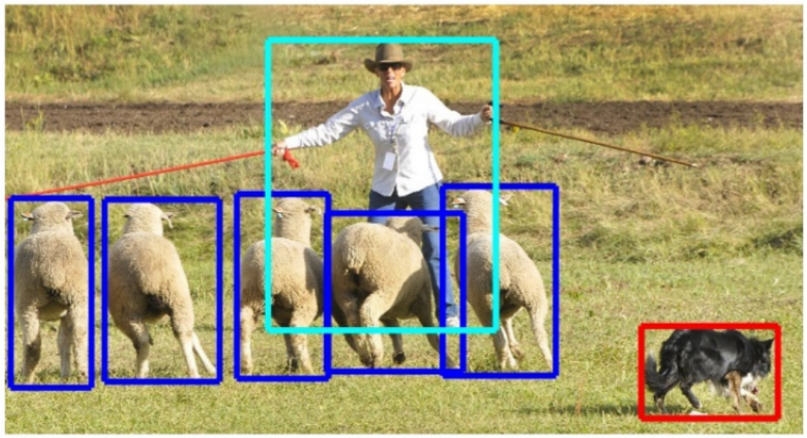
Mask R-CNN

Kaiming He, Georgia Gkioxari, Piotr Dollar,
Ross Girshick,

Facebook AI Research (FAIR)

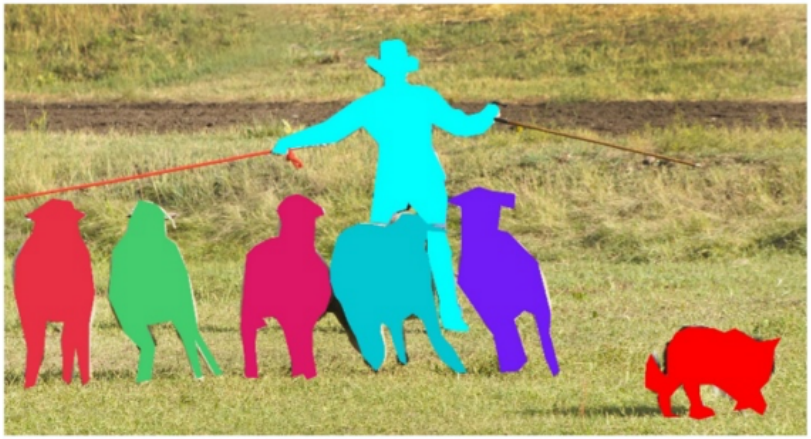
Classification, Detection and Segmentation

Classification



Object detection

Semantic segmentation



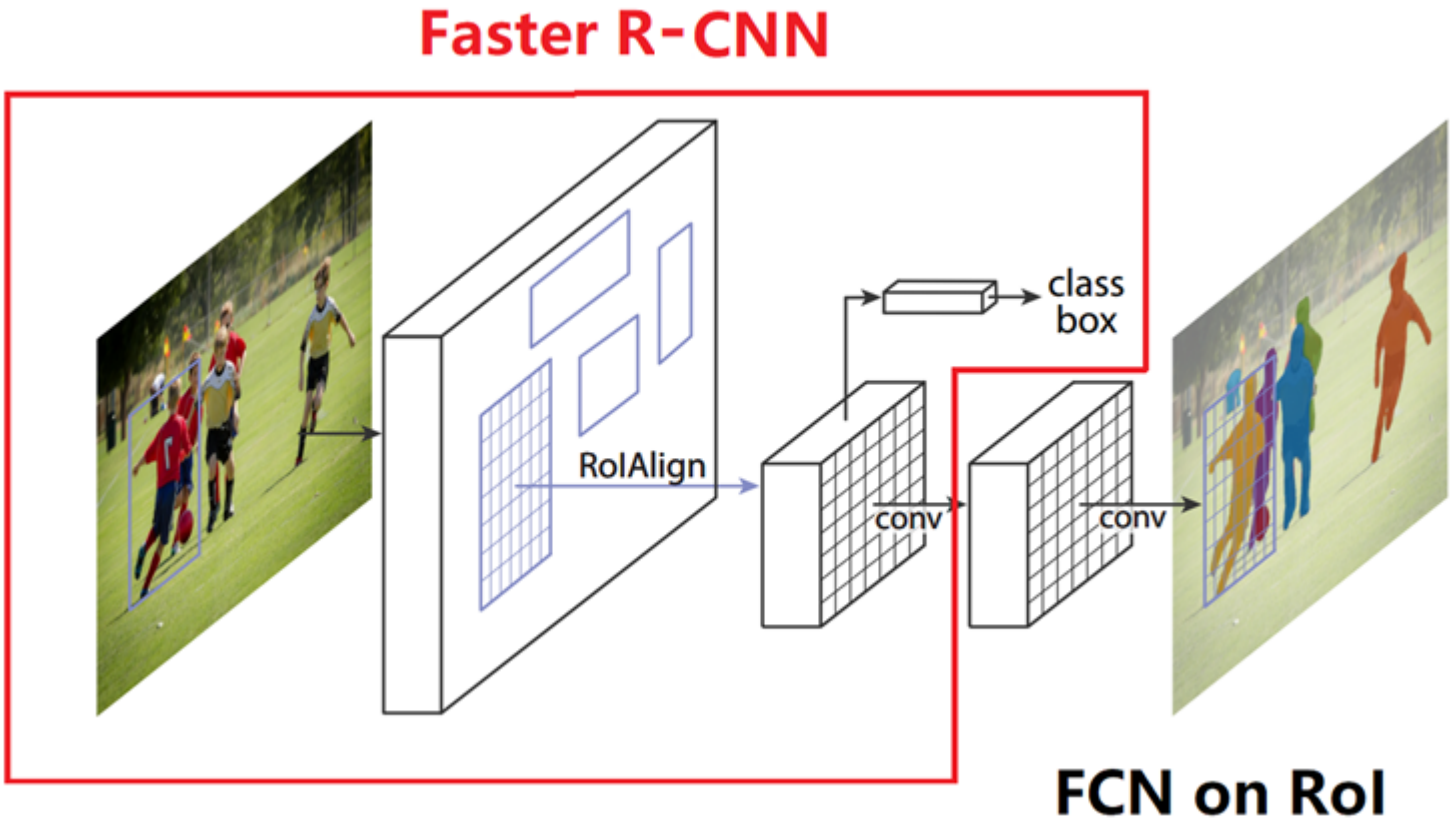
Instance segmentation

Overview of Mask R-CNN

- Goal: to create a framework for instance segmentation
- Builds on top of Faster R-CNN by adding a parallel branch
- Predicts segmentation mask using a small FCN for each Region of Interest (RoI)
- Generates a binary mask for each class independently: decouples segmentation and classification
- Changes RoIPool in Faster R-CNN to a quantization-free layer called RoIAlign
- Result: performs better than state-of-the-art models in instance segmentation, bounding box detection and person keypoint detection
- Easy to generalize to other tasks: Human pose detection

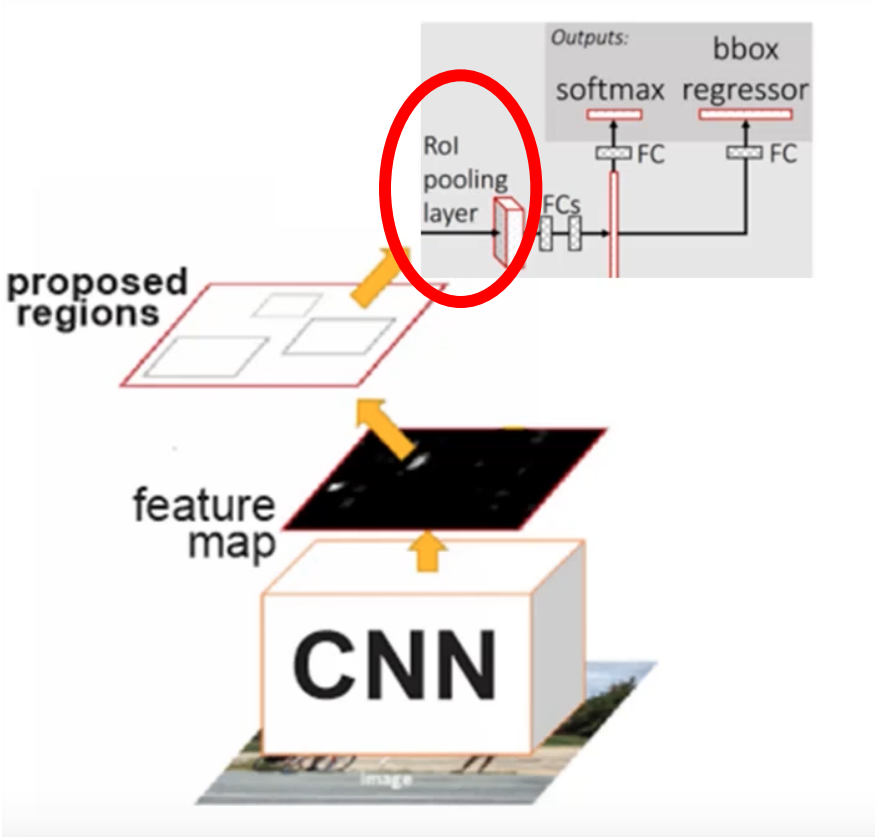
Architecture

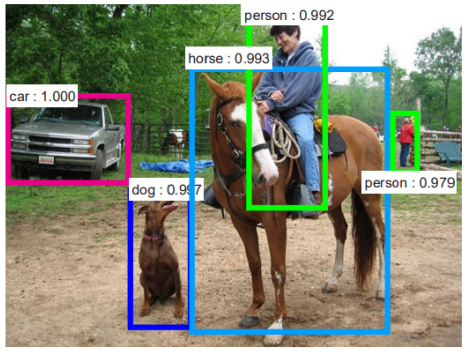
- Mask R-CNN ~ Faster R-CNN + FCN

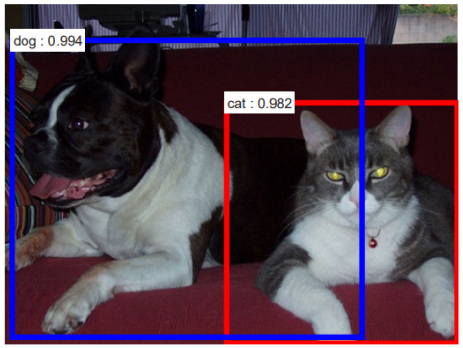


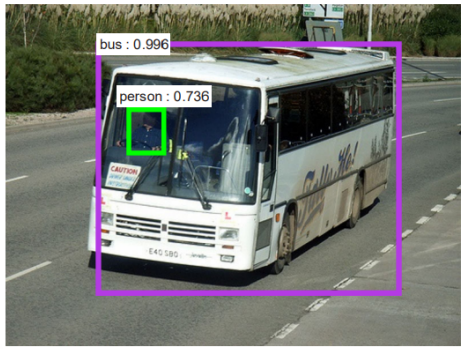
Architecture

- Faster R-CNN





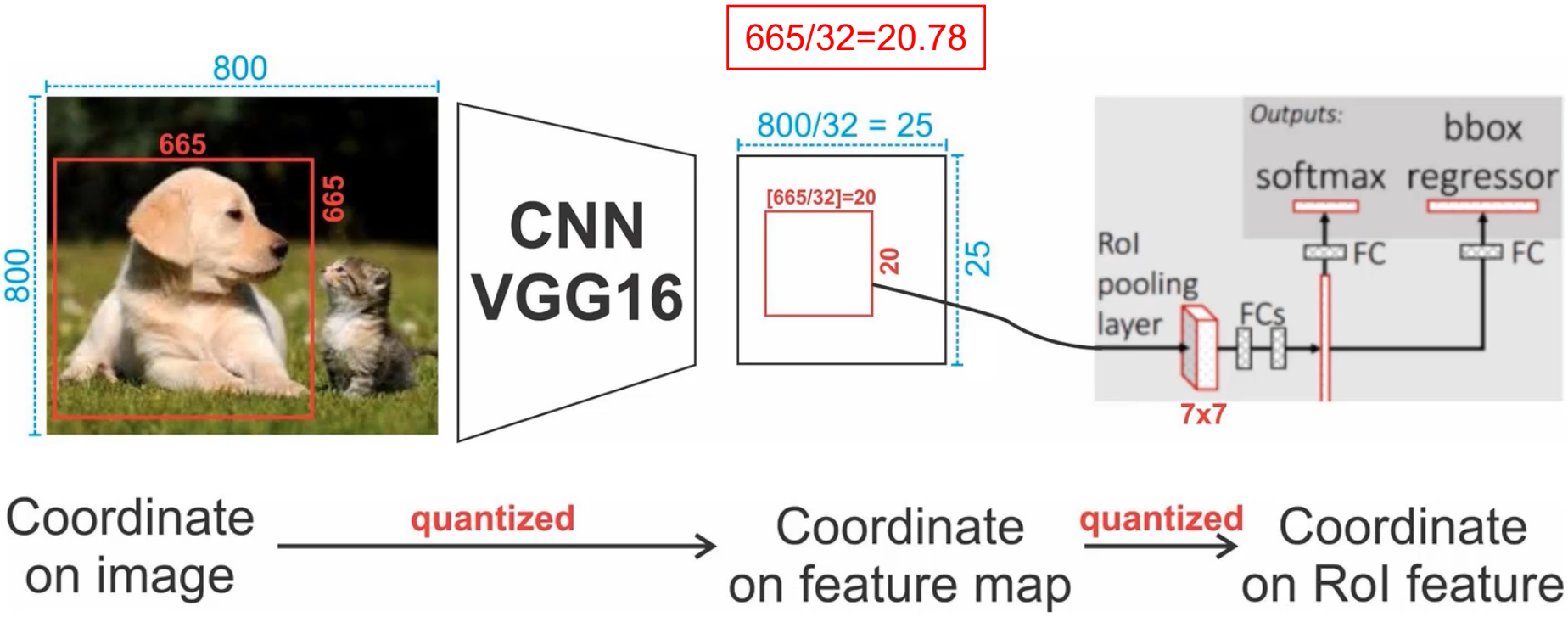




Ren et al,, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, 2015

Architecture

- Modification #1: RoIPool changed RoIAlign

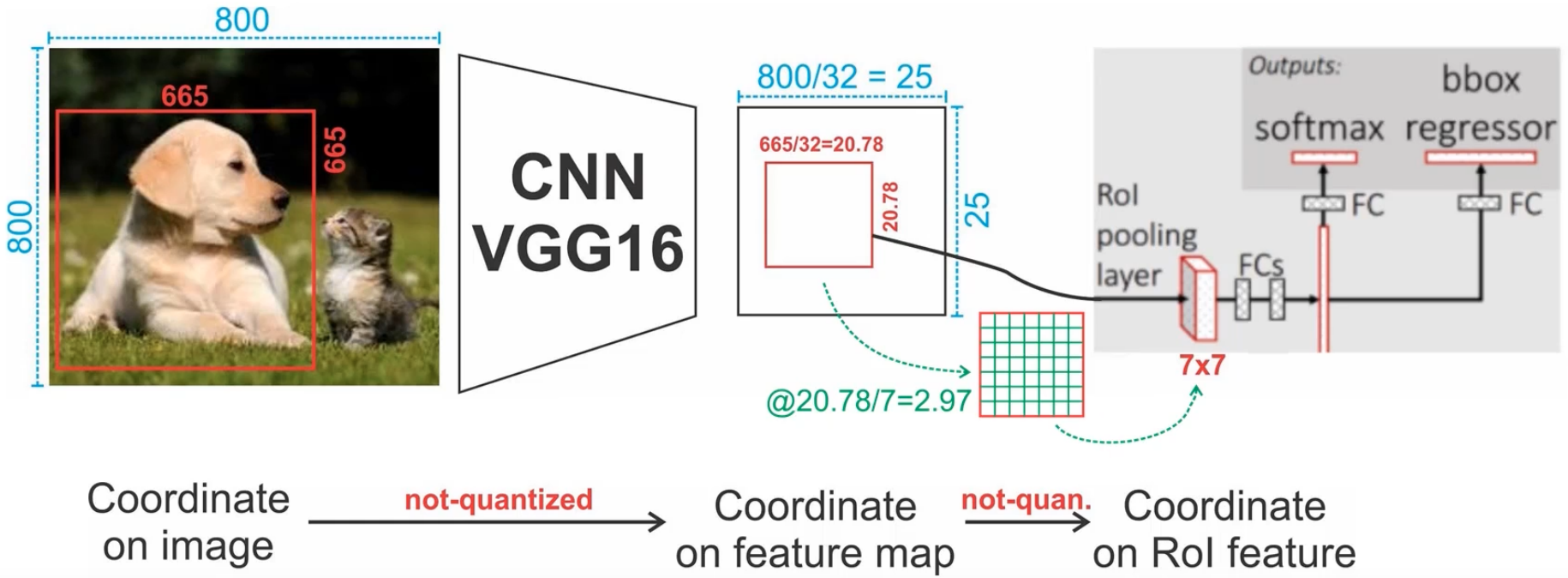


RoIPool

https://www.youtube.com/watch?v=UI25zSysk2A&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

Architecture

- Modification #1: RoIPool changed RoIAlign

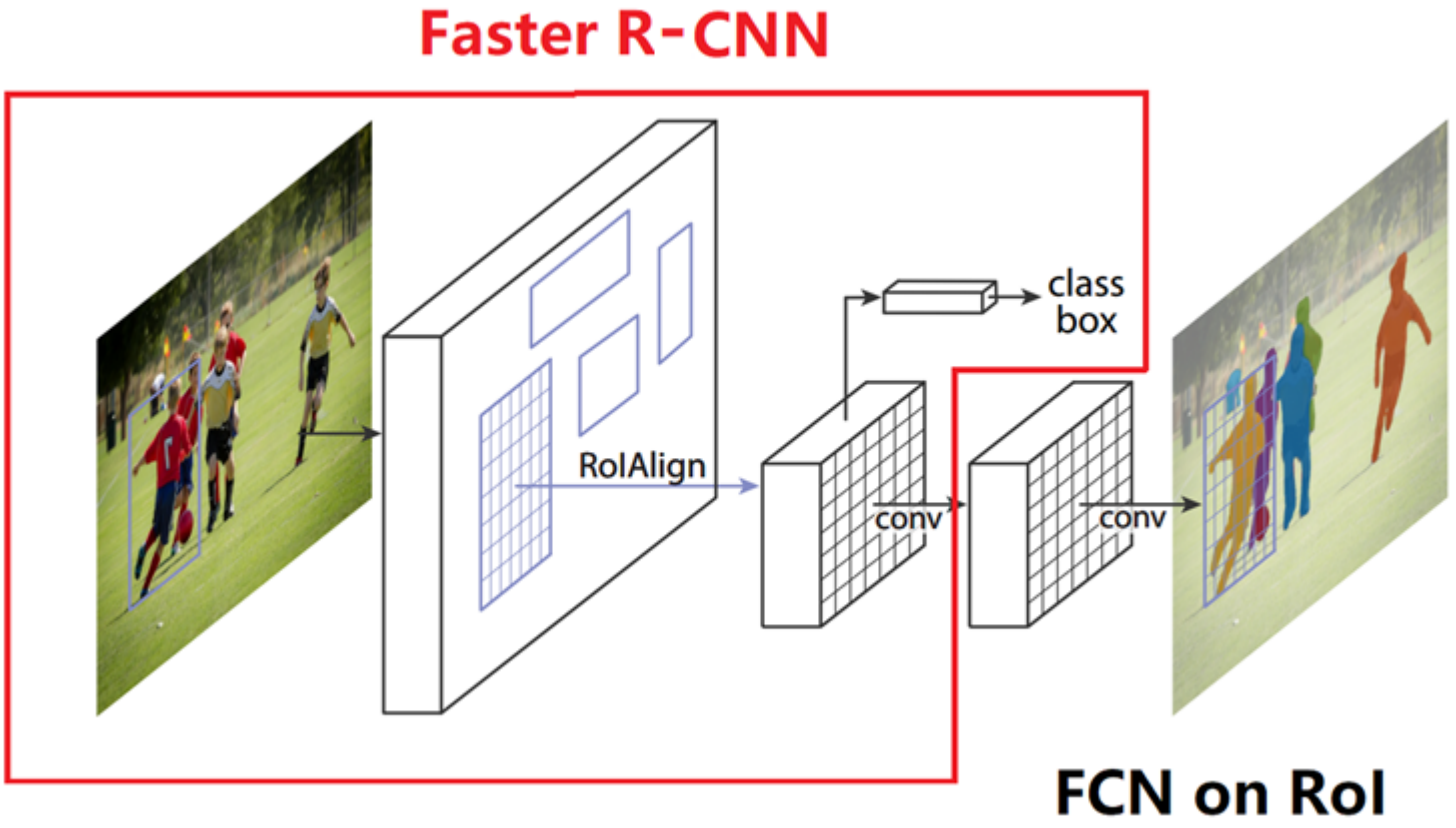


RoIAlign

https://www.youtube.com/watch?v=UI25zSysk2A&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

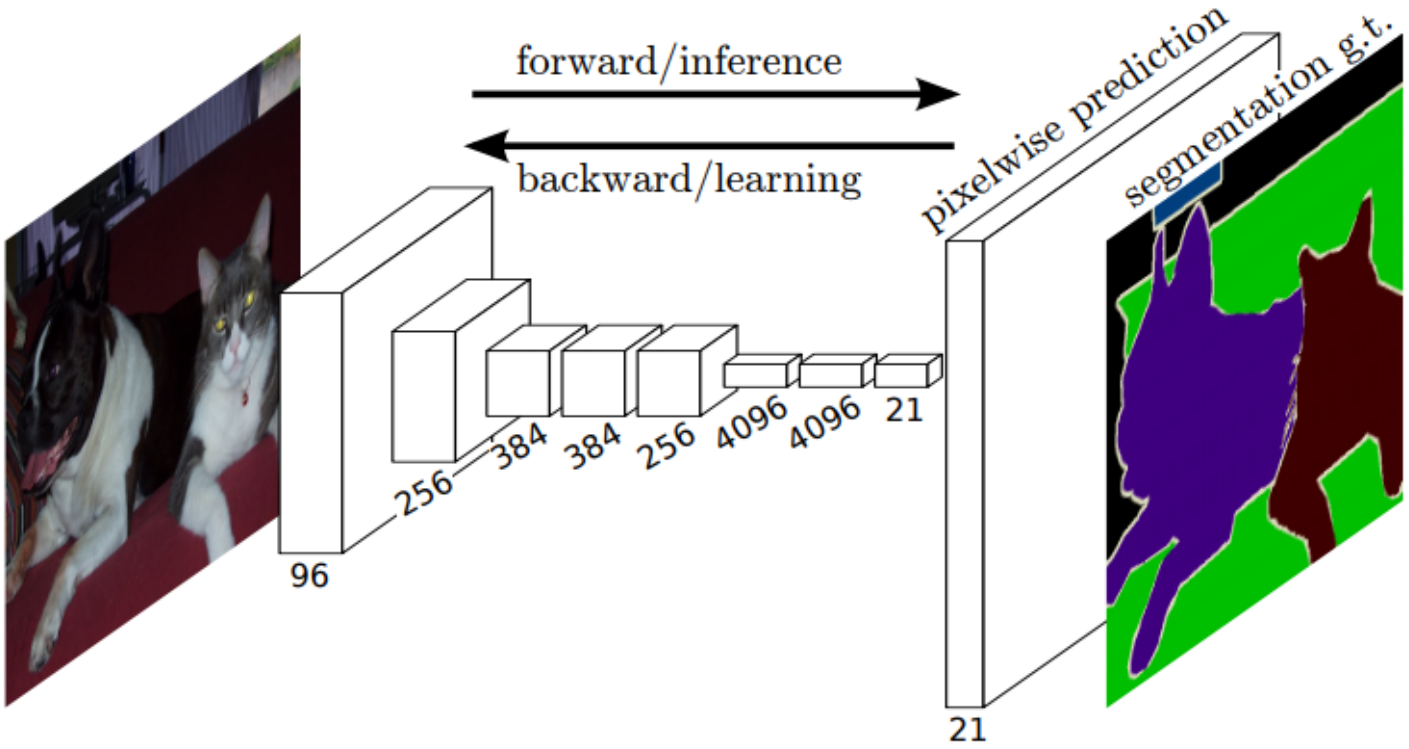
Architecture

- Mask R-CNN ~ Faster R-CNN + FCN



Architecture

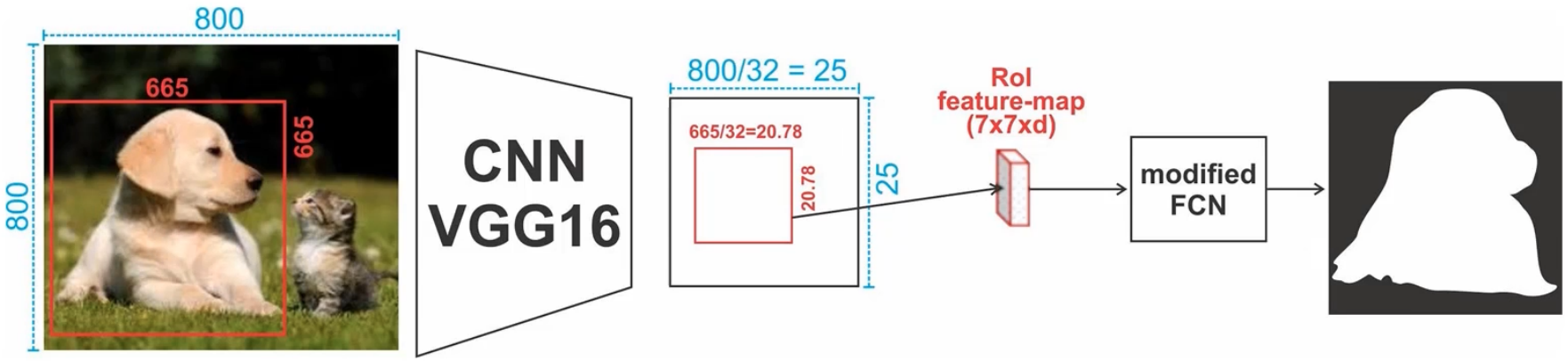
- FCN



Long et al,, Fully Convolutional Networks for Semantic Segmentation, 2014

Architecture

- Modification #2: Decouple mask and classification prediction
  - Predict binary mask for each class
  - Mask branch can predict K (K= number of classes) masks per RoI. Only use the k-th mask, where k is the predicted class by the classification branch



https://www.youtube.com/watch?v=UI25zSysk2A&list=PLkRkKTC6HZMxZrxnHUDYSLiPZxiUUFD2C

Architecture

- Mask R-CNN ~ Faster R-CNN + FCN
  - Faster R-CNN with RoIAlign
  - FCN for binary mask
  - Parallel prediction for the class, BB and binary mask for each RoI
- Loss function for each sampled RoI

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

Implementation details

a. Hyperparameter : using existing Faster R-CNN
b. Backbone architectures : ResNet50, ResNet101, ResNeXt50 ResNeXt101, FPN (Feature Pyramid Networks)
c. Input image : resized into 800px for its shorter size
d. GPU : 8 GPU @2 images on training
e. Training time : 32 hours (ResNet50-FPN), 44 hours (ResNet101-FPN), not end-to-end training.
f. Testing time (ResNet101-FPN): 195ms per image on an Nvidia Tesla M40 GPU (plus 15ms CPU time resizing the outputs to the original resolution)
g. Dataset : MS COCO (80k train, 35k val, 5k test)

Results

- Object detection

	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**

Results

- Object detection

Results

- Ablation experiments

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**

	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*

	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*

Results

- Human pose estimation

	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}	AP^{kp}_{M}	AP^{kp}_{L}
CMU-Pose+++ [6]	61.8	84.9	67.5	57.1	68.2
G-RMI [32]†	62.4	84.0	68.5	**59.1**	68.1
Mask R-CNN, keypoint-only	62.7	87.0	68.4	57.4	71.1
Mask R-CNN, keypoint & mask	**63.1**	**87.3**	**68.7**	57.8	**71.4**

Conclusion

- Mask R-CNN outperforms all existing, single-model entries on every task, including the COCO 2016 challenge winners.
- Keys of improvements: (i) RoIAlign, (ii)decouple mask with class prediction that generates binary mask, and (iii) parallel object detection and mask generation

FYI

https://github.com/matterport/Mask_RCNN