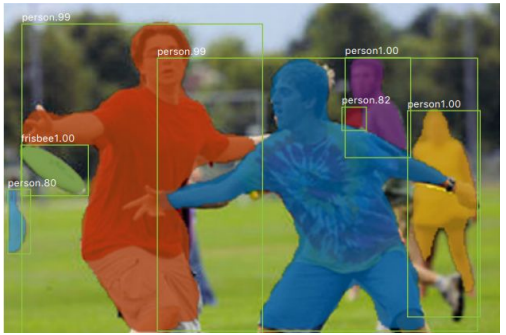
Mask R-CNN

Presenter: Zichao Hu

08/30/2022

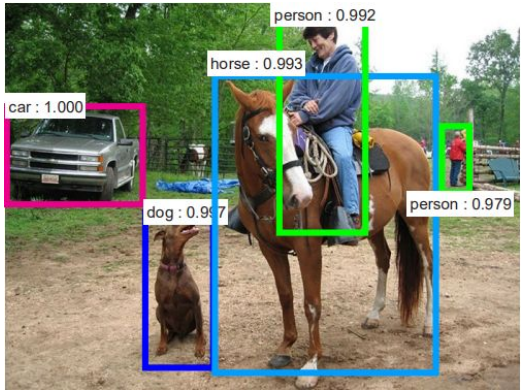
What is Mask R-CNN

- Mask R-CNN [1] is framework to solve the instance segmentation problem.
  - Instance segmentation is a task of detecting and delineating each object in an image in a fine-grained pixel level
  - Instance segmentation can estimate object position given an image, so tasks such as robot manipulation can perform grasp planning
- It is an extension of the Faster R-CNN framework.



Instance Segmentation: Mask R-CNN

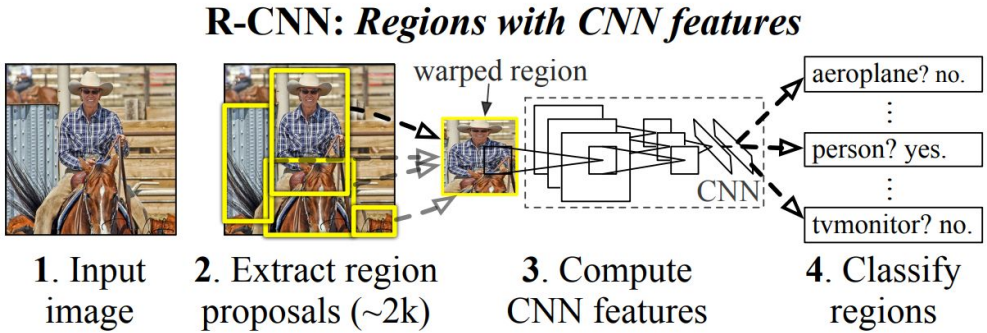
Background: R-CNN, Fast R-CNN, Faster R-CNN



Object Detection: R-CNN, Fast R-CNN, Faster R-CNN [2][3][4]

R-CNN (Region CNN)

- Region proposal: selective search
- Warp each region to a fixed size and fit through a convnet
- SVM and bounding box regression
- Slow, because computation for cnn forward pass is not shared



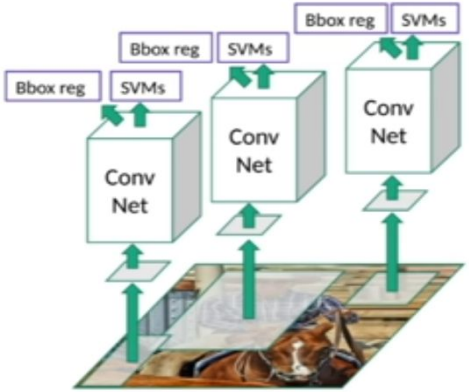
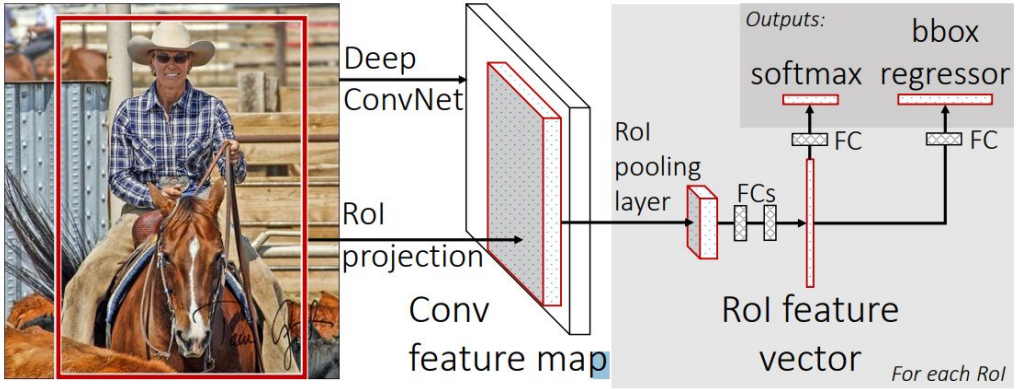
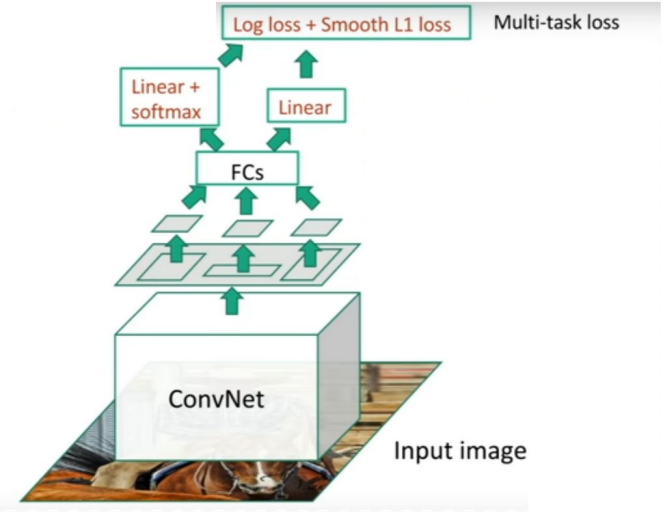


Image excerpted from [5]

Fast R-CNN

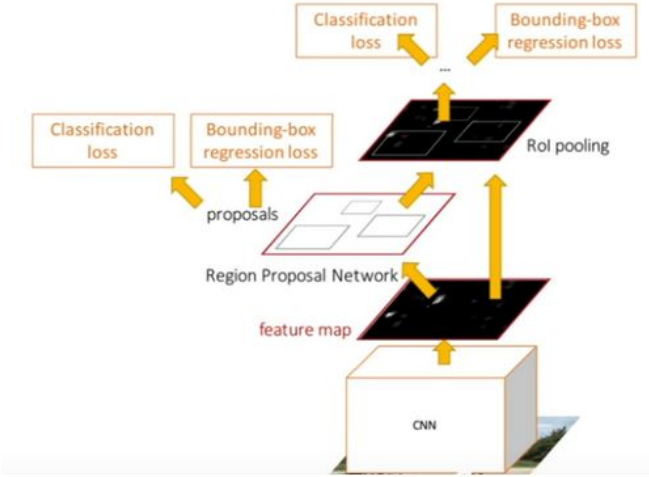
- Switch the order of region proposal and convnet
- Project the region proposal onto the latent space
  - Region of Interest (RoI), RoI pooling layer
- Fast, but the region proposal algorithm becomes the bottleneck





Faster R-CNN

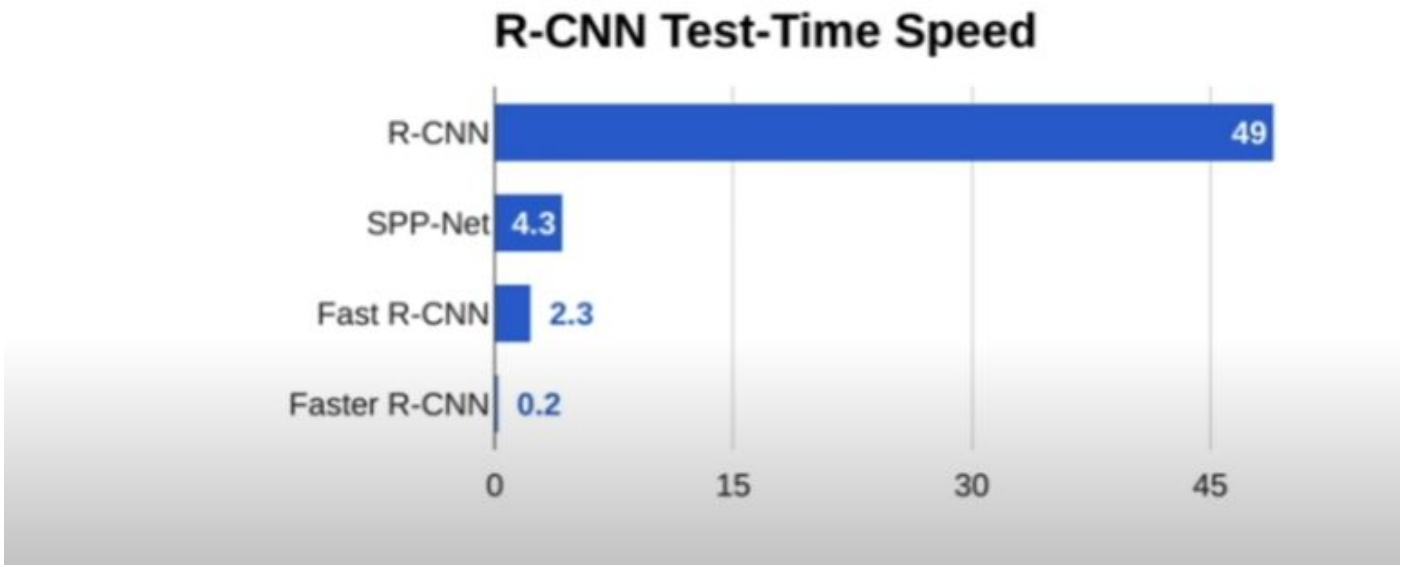
- Region Proposal Network (RPN)
- Multi-objective optimization
- Faster and support real time computation



Speed Performance

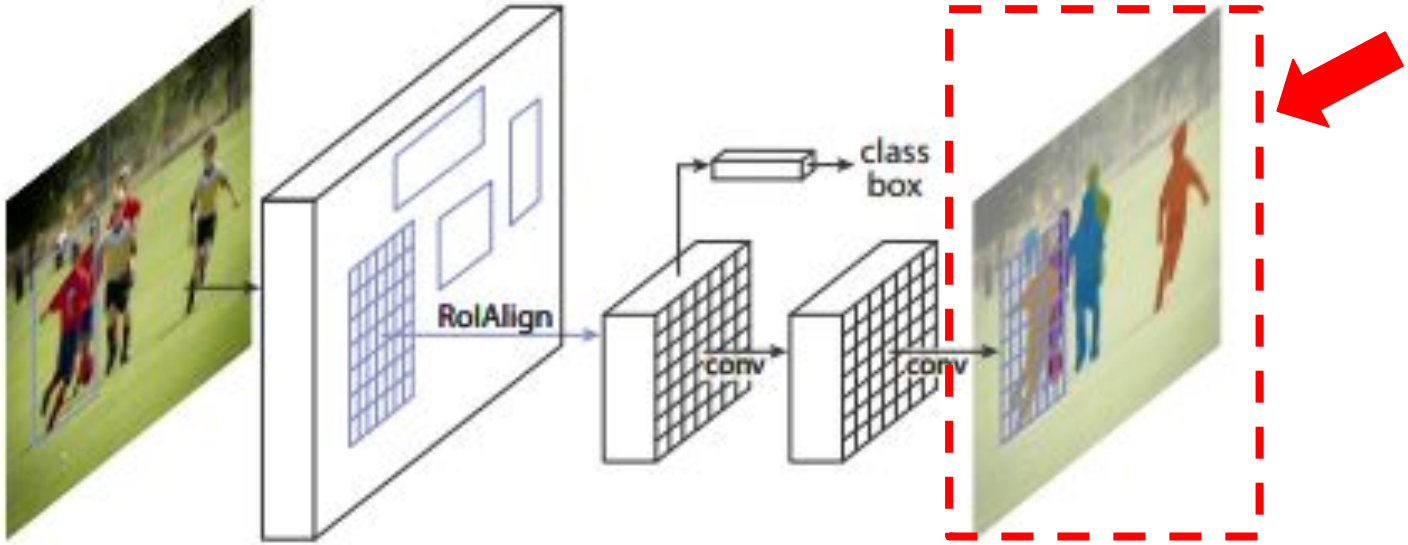
Faster R-CNN:

Make CNN do proposals!



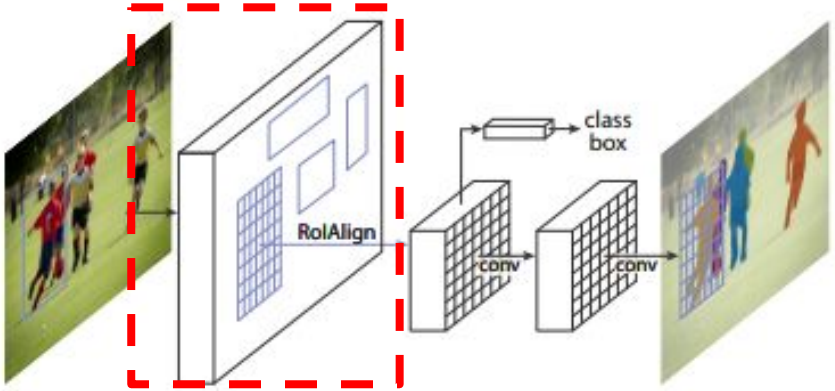
Back To Mask R-CNN!

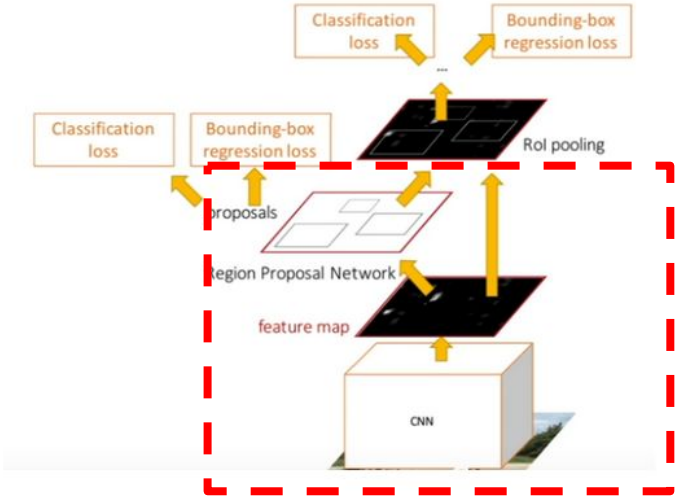
Mask R-CNN outputs a binary mask for each RoI on top of the Faster R-CNN



A Little More Detail: Architecture

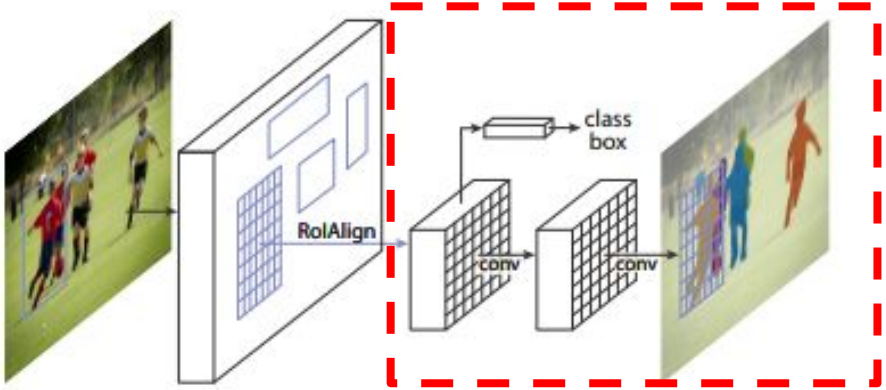
- Mask R-CNN adopts the same two-stage procedure with the identical RPN first stage

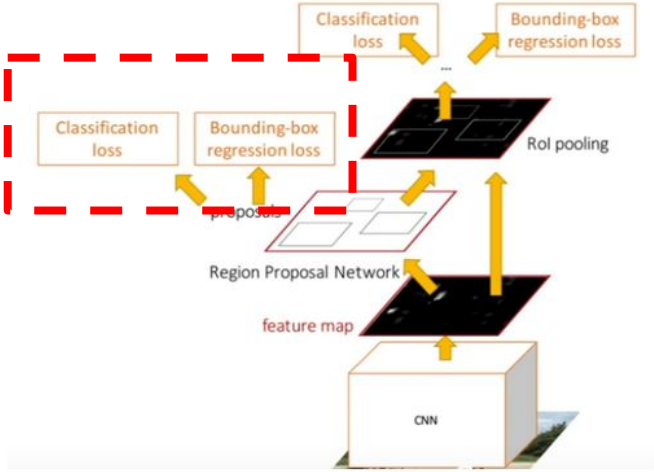




A Little More Detail: Architecture

- Mask R-CNN adopts the same two-stage procedure with the identical RPN first stage
- Mask R-CNN outputs a binary mask for each RoI in parallel to predicting the class and box offset in the second stage





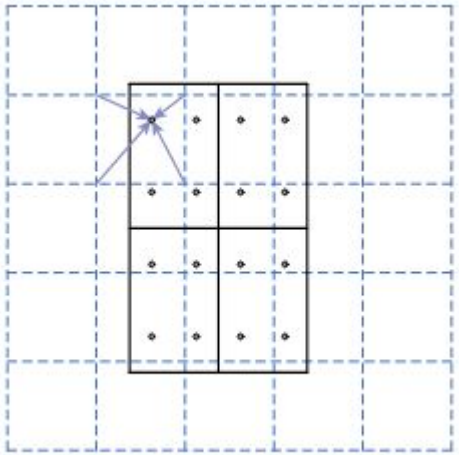
A Little More Detail: Loss Function

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

Note: the mask branch predicts encodings for K classes of each RoI so that the network can generate masks without competition, which decouples mask and class prediction

A Little More Detail: RoIAlign

- To extract the pixel-pixel mask, the RoI to be well aligned to preserve the explicit per-pixel spatial correspondence
- RoIPool: quantize a floating-number RoI to the discrete granularity of the feature map
- RoIAlign: bilinear interpolation to compute the exact values of the input features



Experimental Setup: Instance Segmentation

- Evaluate on COCO dataset
  - Metrics composed of variants of AP (average precision)
- Train using the union of 80k train images and a 35k subset of val images

Experimental Setup: Instance Segmentation

All instantiations of Mask R-CNN outperforms baseline variants of previous state-of-the-art instance segmentation models

	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**

Experimental Setup: Ablation Studies

- **multinomial vs independent**: decoupled mask prediction performs better
- **class specific vs class agnostic mask**: not much difference
  - once the instance has been classified by the box branch, it is sufficient to predict a mask
- **RoIAlign vs RoIPool**: RoIAlign performs better
- **object detection with only RoIAlign**: performs better
- **timing**: a little slower than Faster R-CNN but still comparably fast

Limitation

- Still only work on images, so it can’t explore temporal information of the object of interest in a dynamic setting [6]
- Mask R-CNN usually suffers from motion blur at low resolution and encounters failures
- Supervised training, so getting data labels is very difficult

Future Work

- Hand segmentation under different viewpoints by combination of Mask R-CNN with tracking [6]
  - apply the Mask R-CNN to perform hand segmentation
- Embodied Amodal Recognition: Learning to Move to Perceive Objects [7]
  - apply the Mask R-CNN for agents to learn to move strategically to improve their visual recognition abilities

Summary

- This paper introduces a new framework built on top of Faster R-CNN to perform instance segmentation
- Instance segmentation is important to localize object that can be used to perform planning
- Prior R-CNN family of work only focuses on the object detection but not the instance segmentation problem
- This work is an simple extension of Faster R-CNN without much overhead
- This work outperforms all the state-of-the-art instance segmentation approaches by a large margin

Citations

[1] He et al, “Mask R-CNN”, ICCV 2017
[2] Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014
[3] Ross Girshick, “Fast R-CNN”, ICCV 2015
[4] Ren et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NIPS 2015
[5] Johnson et al, “Lecture 11 | Detection and Segmentation”, *Youtube*, Aug 11, 2017. Available: https://www.youtube.com/watch?v=nDPWywWRIRo
[6] Nguyen et al, “Hand segmentation under different viewpoints by combination of Mask R-CNN with tracking”, ACDT 2018
[7] yang et al, “Embodied Amodal Recognition: Learning to Move to Perceive Objects”, ICCV 2019

Motivation and Main Problem

1-5 slides

High-level description of problem being solved

Why is the problem important?

- its significance towards general-purpose robot autonomy
- its potential application and societal impact of the problem

Technical challenges arising from the problem

- the role of the AI and machine learning in tackling this problem

High-level idea of why prior approaches didn't already solve

Key insight(s) (try to do in 1-3) of the proposed work

Problem Setting

1 or more slides

Problem formulation, key definitions and notations

- Be precise -- should be as formal as in the paper

Context / Related Work / Limitations of Prior Work

1 or more slides

Which other papers have tried to tackle this problem or a related problem?

- The paper's related work is a good start, but there may be others
- What is the key limitations of prior work(s)?

Proposed Approach / Algorithm / Method

1-5 slides

Describe algorithm or framework (pseudocode and flowcharts can help)

- What is the optimization objective?
- What are the core technical innovations of the algorithm/framework?

Implementation details should be left out here, but may be discussed later if its relevant for limitations / experiments

Theory (if relevant)

What are the assumptions made for the theory? Are these reasonable? Realistic?

If the theory build strongly on other prior theory / results, reference those and state them here.

Theory (if relevant, continued)

State main results formally

Give proof sketches

Refer students to the full proofs in paper

Experimental Setup

1-3 slides

Description of the experimental evaluation setting

- What is the domain(s), e.g., datasets, tasks, robot hardware setups?
- What are the baseline(s)?
- What scientific hypotheses are tested?

How did the authors evaluate the success of their approach?

- Clear description of the metrics that will be used

Experimental Results

>1 slide

Present the quantitative and qualitative results

Show figures / tables / plots / robot demos

Pinpoint the most interesting / significant results

Discussion of Results

1-2 slides

What conclusions are drawn from the results by the authors?

- What insights are gained from the experiments?
- What strengths and weaknesses of the proposed method are illustrated by the results?

Are the stated conclusions fully backed by the results and references?

- If so, why? (Recap the relevant supporting evidences from the given results + refs)
- If not, what are the additional experiments / comparisons that can further support/repudiate the conclusions of the paper?

Critique / Limitations / Open Issues

1-2 slides

What are the key limitations of the proposed approach / ideas? (e.g. does it require strong assumptions that are unlikely to be practical? Computationally expensive? Require a lot of data?)

Are there any practical challenges in deploying the approach on physical robots in the real world? Are there any safety or ethical concerns of using such approach?

If follow-up work has addressed some of these limitations, include pointers to that. But don't limit your discussion only to the problems / limitations that have already been addressed.

Future Work for Paper / Reading

1-2 slides

What interesting questions does it raise for future work?

- Your own ideas for future work
- Others' ideas (if others have already built on this idea)

Extended Readings

1-2 slides

Pointers to papers that use this paper as a reference and/or other very related papers that others may want to read

Point classmates to where they can go for further reading on this paper/reading

Summary

1 slide

Approximately one bullet for each of the following

- Problem the reading is discussing
- Why is it important and hard
- What is the key limitation of prior work
- What is the key insight(s) (try to do in 1-3) of the proposed work
- What did they demonstrate by this insight? (tighter theoretical bounds, state of the art performance on X, etc)