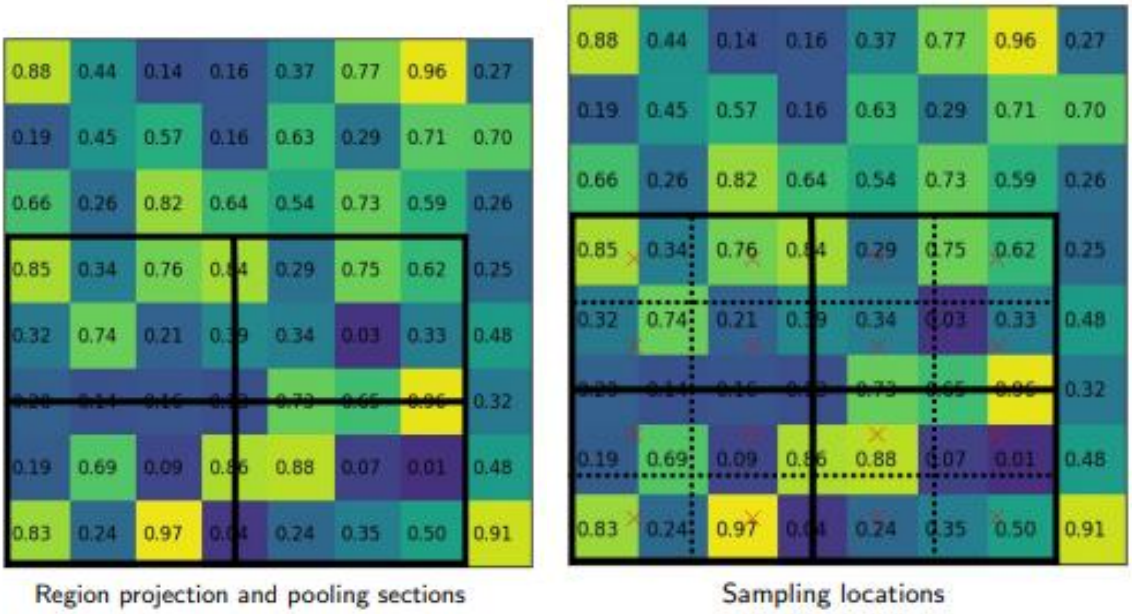


Fig 4. - Region Proposal Network - Anchor Generation Mask R-CNN [4]

Mask Representation, is a small Fully Convolutional Network (FCN) applied to each Region of Interest (RoI). The FCN outputs a fixed-size mask for each RoI and each class. The mask is used to predict the pixel-level segmentation of the object instances in the test images.

RoI Align, is a technique that improves the accuracy of object detection and instance segmentation fmodels in RoI Pooling.



0.88	0.44	0.14	0.16	0.37	0.77	0.96	0.27
0.19	0.45	0.57	0.16	0.63	0.29	0.71	0.70
0.66	0.26	0.82	0.64	0.54	0.73	0.59	0.26
0.85	0.34	0.76	0.84	0.29	0.75	0.62	0.25
0.32	0.74	0.21	0.39	0.34	0.03	0.33	0.48
0.20	0.14	0.16	0.13	0.73	0.65	0.96	0.32
0.19	0.69	0.09	0.86	0.88	0.07	0.01	0.48
0.83	0.24	0.97	0.04	0.24	0.35	0.50	0.91

Fig. 5 - ROI Align [4]

References

[1] https://merveelifsarac.medium.com/cnn-r-cnn-fast-r-cnn-mask-r-cnn

[2] https://www.analyticsvidhya.com/blog/2019/07/computer-vision-implementing-mask-r-cnn-image-segmentation

[3] https://www.geeksforgeeks.org/r-cnn-vs-fast-r-cnn-vs-faster-r-cnn-ml

[4] https://www.geeksforgeeks.org/mask-r-cnn-ml

[5] https://arxiv.org/abs/1703.06870