Mask R-CNN

Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick
Facebook AI Research (FAIR)

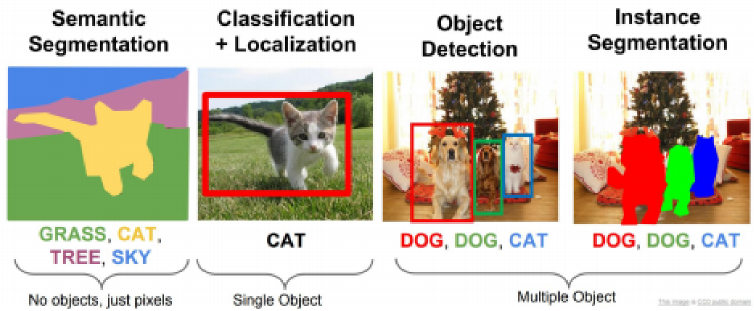
Rishabh Tiwari

Novosibirsk State university

10 November 2020

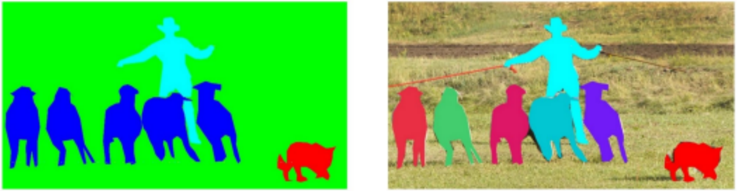
Introduction

Types of Computer Vision Tasks



Introduction

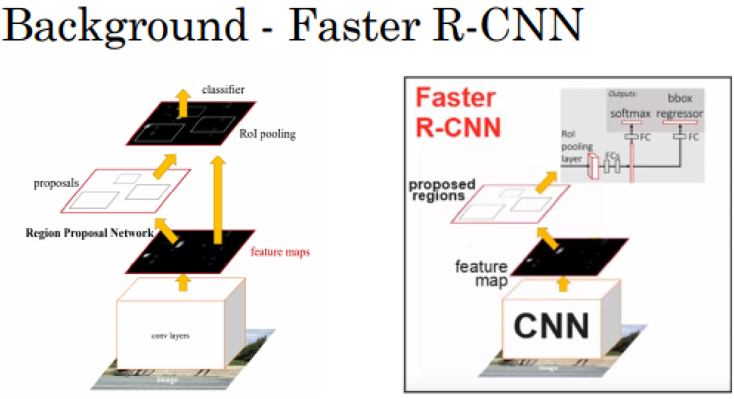
Semantic vs Instance Segmentation



OverView Of Mask R-CNN

- To create a framework for Instance segmentation.
- Builds on top of Faster R-CNN by adding a parallel branch.
- For each Region of Interest (RoI) predicts segmentation mask using a small FCN.
- Changes RoI pooling in Faster R-CNN to a quantization-free layer called RoI.
- Generate a binary mask for each class independently: decouples segmentation and classification.
- Easy to generalize to other tasks: Human pose detection.

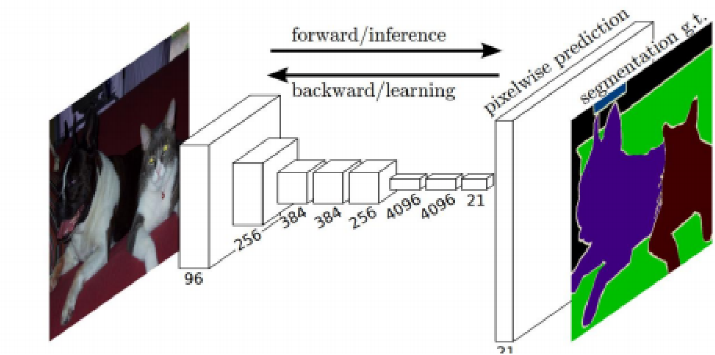
Background - Faster R-CNN



Background - Faster R-CNN

Background FCN

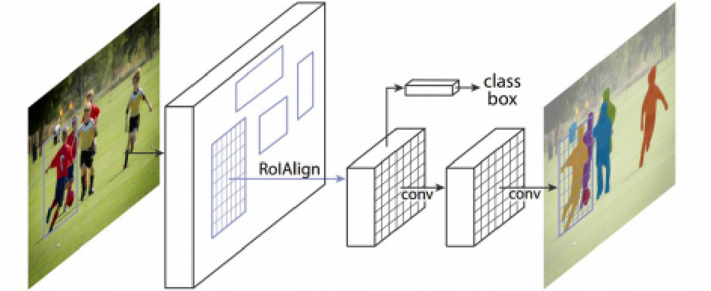
Background - FCN



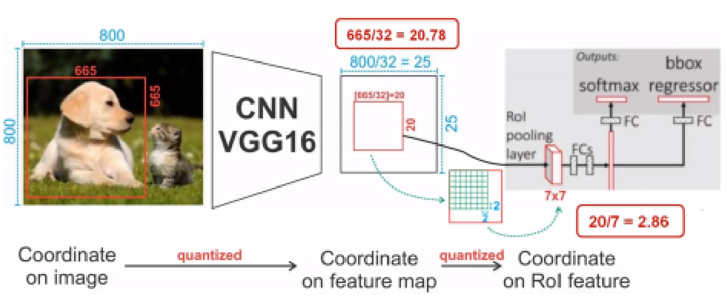
Basic Architecture Of Mask R-CNN

- Procedure:
- RPN ,RoI Align,Parallel prediction for the class, box and binary mask for each RoI.
- Segmentation is different from most prior systems where classification depends on mask prediction.
- Loss function for each sampled RoI is a sum of losses of Box, Class and Mask.

Mask R-CNN Framework



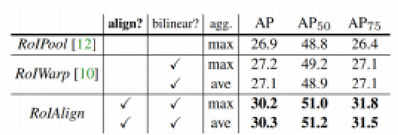
RoI Align – Motivation



ROI ALIGN

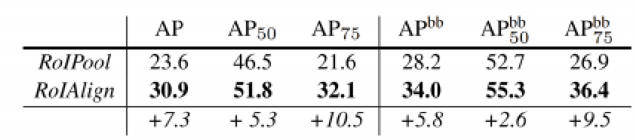
- Removes quantization which causes this misalignment.
- For each bin, you regularly sample 4 locations and do bilinear interpolation.
- Result are not sensitive to exact sampling location or the number of samples
- Compare results with RoI wrapping: Which basically does bilinear interpolation on feature map only.

ROI RESULTS



	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**

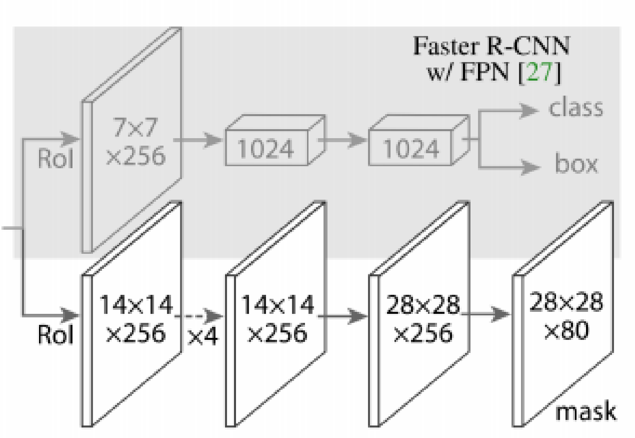
(a) RoIAlign (ResNet-50-C4) comparison



	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*

(b) RoIAlign (ResNet-50-C5, stride 32) comparison

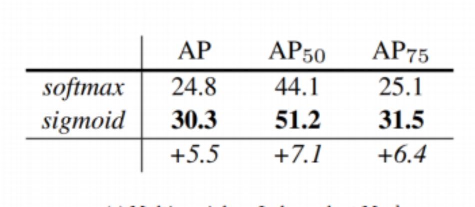
FCN MASK HEAD



Loss Funtion

- To each map a per-pixel sigmoid is applied.
- The map loss is then defined as average binary cross entropy loss.
- Decouples class prediction and mask generation.
- Empirically better results and model becomes easier to train.

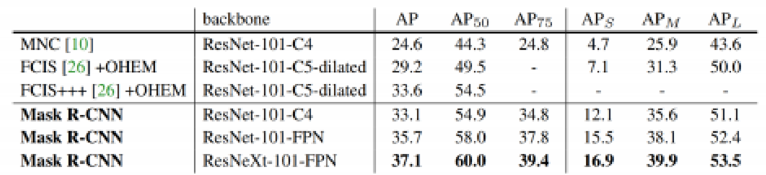
Loss Function - Results



	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*

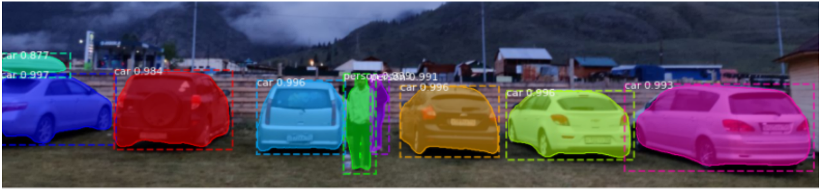
(a) Multinomial vs. Independent Masks

Result

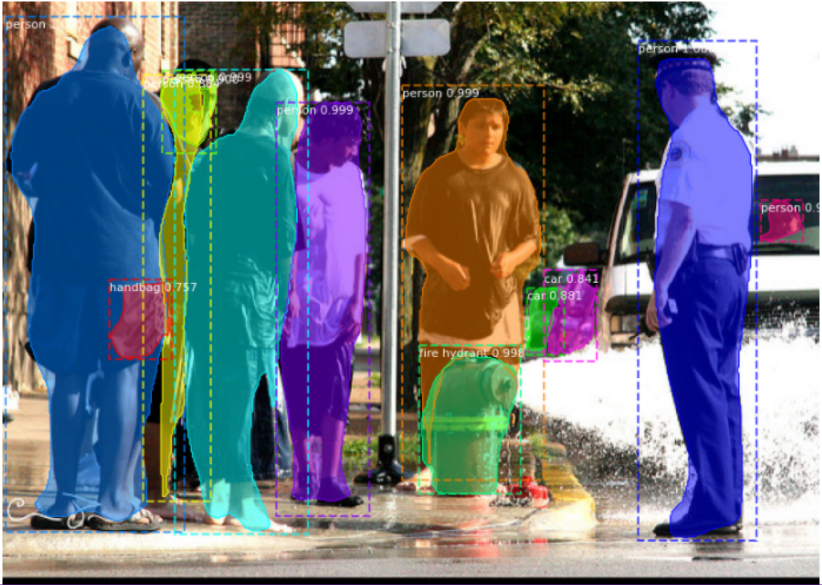


	backbone	AP	AP$_{50}$	AP$_{75}$	AP$_S$	AP$_M$	AP$_L$
MNC [10]	ResNet-101-C4	24.6	44.3	24.8	4.7	25.9	43.6
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-	7.1	31.3	50.0
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-	-	-	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8	12.1	35.6	51.1
Mask R-CNN	ResNet-101-FPN	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**

My Results


car 0.877
car 0.997
car 0.984
car 0.996
person 0.991
car 0.996
car 0.996
car 0.993

My Results


person 1.0
person 0.999
person 0.999
person 0.999
person 1.00
person 0.9
handbag 0.757
car 0.841
car 0.881
fire hydrant 0.998

Conclusion

- A framework to do state-of-art instance segmentation.
- Generates high-quality segmentation mask.
  does Object Detection, Instance Segmentation and can also be extended to human pose estimation.
- All of them are done in parallel.
- Simple to train and adds a small overhead to Faster R-CNN.

THANK YOU