FASTER R-CNN

MASK R-CNN

Michał Tyrolski

Collaboration and Research Group Meeting
17.12.2020
Machine Learning Society
@ University of Warsaw

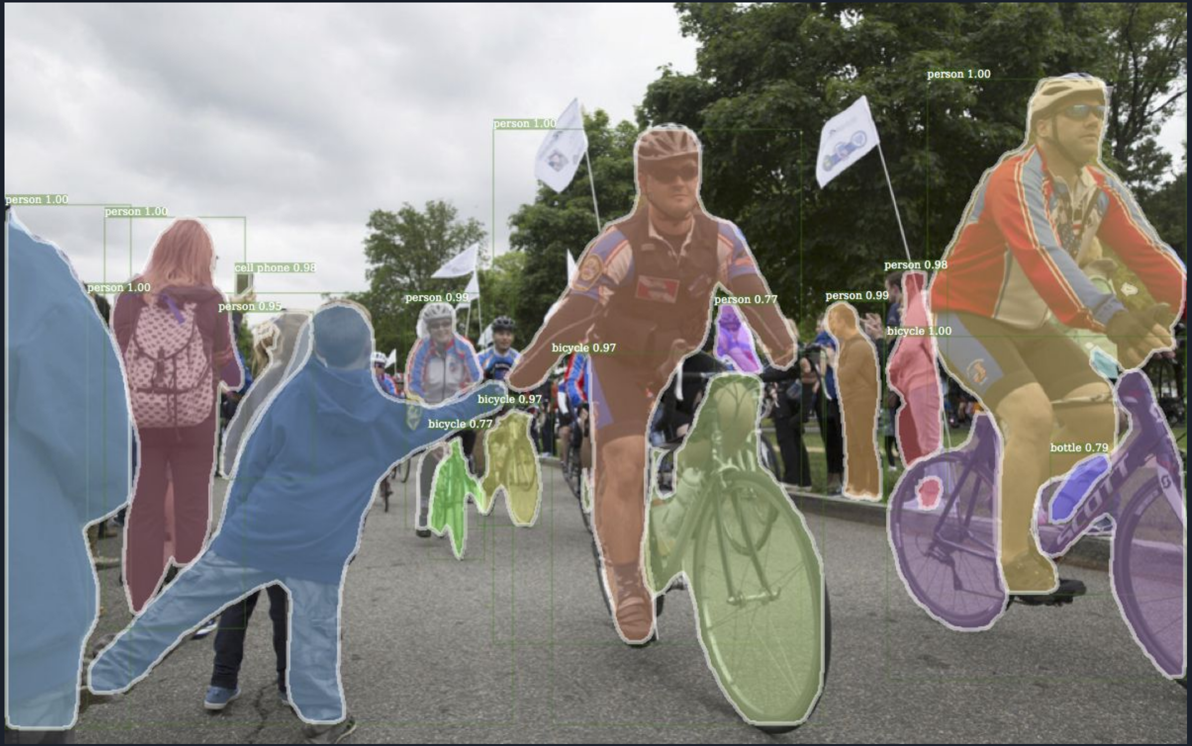
Presentation

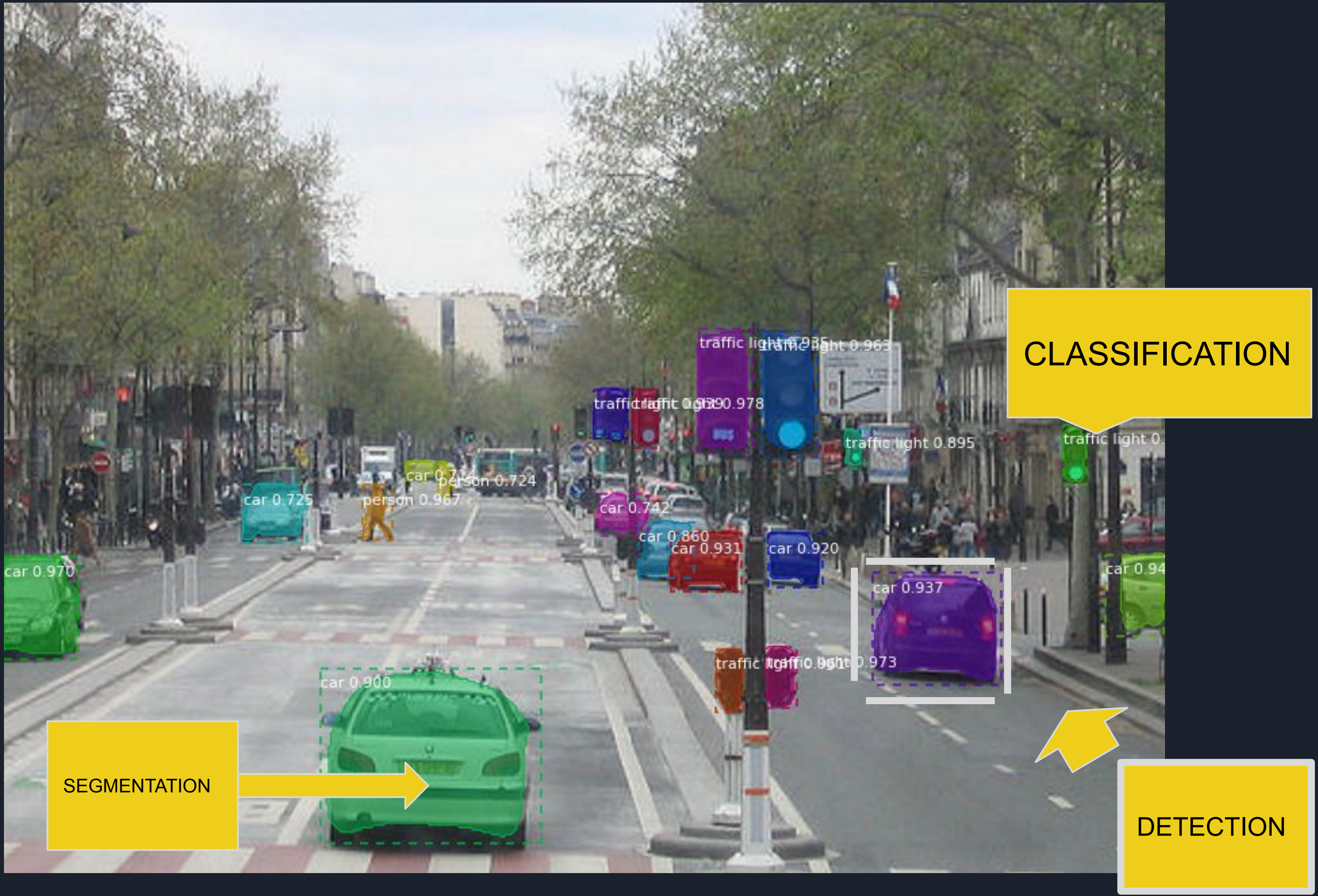
- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

Presentation

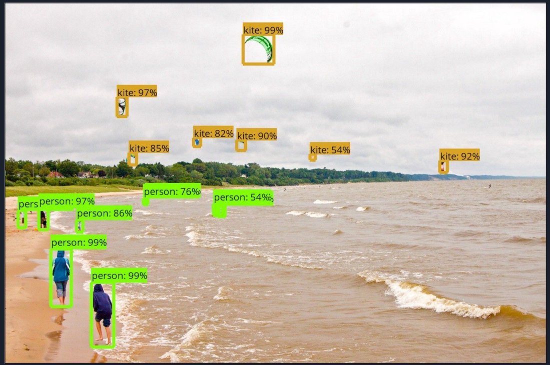
- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

Target: Object Image Segmentation




CLASSIFICATION
SEGMENTATION
DETECTION
traffic light 0.895
person 0.724
car 0.725
person 0.967
car 0.742
car 0.860
car 0.931
car 0.920
car 0.970
car 0.937
car 0.900

TIMELINE



R-CNN (2013)

↓

FAST R-CNN (2015)

↓

FASTER R-CNN (2015)

↓

MASK R-CNN (2017)

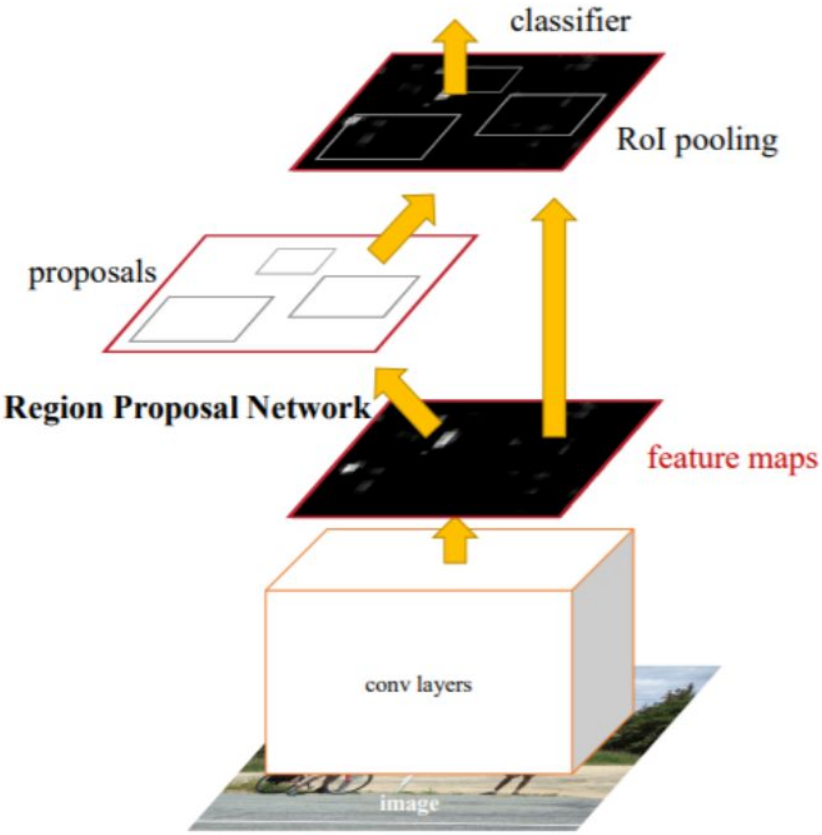
DETECTION + CLASSIFICATION

… + SEMANTIC SEGMENTATION

Presentation

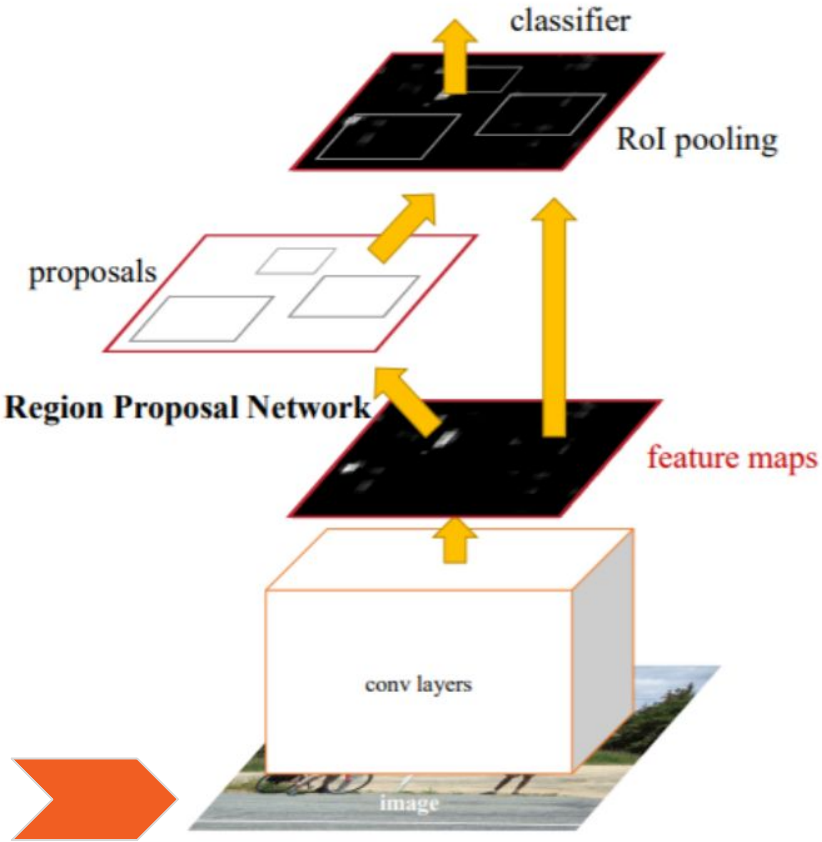
- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

FASTER R-CNN HIGH LEVEL VIEW



Faster R-CNN

FASTER R-CNN HIGH LEVEL VIEW



Faster R-CNN

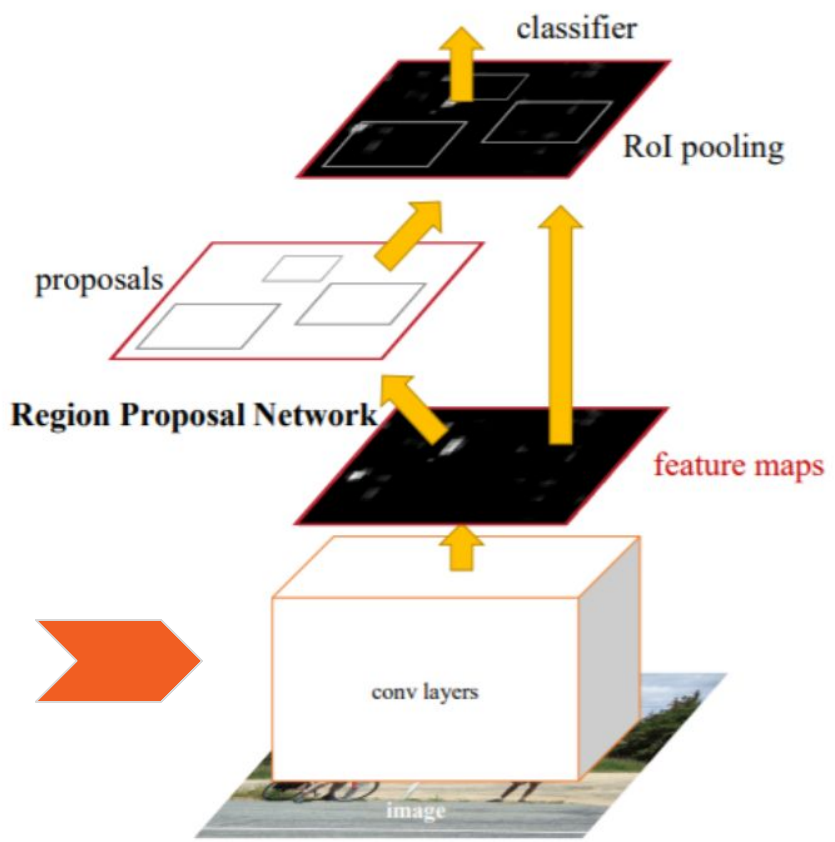
INPUT

Image with shape (h,w,3)

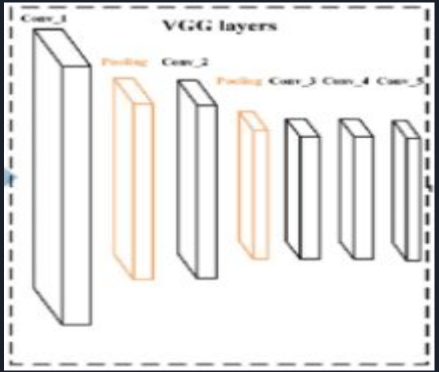
Image size must be divisible by 2 at least 6 times to avoid fractions when downscaling and upscaling.

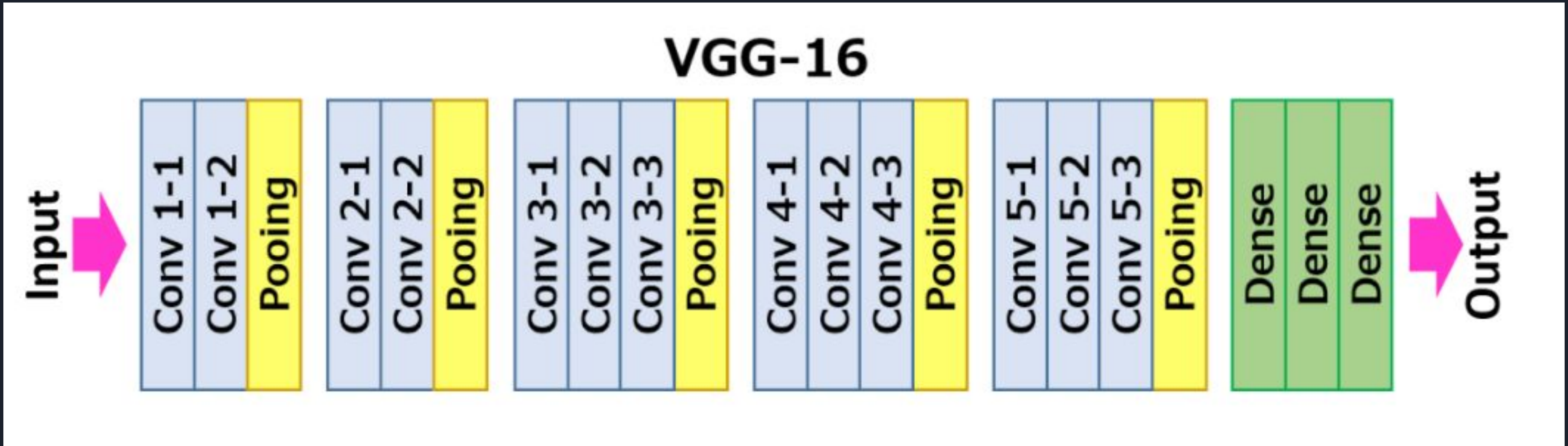
256, 320, 384, 448, 512, ...

FASTER R-CNN HIGH LEVEL VIEW

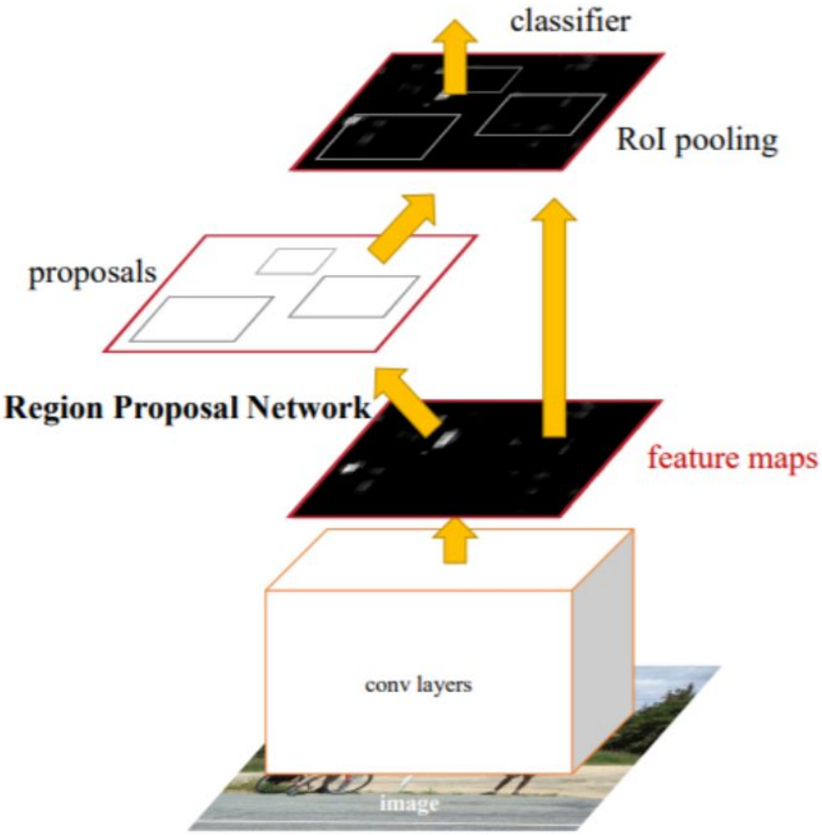


Faster R-CNN


VGG layers

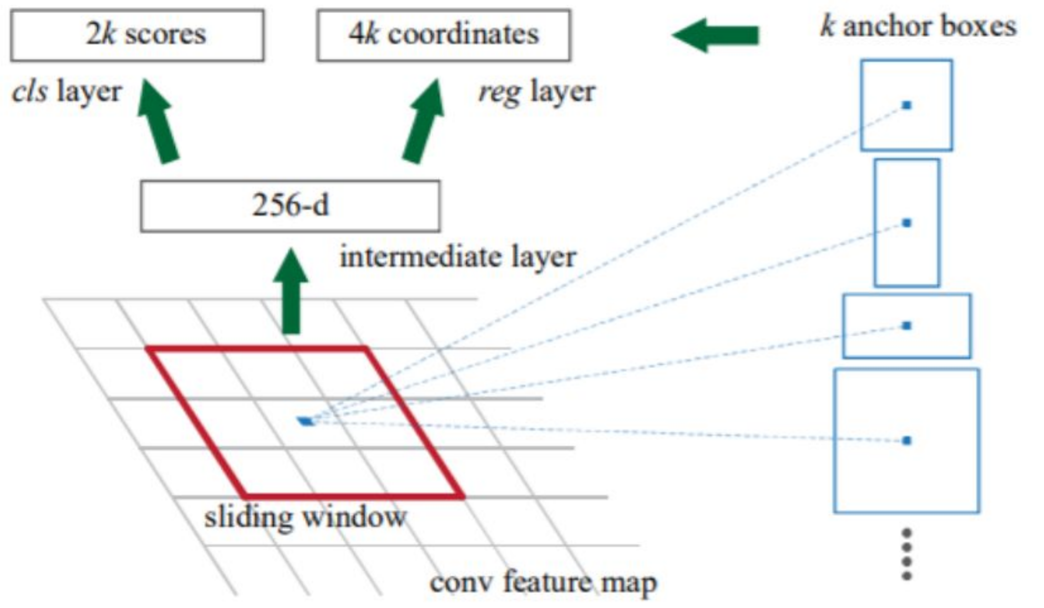

VGG-16
Input
Conv 1-1
Conv 1-2
Pooing
Conv 2-1
Conv 2-2
Pooing
Conv 3-1
Conv 3-2
Conv 3-3
Pooing
Conv 4-1
Conv 4-2
Conv 4-3
Pooing
Conv 5-1
Conv 5-2
Conv 5-3
Pooing
Dense
Dense
Dense
Output

FASTER R-CNN HIGH LEVEL VIEW



Faster R-CNN

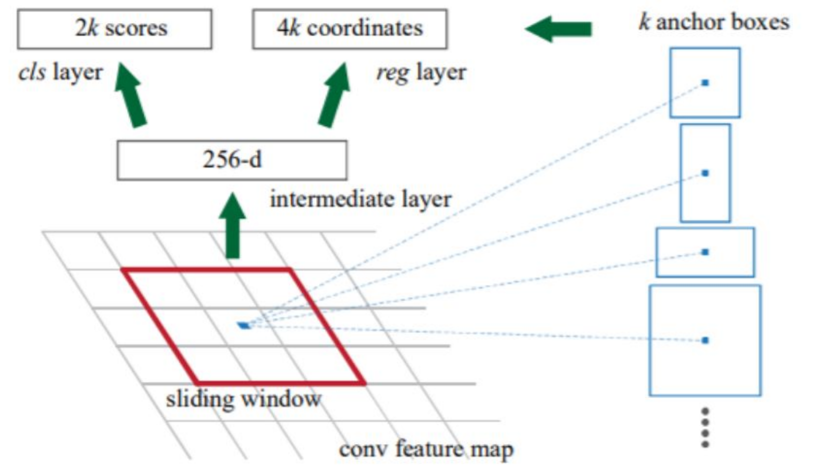
RPN - REGION PROPOSAL NETWORK



Region Proposal Network

RPN - REGION PROPOSAL NETWORK

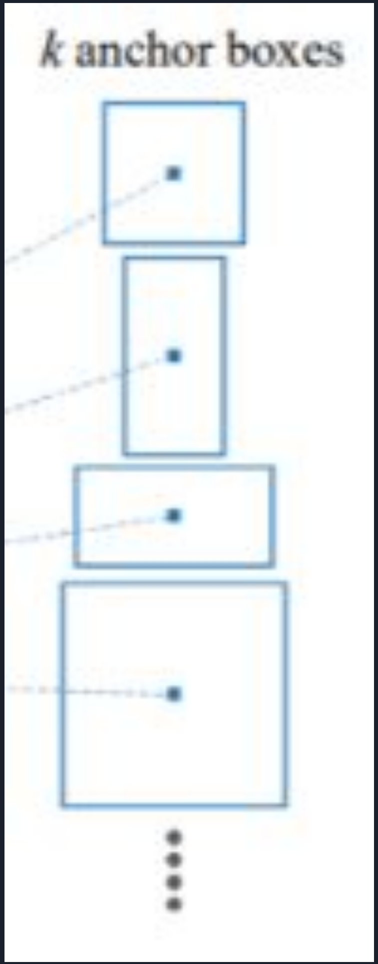
- Small network.
- Slide the network over the convolutional feature map output.
- At each sliding-window location, we simultaneously predict multiple region proposals.
- Each sliding window is mapped to a lower-dimensional feature (512-d for VGG)



Region Proposal Network

ANCHOR BOXES

- From the previous slide...
  At each sliding-window location, we simultaneously predict multiple region proposals.
- An anchor is centered at the sliding window.
- 3 scales and 3 ratios => k==9
- Result of RPN is 4k boxes and 2k objectness scores (can be also just k).
- Feature map W x H resulting with WHk anchors in total (~2400k)



RPN - Training #1 Vocabulary

- Anchor is positive if…
  - has highest IoU overlap with a ground-truth box or
  - IoU > 0.7 (do the job for most positive cases)
- Anchor is negative if…
  - IoU < 0.3 for all ground-truth boxes

RPN - Training #2 Minibatch Loss

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*).$$

RPN - Training #2 Minibatch Loss

$$L(\{p_i\}, \{t_i\}) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*).$$

- p_i* - ground-truth label for anchor type
    - (1 - positive, 0 - negative)
- L_{cls} - Log loss (cross entropy) over 2 classes
    - objects vs not object
- L_{reg} (t_i, t_i*) = R(t_i - t_i*)
    - R - Smooth L1

$$L_{1;smooth} = \begin{cases} |x| & \text{if } |x| > \alpha; \\ \frac{1}{|\alpha|} x^2 & \text{if } |x| \leq \alpha \end{cases}$$

- N_{cls} - mini-batch size (~256)
- N_{reg} - number of anchors (~2.4k)
- \lambda - magic 10 (~ N_reg / N_cls)
- t_i, t_i*

$$\begin{aligned} t_x &= (x - x_a)/w_a, \quad t_y = (y - y_a)/h_a, \\ t_w &= \log(w/w_a), \quad t_h = \log(h/h_a), \\ t_x^* &= (x^* - x_a)/w_a, \quad t_y^* = (y^* - y_a)/h_a, \\ t_w^* &= \log(w^*/w_a), \quad t_h^* = \log(h^*/h_a), \end{aligned}$$

RPN - Training #2 Minibatch Loss

$$L(\{p_i\},\{t_i\}) = \frac{1}{N_{cls}}\sum_i L_{cls}(p_i, p_i^*) + \lambda\frac{1}{N_{reg}}\sum_i p_i^* L_{reg}(t_i, t_i^*).$$

only for positive

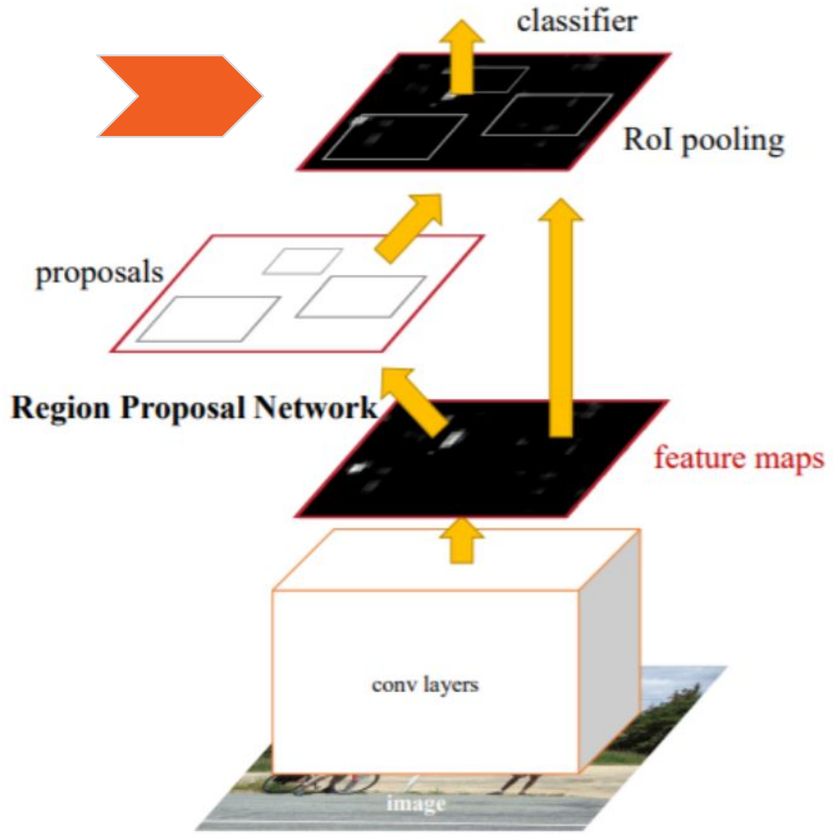
Both RPN and Fast R-CNN trained independently?

Both RPN and Fast R-CNN trained independently?

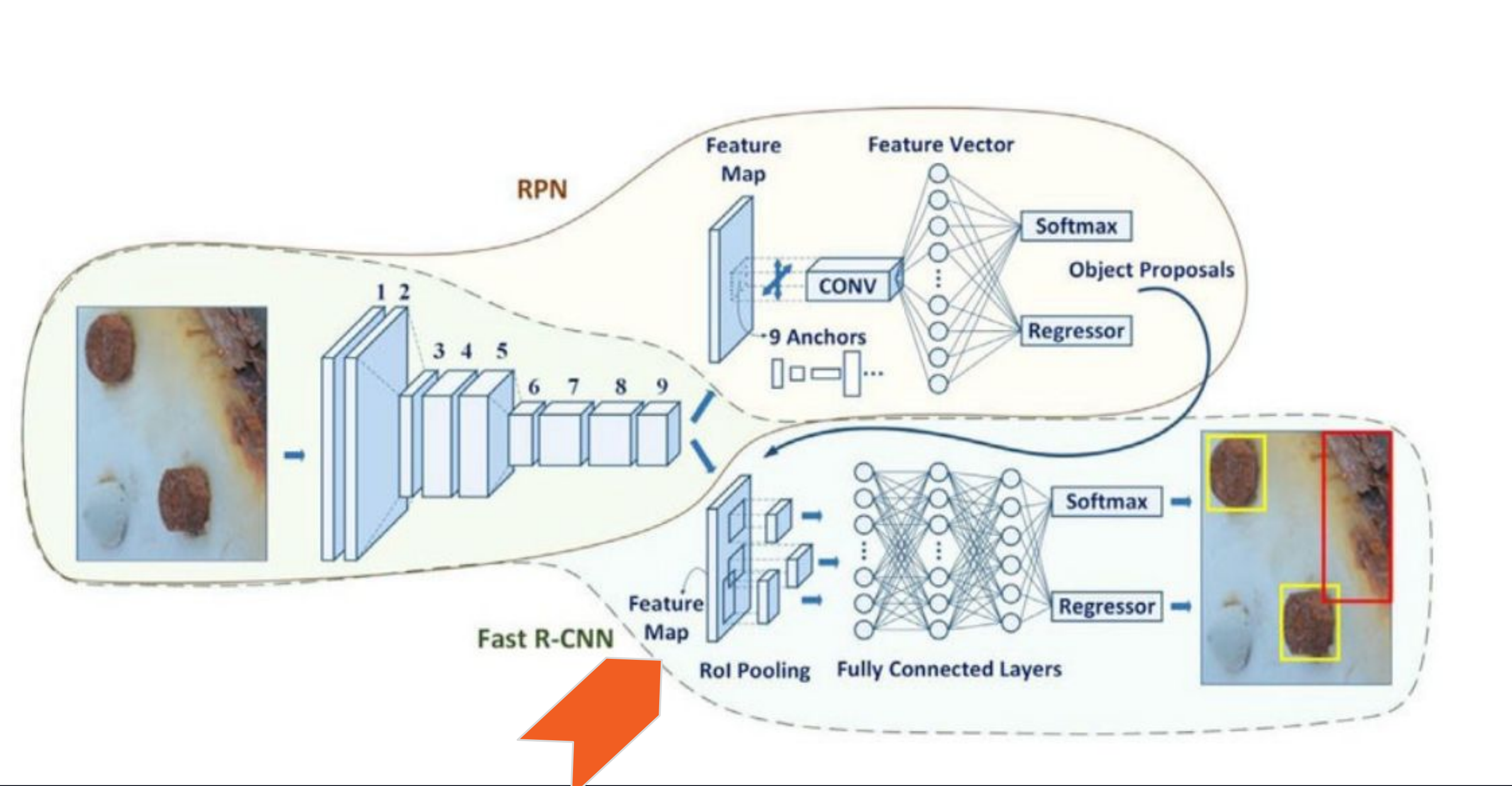
Alternating training

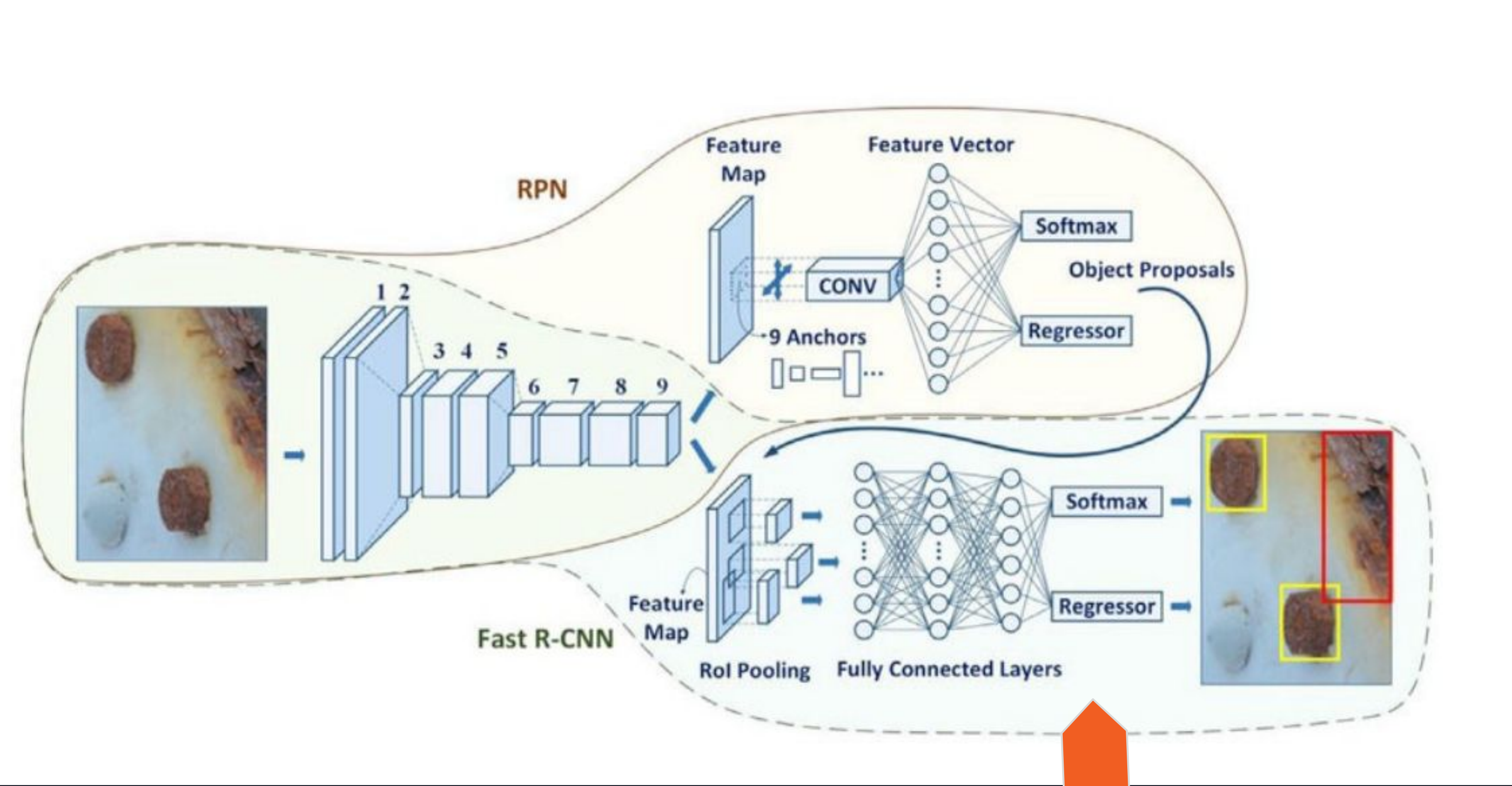
- Iter:
  - Train RPN
  - Train Fast R-CNN

FASTER R-CNN HIGH LEVEL VIEW



Faster R-CNN


RPN
Feature Map
Feature Vector
Softmax
CONV
Object Proposals
Regressor
9 Anchors
1 2
3 4 5
6 7 8 9
Feature Map
Fast R-CNN
RoI Pooling
Fully Connected Layers
Softmax
Regressor

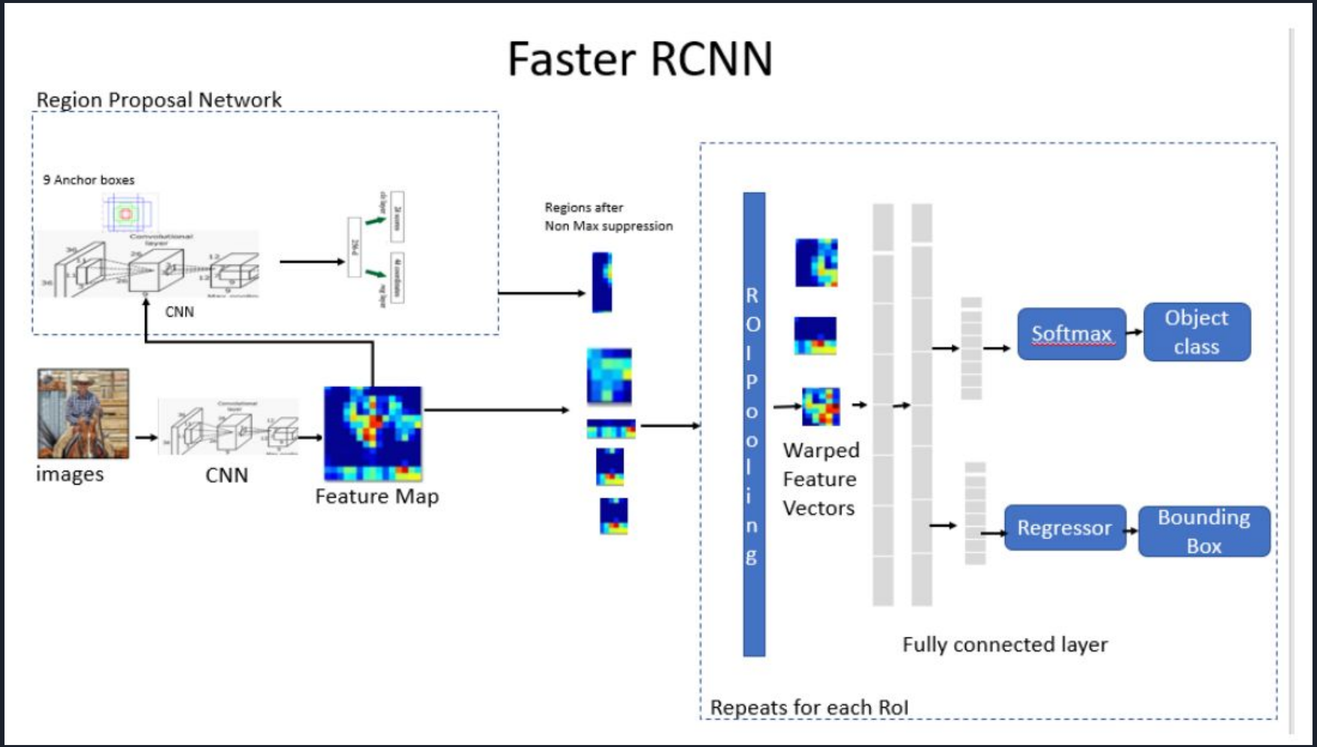


RPN
Feature Map
Feature Vector
Softmax
CONV
Object Proposals
9 Anchors
Regressor
1 2
3 4 5
6 7 8 9
Softmax
Regressor
Feature Map
Fast R-CNN
RoI Pooling
Fully Connected Layers

Presentation

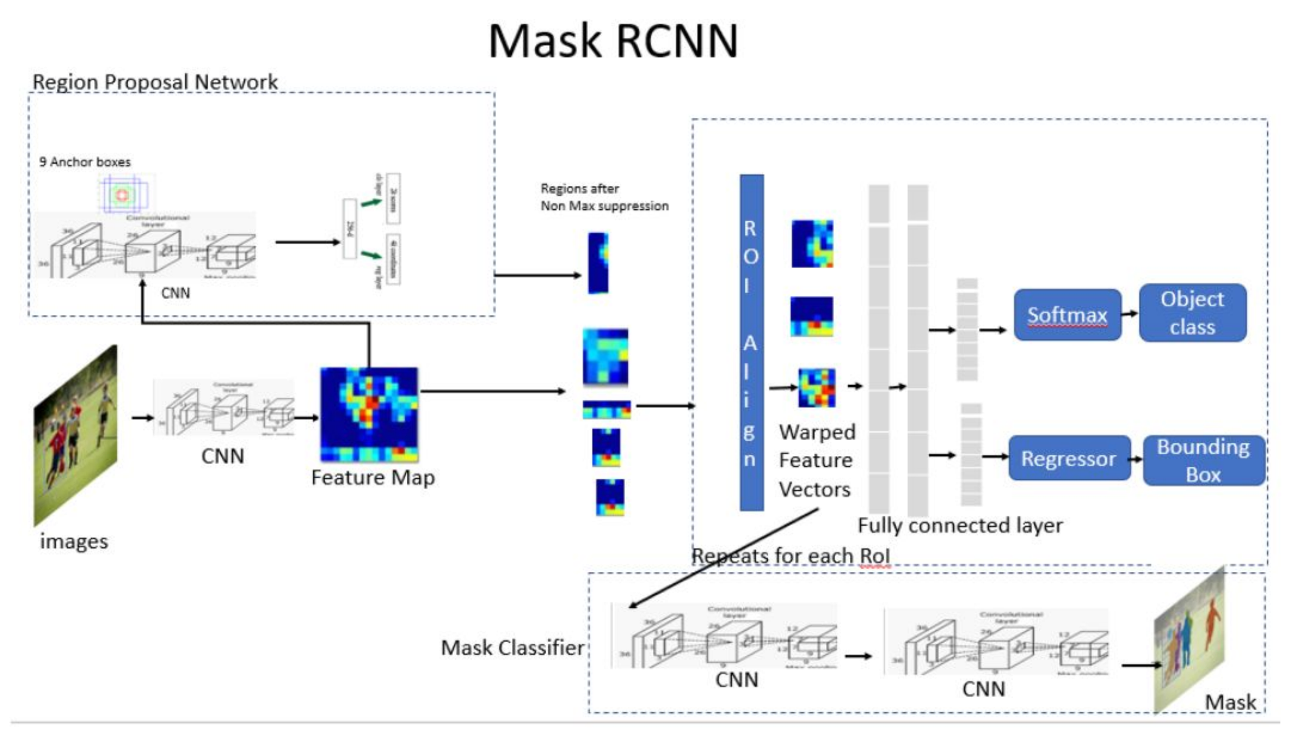
- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

MASK R-CNN



Faster RCNN
Region Proposal Network
9 Anchor boxes
CNN
Regions after
Non Max suppression
images
CNN
Feature Map
R
O
I
P
o
o
l
i
n
g
Warped
Feature
Vectors
Softmax
Object
class
Regressor
Bounding
Box
Fully connected layer
Repeats for each RoI

MASK R-CNN



Mask RCNN
Region Proposal Network
9 Anchor boxes
CNN
Regions after
Non Max suppression
R
O
I
A
l
i
g
n
Warped
Feature
Vectors
Softmax
Object
class
Regressor
Bounding
Box
images
CNN
Feature Map
Fully connected layer
Repeats for each RoI
Mask Classifier
CNN
CNN
Mask

MASK R-CNN

- Faster R-CNN has two outputs for each candidate object, a class label and a bounding-box offset

Mask R-CNN = Faster R-CNN + MASK BRANCH

Mask branch #1

- For each RoI Mask R-CNN also outputs a binary mask.
- Multitask Loss for each RoI:

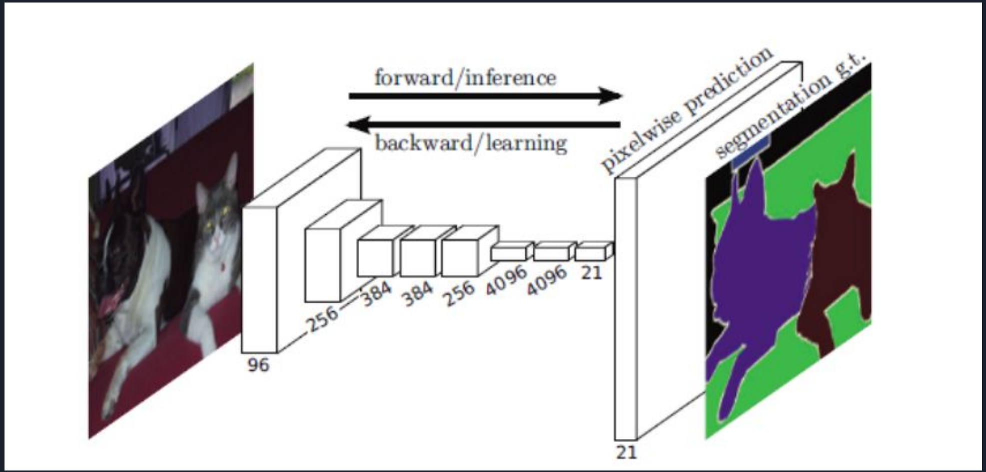
$$L = L_{cls} + L_{reg} + L_{mask}$$

Mask branch #2

- Output of mask branch: Km^2
- K classes
- m x m - mask resolution - The $m \times m$ floating-number mask output is then resized to the RoI size, and binarized at a threshold of 0.5.
- K'th class RoI - L_{mask} only defined for k'th mask (other masks don't contribute to the loss).
- Loss: average binary cross-entropy loss.
- There is no competition among classes.

Mask branch #3

- Spatial structure of masks can be addressed naturally by the pixel-to-pixel correspondence provided by convolutions.
- Masks are produced in use of FCNs: J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015.



RESULTS

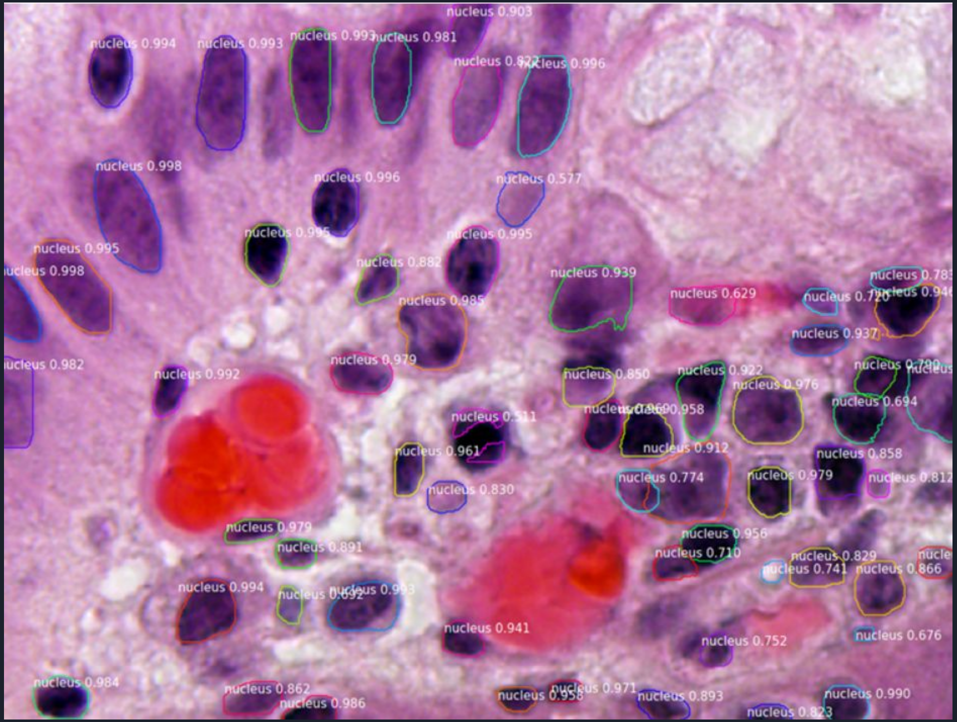
TASK	DATASET	MODEL	METRIC NAME	METRIC VALUE	GLOBAL RANK	RESULT	BENCHMARK
Object Detection	PASCAL VOC 2007	Faster R-CNN	MAP	73.2%	# 23		Compare
Real-Time Object Detection	PASCAL VOC 2007	Faster R-CNN	MAP	73.2%	# 4		Compare
			FPS	7	# 5		Compare

TASK	DATASET	MODEL	METRIC NAME	METRIC VALUE	GLOBAL RANK	RESULT	BENCHMARK
Nuclear Segmentation	Cell17	Mask R-CNN	F1-score	0.8004	# 3		Compare
			Dice	0.707	# 3		Compare
			Hausdorff	12.6723	# 3		Compare
Panoptic Segmentation	Cityscapes val	Mask R-CNN+COCO	PQth	54.0	# 12		Compare
Keypoint Detection	COCO	Mask R-CNN	Validation AP	69.2	# 7		Compare
			Test AP	63.1	# 9		Compare

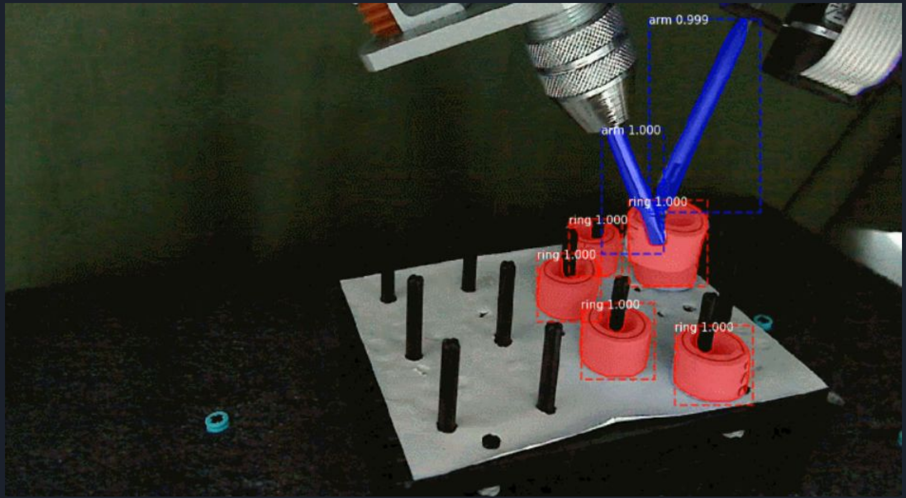
Presentation

- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

EXTRA USE CASES



EXTRA USE CASES



Presentation

- TARGETS
- FASTER R-CNN
- MASK R-CNN
- EXTRA USE-CASES

THANK YOU!

BIBLIOGRAPHY

- https://arxiv.org/abs/1411.4038
- https://arxiv.org/abs/1703.06870
- https://arxiv.org/abs/1506.01497
- https://towardsdatascience.com/computer-vision-instance-segmentation-with-mask-r-cnn-7983502fcad1
- https://towardsdatascience.com/faster-rcnn-object-detection-f865e5ed7fc4
- https://towardsdatascience.com/computer-vision-a-journey-from-cnn-to-mask-r-cnn-and-yolo-1d141eba6e04