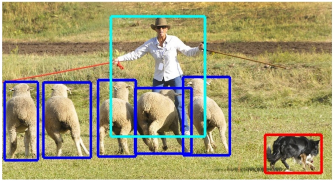
Instance segmentation with Mask R-CNN

Seminar: Current Works in Computer Vision

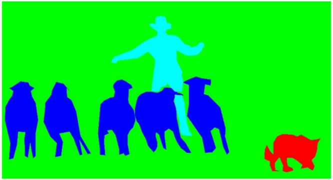
Silvio Galesso

Scene understanding

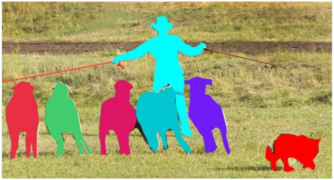
Image classification



Object detection



Semantic segmentation



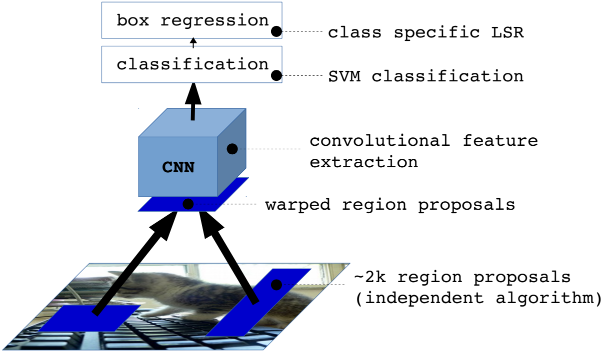
Instance segmentation

Source: Lin et al., Microsoft COCO: Common Objects in Context

Mask R-CNN: Motivation and goals

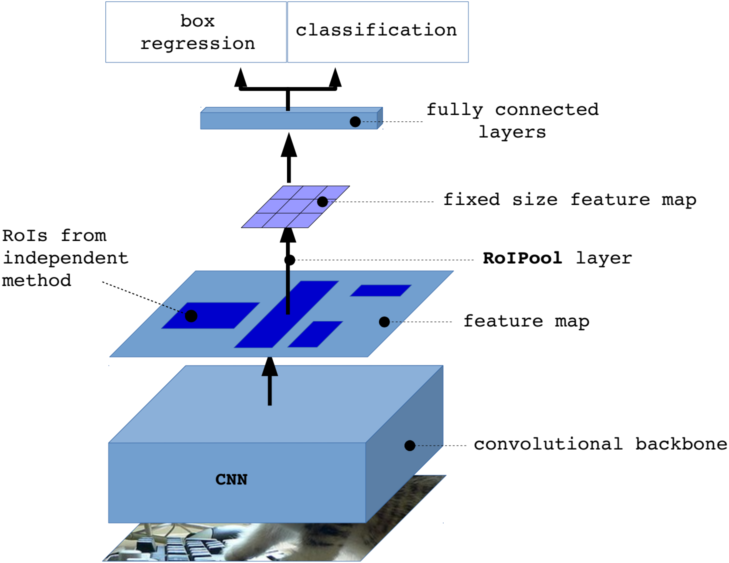
- State of the art multitask model for visual scene understanding:
  - object detection
  - classification
  - instance segmentation
- Highly modular and easy to train
- Flexible: e.g. human pose estimation with minor changes

Background: R-CNN architechture

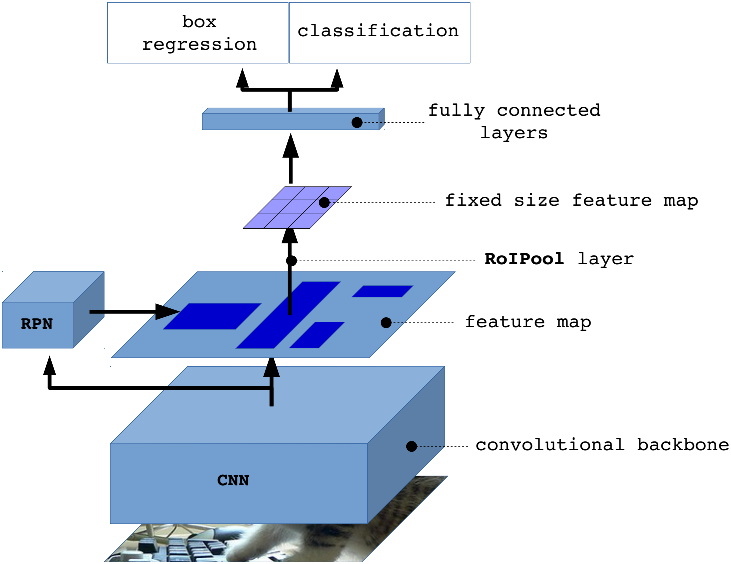


- Based on proposed Regions of Interest (RoI)
- Requires region warping for fixed size features
- Very inefficient pipeline

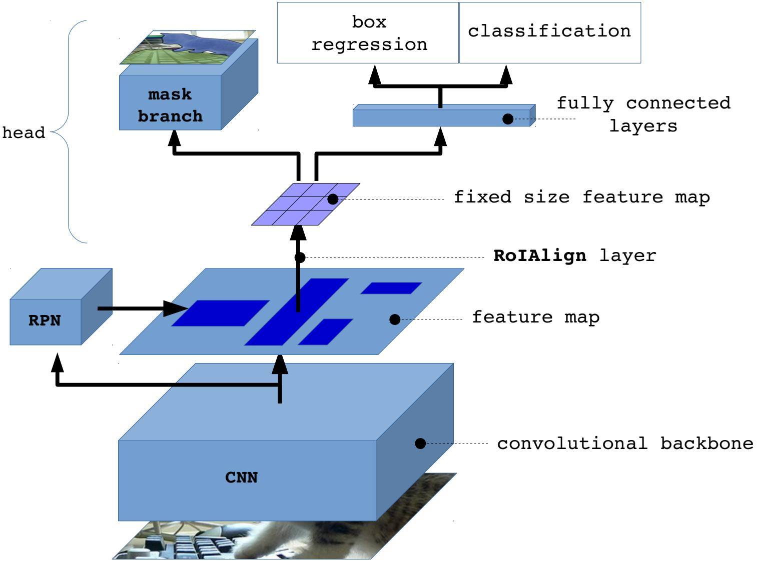
Background: Fast R-CNN



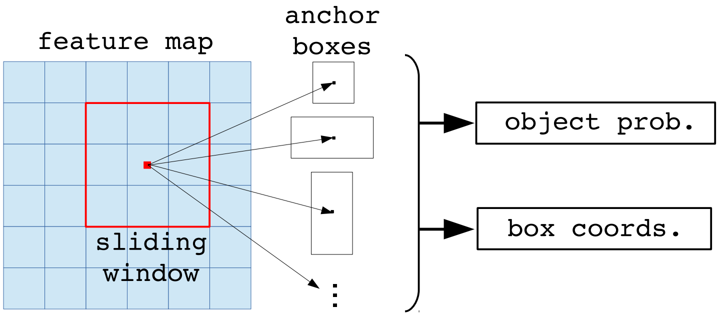
Background: Faster R-CNN



Mask R-CNN: overview

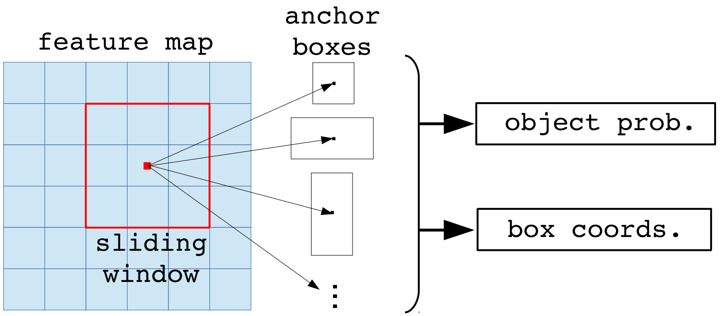


Region Proposal Network



- Shared conv layers with main model
- $n \times n$ sliding window, large receptive field
- Parallel branches:
    - object probability classification
    - box regression

Region Proposal Network

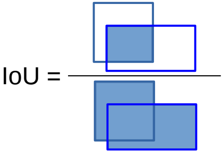


Anchor boxes:

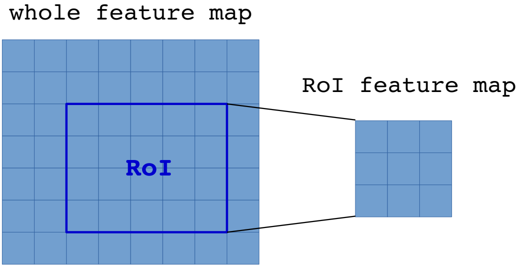
- for every window position, k region prototypes
- multiple scales, e.g. 128^2, 256^2, 512^2
- multiple ratios, e.g. $1:1$, $1:2$, $2:1$

Region Proposal Network

- Multiple anchor scales and ratios $\rightarrow$ single scale images
- Proposal evaluation based on Intersection over Union with ground truth boxes:
  - best regions are kept as positive examples
  - worst (IoU < 0.3) are kept as negatives for training



RoI feature extraction



- RoIPool: quantized bins + pooling
- RoIAlign: continuous bins + bilinear interpolation + pooling
  $\Rightarrow$ better preserved spatial correspondence

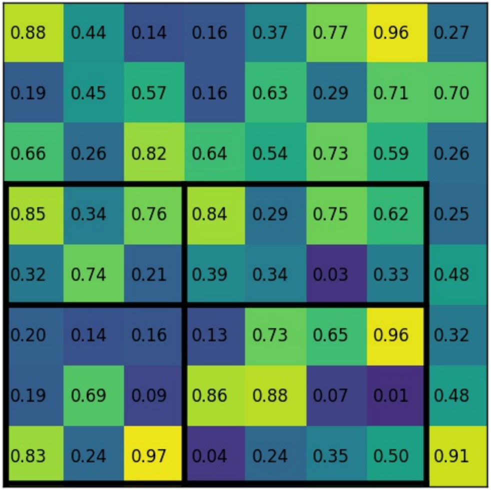
RoIPool (Faster R-CNN)

Source: https://deepsense.io/region-of-interest-pooling-explained/

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

Input activation

RoIPool (Faster R-CNN)



Region projection and pooling sections

RoIPool (Faster R-CNN)

0.85	0.84
0.97	0.96

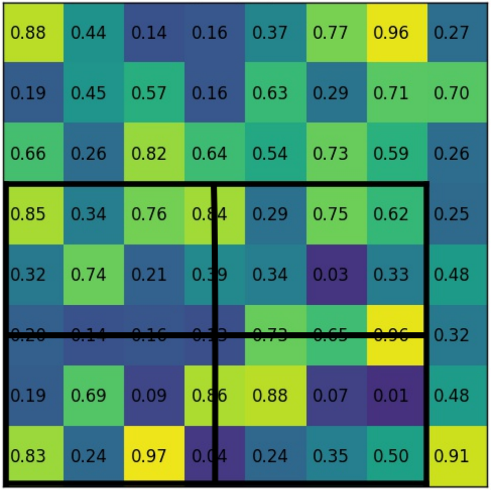
Max pooling output

RoIAlign (Mask R-CNN)

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

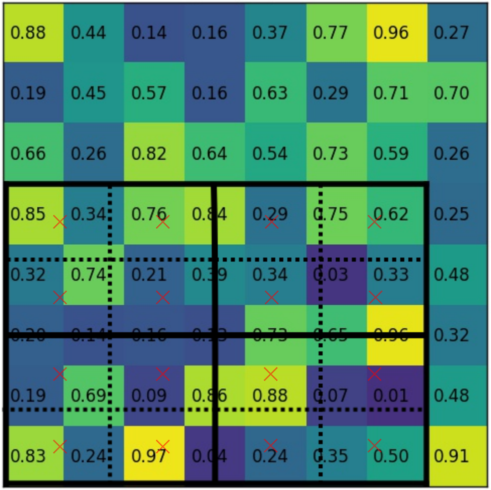
Input activation

RoIAlign (Mask R-CNN)



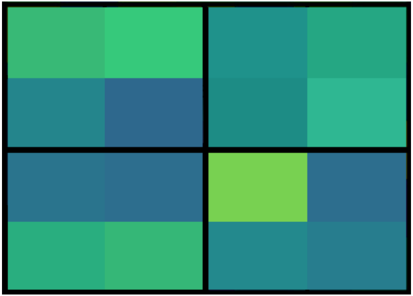
Region projection and pooling sections

RoIAlign (Mask R-CNN)



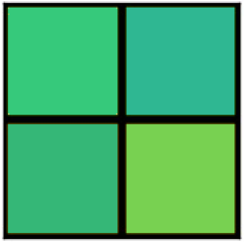
Sampling locations

RoIAlign (Mask R-CNN)



Bilinear interpolated values

RoIAlign (Mask R-CNN)



Max pooling output

Class prediction & box regression

- Fully connected branch:
  - $K+1$ softmax for classification
  - $4 \cdot K$ box regression targets:
    $\mathbf{t}^k = (t_x^k, t_y^k, t_w^k, t_h^k)$

- Multitask loss:
  - L_{cls}: negative log likelihood
  - L_{reg}: smooth $L1$ loss

$$L_{box} = L_{cls} + \lambda \mathbf{1}_{[u=u^*]} L_{reg}$$

Segmentation

Mask branch features:

- Fully convolutional
- $K \cdot (m \times m)$ sigmoid outputs:
  - $\rightarrow$ pixel-wise binary classification
  - $\rightarrow$ one mask for each class, no competition
- L_{mask}: mean binary cross-entropy

Overall head loss:

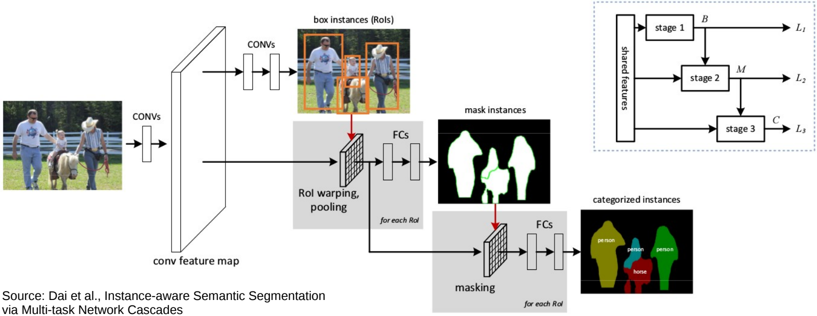
$$L = L_{box} + L_{mask}$$

Experiments

Dataset & metrics

- Main dataset: MS COCO
  - 80 classes
  - 115k training images
  - 5k images for ablation experiments
  - undisclosed ground truth `test-dev` for main results
- Similarity measure: Intersection over Union (IoU)
- AP_{50} & AP_{75} (PASCAL VOC metrics):
  Average Precision: IoU threshold $(.50, .75)$ for true positives
  $\rightarrow$ precision-recall curve $\rightarrow$ area under curve
- AP (MS COCO metric):
  mean Average Precision over different IoU thresholds.

Related methods: MNC

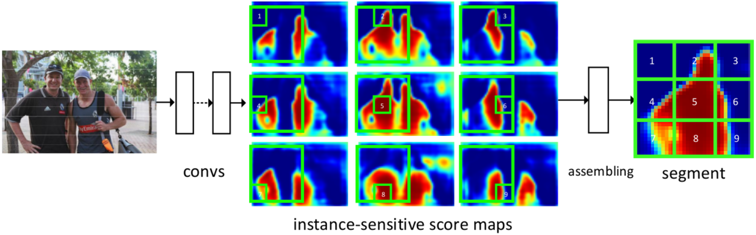


Source: Dai et al., Instance-aware Semantic Segmentation via Multi-task Network Cascades

Multi-task **N**etwork **C**ascade:

- bounding box regression
- mask estimation
- classification

Related methods: FCIS



Source: Dai et al., Instance-sensitive Fully Convolutional Networks

- **F**ully **C**onvolutional **I**nstance **S**egmentation
- Challenge: translation invariance $\rightarrow$ no instance awareness
- Proposed solution: positional aware sliding masks

Segmentation results

Model	backbone	AP	AP_{50}	AP_{75}
MNC	ResNet-101-C4	24.6	44.3	24.8
FCIS+++	ResNet-101-C5-dil.	33.6	54.5	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**

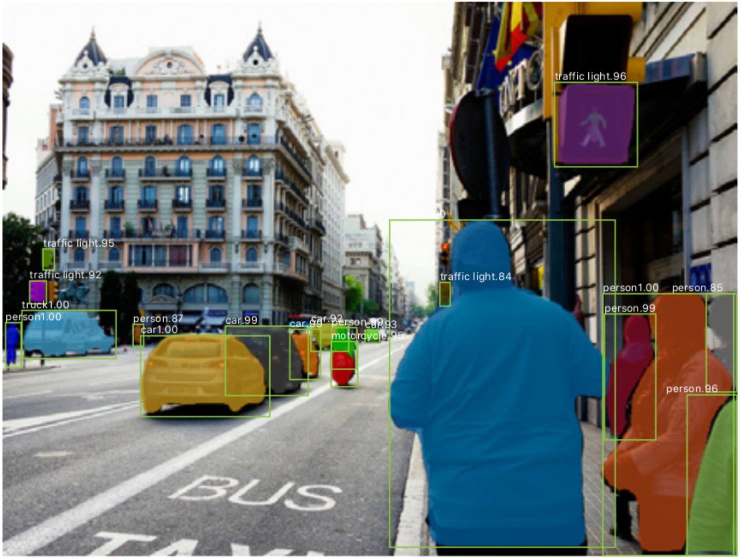
FPN (Feature Pyramid Network):
multi-scale hierarchical convolutional features, good for detection

Cityscape results

Cityscape dataset: smaller ($5k$ `fine` + $20k$ `coarse`), urban scenery for segmentation

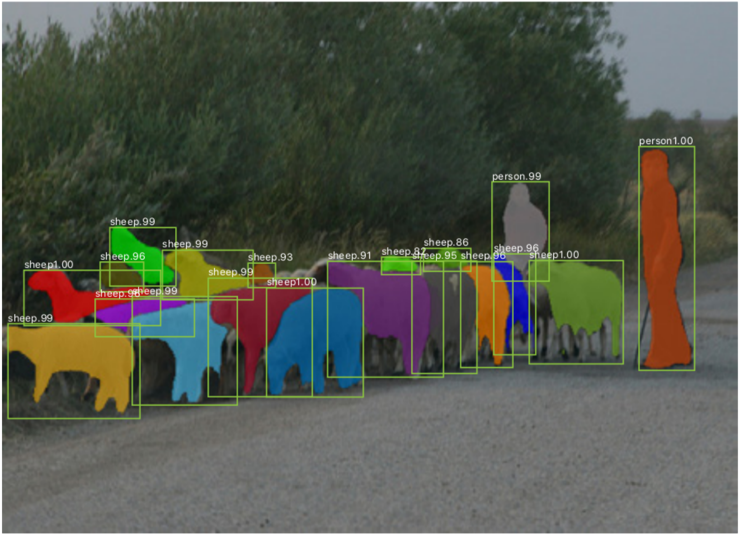
Model	training set	AP	AP_{50}
SAIS	`fine`	17.4	36.7
DIN	`fine+coarse`	20.0	38.8
Mask R-CNN	`fine`	26.2	49.9
Mask R-CNN	`fine`+COCO pretrain	**32.0**	**58.1**

Example results

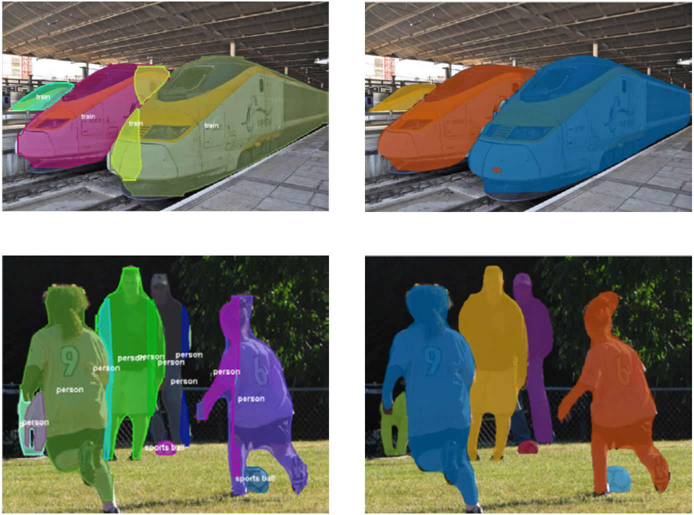


Example results

Example results



FCIS vs Mask R-CNN



Detection results

Model	backbone	AP^{bb}	AP^{bb}_{50}	AP^{bb}_{75}
Faster R-CNN	ResNet-101-FPN	36.2	59.1	39.0
Faster R-CNN	same + RoIAlign	37.3	59.6	40.3
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**

- Backbone (ResNeXt): $+1.6 AP^{bb}$
- RoiAlign: $+1.1 AP^{bb}$
- Multitask training: $+0.9 AP^{bb}$

Ablation experiments

Backbone architecture:

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**

Ablation experiments

RoIAlign layer:

	stride 16			stride 32*		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
RoIPool	26.9	48.8	26.4	23.6	46.5	21.6
RoIAlign	**30.2**	**51.0**	**31.8**	**30.9**	**51.8**	**32.1**

* larger stride means larger misalignments

Ablation experiments

Mask branch independence:

last fc layer	AP	AP_{50}	AP_{75}
softmax + multinomial loss	24.8	44.1	25.1
sigmoid + binomial loss	**30.3**	**51.2**	**31.5**
	+5.5	+7.1	+6.4

Human pose estimation

- Task: localize anatomical keypoints
- K keypoint types (e.g. left shoulder, right elbow)
  $\rightarrow K$ *one-hot masks* $\rightarrow$ cross-entropy loss, m^2 softmax
- No other domain knowledge employed

Model	AP^{kp}	AP^{kp}_{50}	AP^{kp}_{75}
CMU-Pose+++	61.8	84.9	67.5
G-RMI	62.4	84.0	68.5
Mask R-CNN, keypoint only	62.7	87.0	68.4
Mask R-CNN, keypoint & mask	**63.1**	**87.3**	**68.7**

Examples

Examples

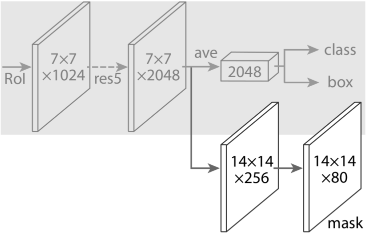
Wrap-up

- Mask R-CNN does detection, classification and instance segmentation.
- Based on Faster R-CNN + mask branch, RoIAlign
- State of the art detection and instance segmentation on MS COCO and Cityscapes
- Can do human pose estimation with small adaptations

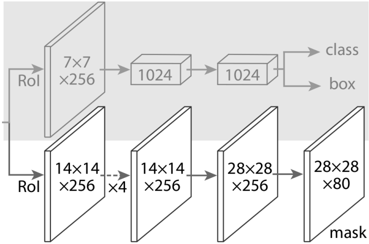
Literature

- K. He, G. Gkioxari, P. Dollr, R. Girshick. Mask R-CNN. ArXiv, Mar. 2017.
- S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In NIPS, 2015.
- R. Girshick. Fast R-CNN. In ICCV, 2015.
- R. Girshick, J. Donahue, T. Darrell and J. Malik. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. In IEEE TPAMI, vol. 38, no. 1, pp. 142-158, Jan. 1 2016.
- Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei. Fully convolutional instance-aware semantic segmentation. In CVPR, 2017.
- J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In ECCV, 2016.
- J. Dai, K. He, and J. Sun. Instance-aware semantic segmentation via multi-task network cascades. In CVPR, 2016.
- T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft COCO: Common objects in context. In ECCV, 2014.
- M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In CVPR, 2016.
- T. Lin, P. Dollr, R. Girshick, K. He, B. Hariharan, S. Belongie. Feature Pyramid Networks for Object Detection. In CoRR, 2016.

Mask R-CNN head architectures



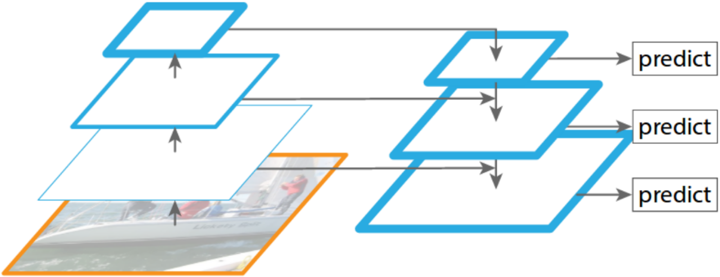
Extended Faster R-CNN head, on ResNet-C4 feature map



Extended Faster R-CNN head, on FPN feature map

Feature Pyramid Network

FPN exploits the inherent heirarchy of CNNs to compute multi-scale features:



Source: Lin et al., Feature Pyramid Networks for Object Detection