

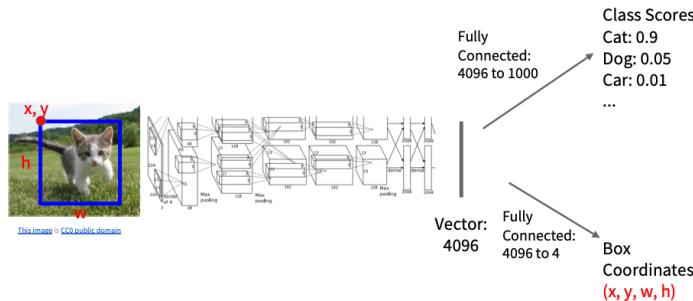
Lecture 9: Object Detection and Image Segmentation

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 1

April 30, 2024

Object Detection: Single Object (Classification + Localization)

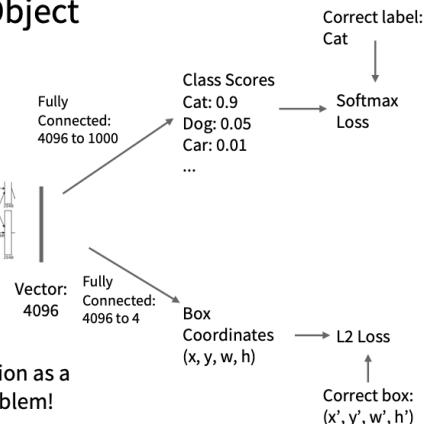
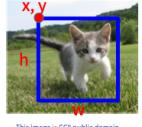


Fei-Fei Li, Ehsan Adeli

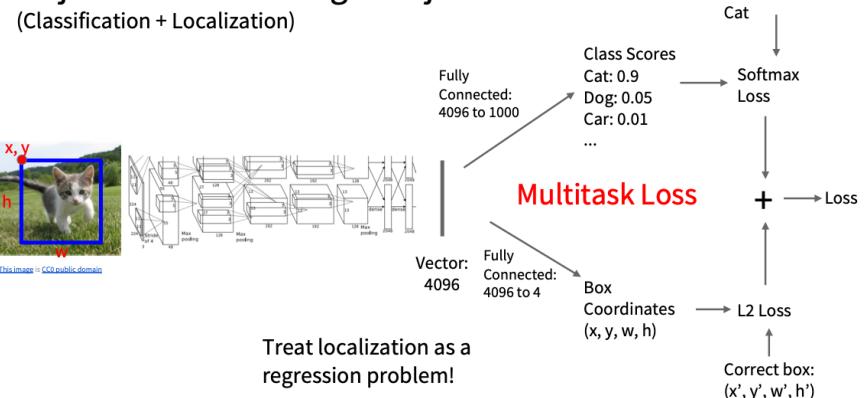
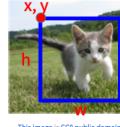
Lecture 11 - 41

April 30, 2024

Object Detection: Single Object (Classification + Localization)



Object Detection: Single Object (Classification + Localization)

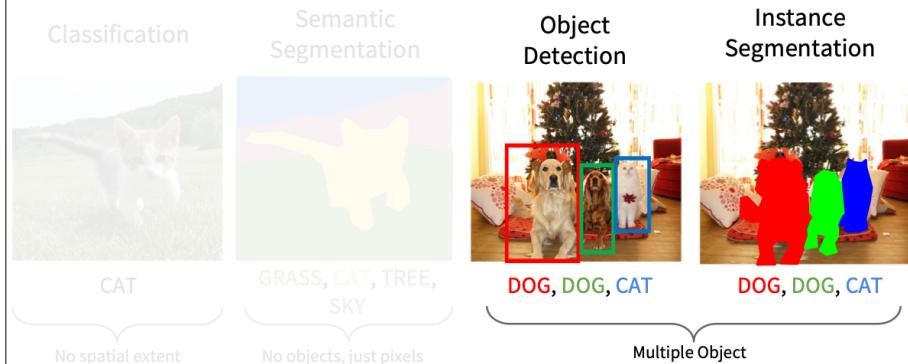


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 43

April 30, 2024

Object Detection

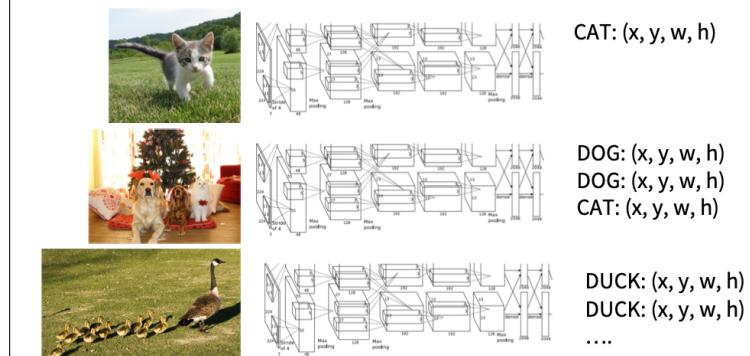


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 40

April 30, 2024

Object Detection: Multiple Objects

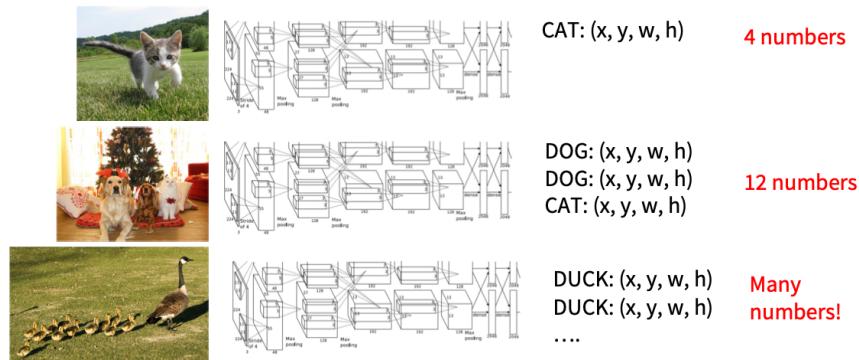


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 44

April 30, 2024

Object Detection: Multiple Objects

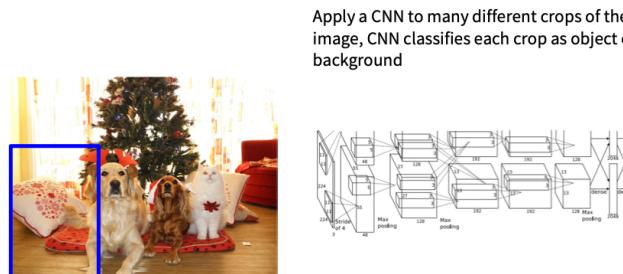


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 45

April 30, 2024

Object Detection: Multiple Objects



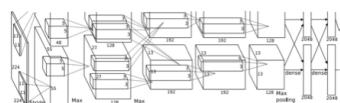
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 46

April 30, 2024

Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

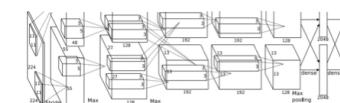
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 47

April 30, 2024

Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

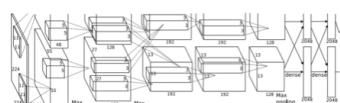
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 48

April 30, 2024

Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? YES
Background? NO

Q: What's the problem with this approach?

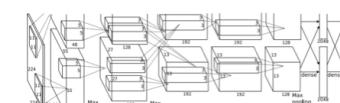
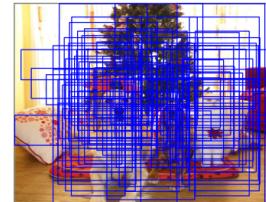
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 49

April 30, 2024

Object Detection: Multiple Objects

Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? YES
Background? NO

Problem: Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

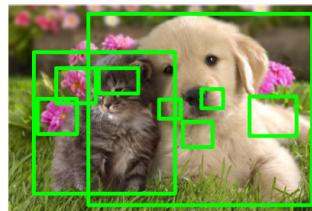
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 50

April 30, 2024

Region Proposals: Selective Search

- Find “blobby” image regions that are likely to contain objects
- Relatively fast to run; e.g. Selective Search gives 2000 region proposals in a few seconds on CPU



Alexe et al., “Measuring the ‘blobbiness’ of image windows”, TPAMI 2012
Uijlings et al., “Selective Search for Object Recognition”, IJCV 2013
Cheng et al., “BING: Binarized normed gradients for objectness estimation at 300fps”, CVPR 2014
Zitnick and Dollar, “Edge boxes: Locating object proposals from edges”, ECCV 2014

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 51

April 30, 2024

R-CNN



Input image

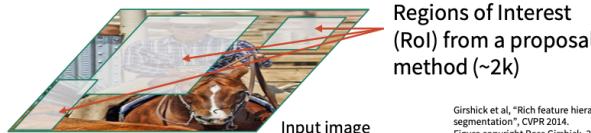
Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 52

April 30, 2024

R-CNN



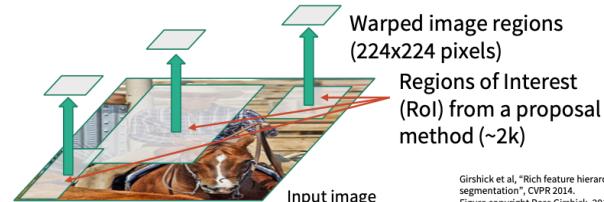
Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 53

April 30, 2024

R-CNN



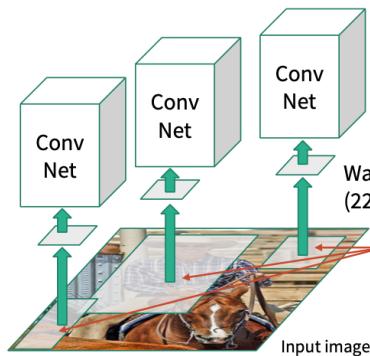
Girshick et al., “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 54

April 30, 2024

R-CNN



Forward each region through ConvNet
(ImageNet-pretrained)

Warped image regions
(224x224 pixels)

Regions of Interest
(RoI) from a proposal
method (~2k)

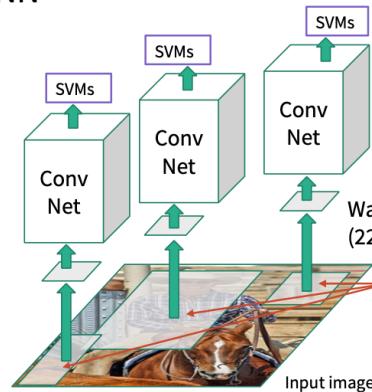
Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 55

April 30, 2024

R-CNN



Classify regions with
SVMs

Forward each region through ConvNet
(ImageNet-pretrained)

Warped image regions
(224x224 pixels)

Regions of Interest
(RoI) from a proposal
method (~2k)

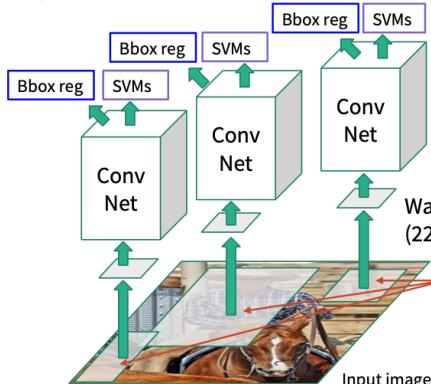
Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 56

April 30, 2024

R-CNN



Predict "corrections" to the RoI: 4 numbers: (dx, dy, dw, dh)

Bbox reg
SVMs

Conv Net

Forward each region through ConvNet
(ImageNet-pretrained)

Warped image regions
(224x224 pixels)

Regions of Interest
(RoI) from a proposal
method (~2k)

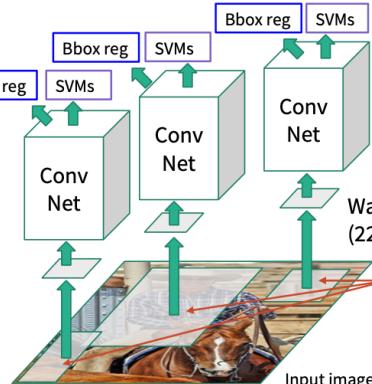
Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 57

April 30, 2024

R-CNN



Predict "corrections" to the RoI: 4 numbers: (dx, dy, dw, dh)

Bbox reg
SVMs

Conv Net

Forward each region through ConvNet

Warped image regions
(224x224 pixels)

Regions of Interest
(RoI) from a proposal
method (~2k)

Problem: Very slow!
Need to do ~2k
independent forward
passes for each image!

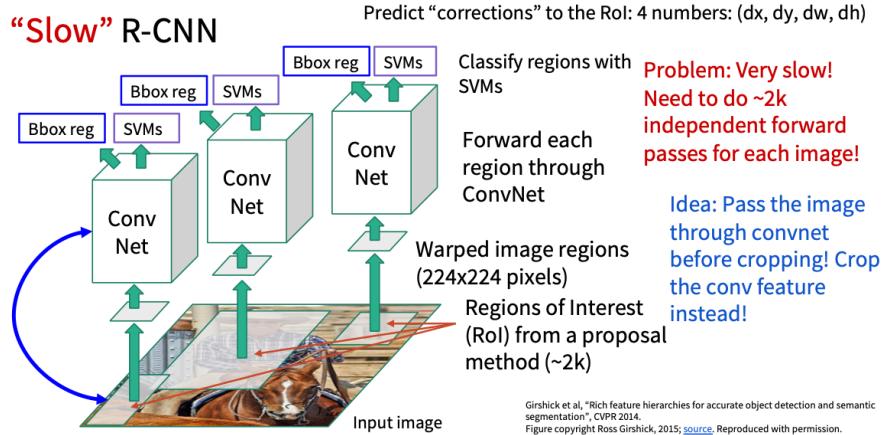
Girshick et al., "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 58

April 30, 2024

“Slow” R-CNN



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 59

April 30, 2024

Fast R-CNN



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 60

April 30, 2024

Fast R-CNN

Fast R-CNN

Regions of Interest (RoI)
from a proposal
method

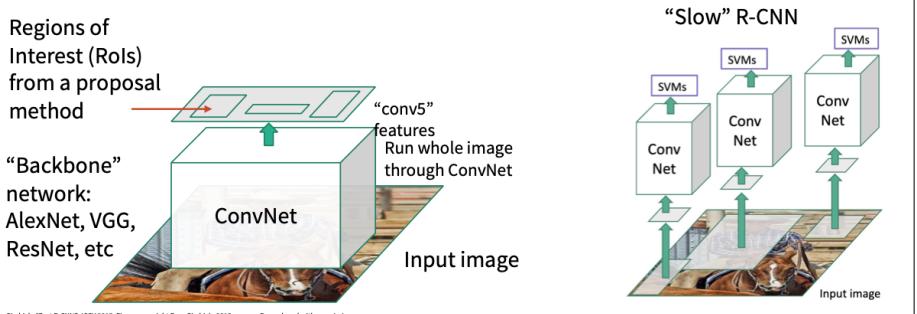
“Backbone”
network:
AlexNet, VGG,
ResNet, etc

Girshick, “Fast R-CNN”, ICCV 2015. Figure copyright Ross Girshick, 2015; [source](#). Reproduced with permission.

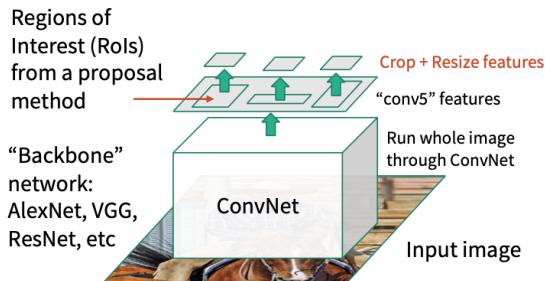
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 62

April 30, 2024



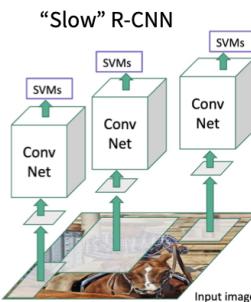
Fast R-CNN



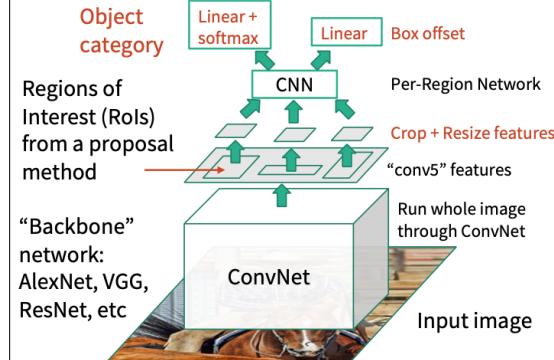
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 63

April 30, 2024



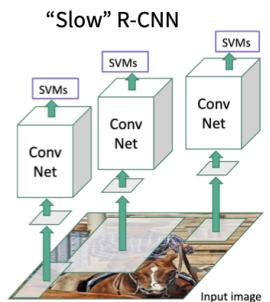
Fast R-CNN



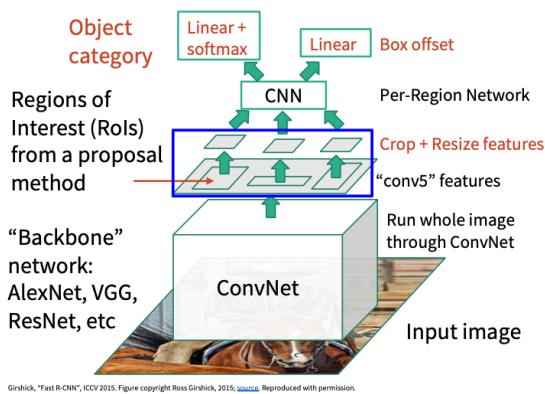
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 64

April 30, 2024



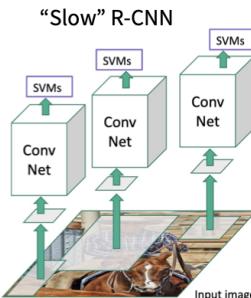
Fast R-CNN



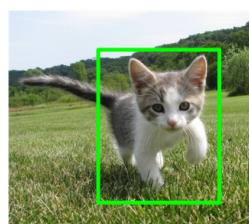
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 65

April 30, 2024



Cropping Features: RoI Pool



Input Image
(e.g. 3 x 640 x 480)

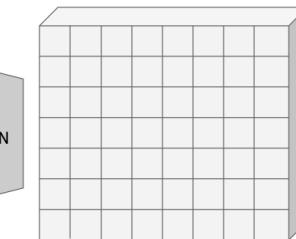


Image features: C x H x W
(e.g. 512 x 20 x 15)

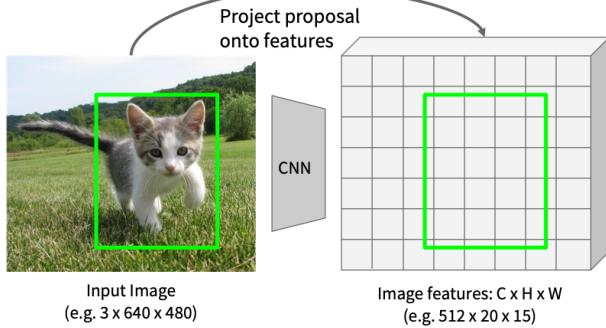
Girshick, “Fast R-CNN”, ICCV 2015.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 66

April 30, 2024

Cropping Features: RoI Pool



Input Image
(e.g. $3 \times 640 \times 480$)

Project proposal
onto features

CNN

Image features: $C \times H \times W$
(e.g. $512 \times 20 \times 15$)

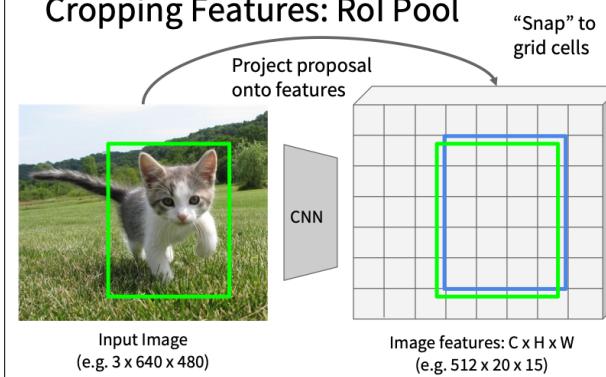
Girshick, "Fast R-CNN", ICCV 2015.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 67

April 30, 2024

Cropping Features: RoI Pool



Input Image
(e.g. $3 \times 640 \times 480$)

CNN

Image features: $C \times H \times W$
(e.g. $512 \times 20 \times 15$)

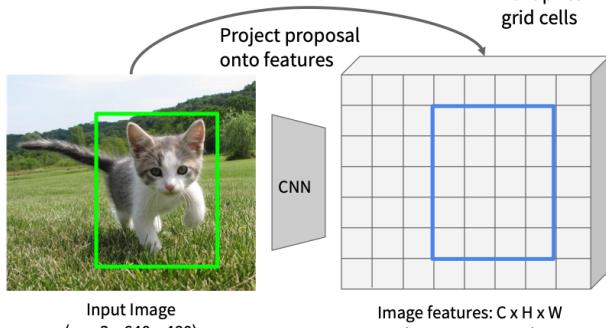
Girshick, "Fast R-CNN", ICCV 2015.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 68

April 30, 2024

Cropping Features: RoI Pool



Input Image
(e.g. $3 \times 640 \times 480$)

Project proposal
onto features

CNN

Image features: $C \times H \times W$
(e.g. $512 \times 20 \times 15$)

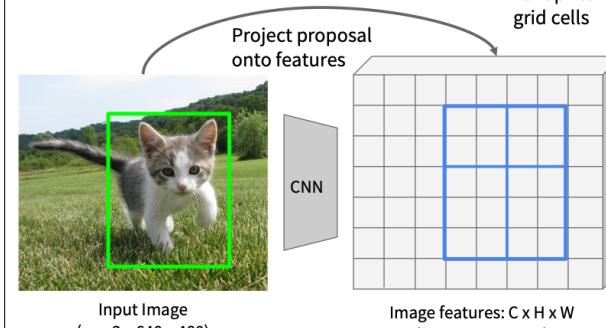
Girshick, "Fast R-CNN", ICCV 2015.

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 69

April 30, 2024

Cropping Features: RoI Pool



Input Image
(e.g. $3 \times 640 \times 480$)

CNN

Image features: $C \times H \times W$
(e.g. $512 \times 20 \times 15$)

Girshick, "Fast R-CNN", ICCV 2015.

Fei-Fei Li, Ehsan Adeli

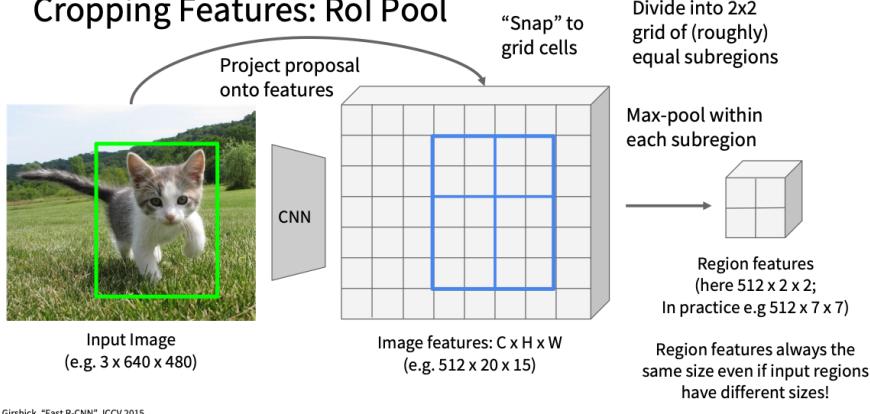
Lecture 11 - 70

April 30, 2024

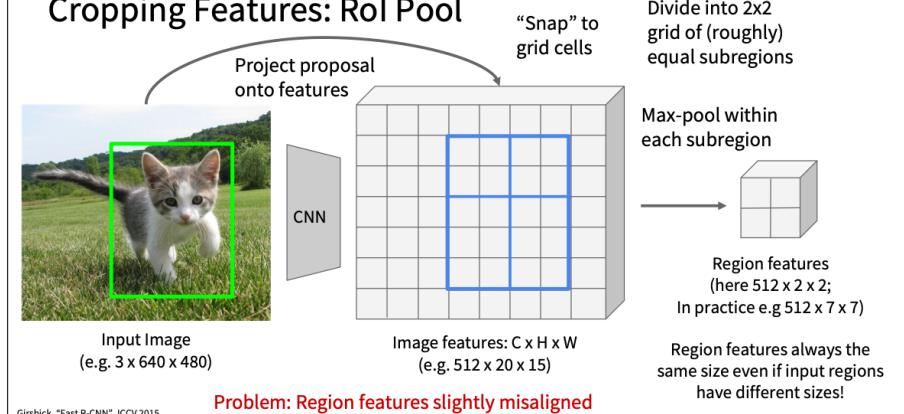
"Snap" to grid cells
Divide into 2×2
grid of (roughly)
equal subregions

Q: how do we resize the 512×5
 $\times 4$ region to, e.g., a $512 \times 2 \times 2$
tensor?

Cropping Features: RoI Pool



Cropping Features: RoI Pool



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 71

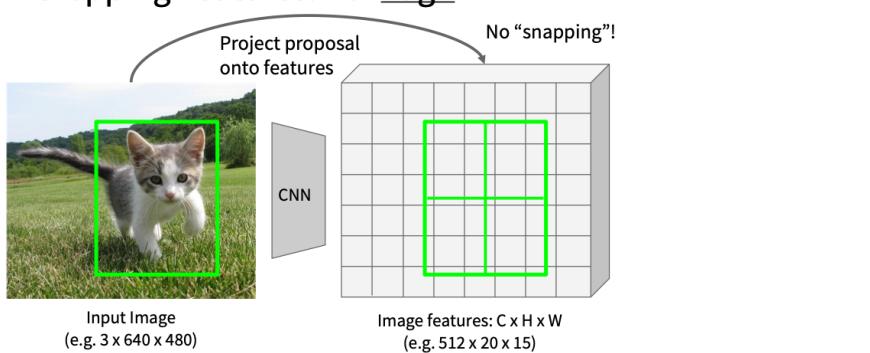
April 30, 2024

Fei-Fei Li, Ehsan Adeli

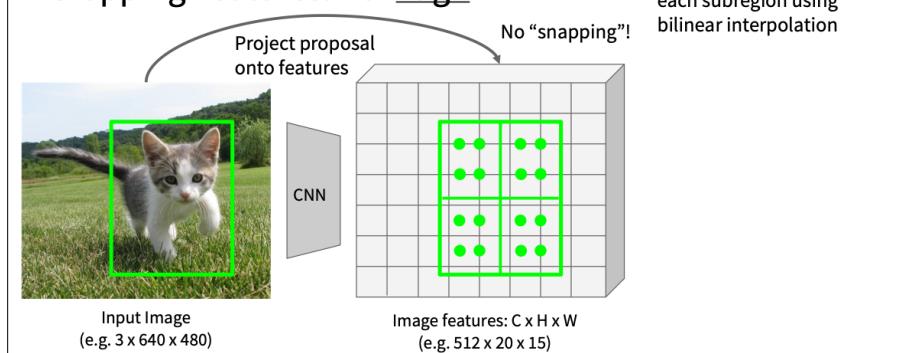
Lecture 11 - 72

April 30, 2024

Cropping Features: RoI Align



Cropping Features: RoI Align



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 73

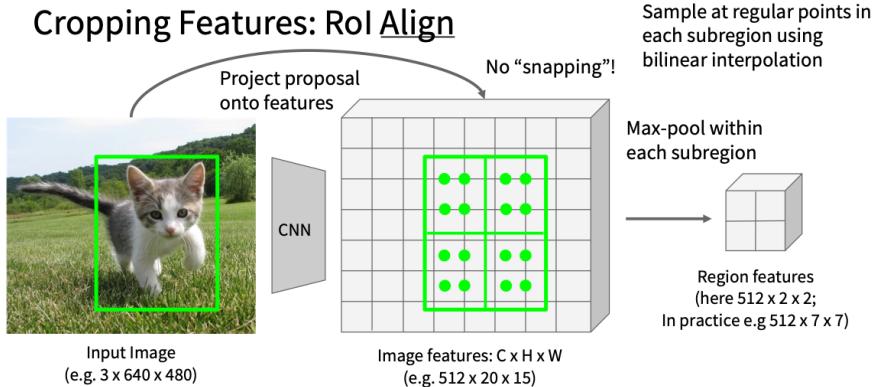
April 30, 2024

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 74

April 30, 2024

Cropping Features: RoI Align



He et al, "Mask R-CNN", ICCV 2017

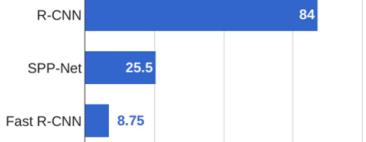
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 77

April 30, 2024

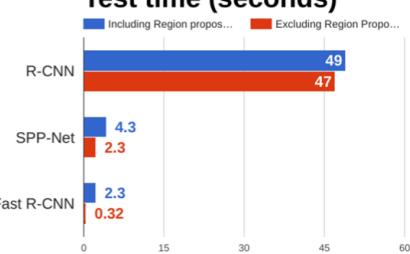
R-CNN vs Fast R-CNN

Training time (Hours)



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
He et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick, "Fast R-CNN", ICCV 2015

Test time (seconds)



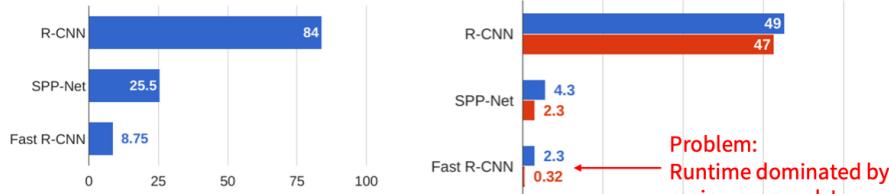
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 78

April 30, 2024

R-CNN vs Fast R-CNN

Training time (Hours)



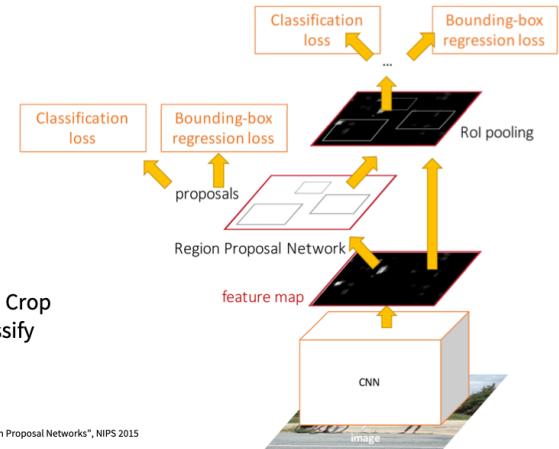
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.
He et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV 2014
Girshick, "Fast R-CNN", ICCV 2015

Faster R-CNN:

Make CNN do proposals!

Insert Region Proposal Network (RPN) to predict proposals from features

Otherwise same as Fast R-CNN: Crop features for each proposal, classify each one



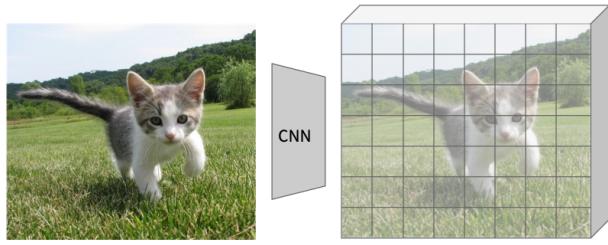
Ren et al, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 80

April 30, 2024

Region Proposal Network



Input Image
(e.g. 3 x 640 x 480)

CNN

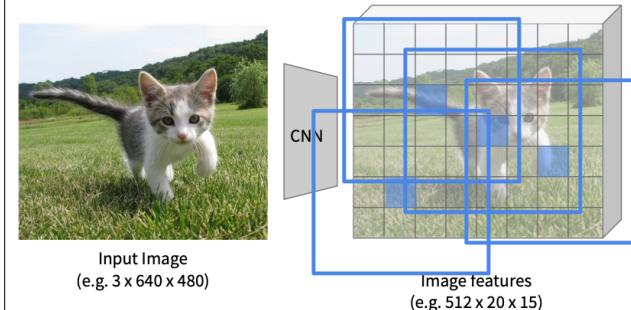
Image features
(e.g. 512 x 20 x 15)

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 81

April 30, 2024

Region Proposal Network



Input Image
(e.g. 3 x 640 x 480)

CNN

Image features
(e.g. 512 x 20 x 15)

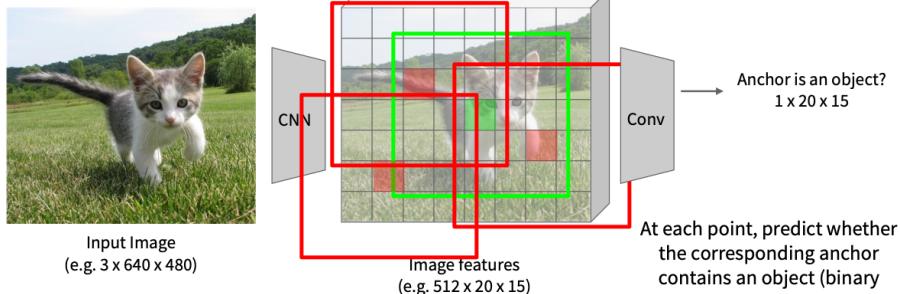
Imagine an anchor box of fixed size at each point in the feature map

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 82

April 30, 2024

Region Proposal Network



Input Image
(e.g. 3 x 640 x 480)

CNN

Image features
(e.g. 512 x 20 x 15)

Imagine an anchor box of fixed size at each point in the feature map

Anchor is an object?
1 x 20 x 15

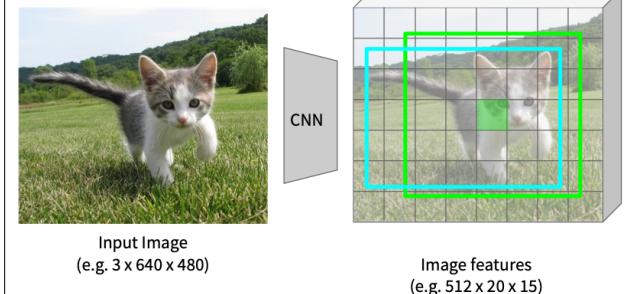
At each point, predict whether the corresponding anchor contains an object (binary classification)

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 83

April 30, 2024

Region Proposal Network



Input Image
(e.g. 3 x 640 x 480)

CNN

Image features
(e.g. 512 x 20 x 15)

Imagine an anchor box of fixed size at each point in the feature map

Anchor is an object?
1 x 20 x 15

Box corrections
4 x 20 x 15

For positive boxes, also predict a corrections from the anchor to the ground-truth box (regress 4 numbers per pixel)

Fei-Fei Li, Ehsan Adeli

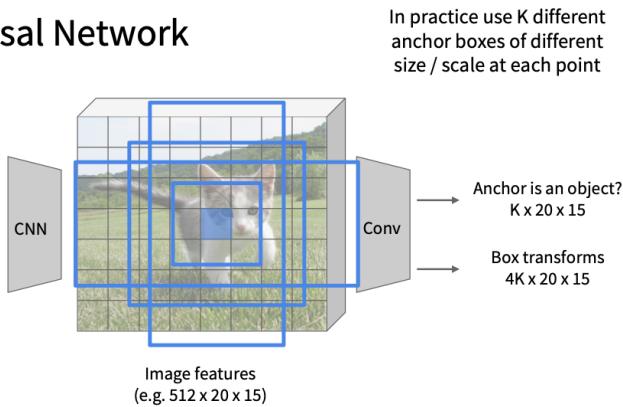
Lecture 11 - 84

April 30, 2024

Region Proposal Network



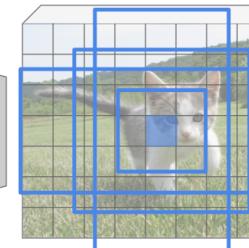
Input Image
(e.g. 3 x 640 x 480)



Region Proposal Network



Input Image
(e.g. 3 x 640 x 480)



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 85

April 30, 2024

Fei-Fei Li, Ehsan Adeli

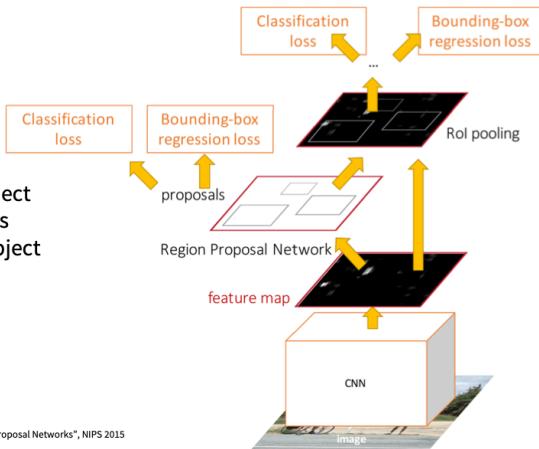
Lecture 11 - 86

April 30, 2024

Faster R-CNN:

Make CNN do proposals!

- Jointly train with 4 losses:
1. RPN classify object / not object
 2. RPN regress box coordinates
 3. Final classification score (object classes)
 4. Final box coordinates



Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 87

April 30, 2024

Faster R-CNN:

Make CNN do proposals!

R-CNN Test-Time Speed



Fei-Fei Li, Ehsan Adeli

Lecture 11 - 88

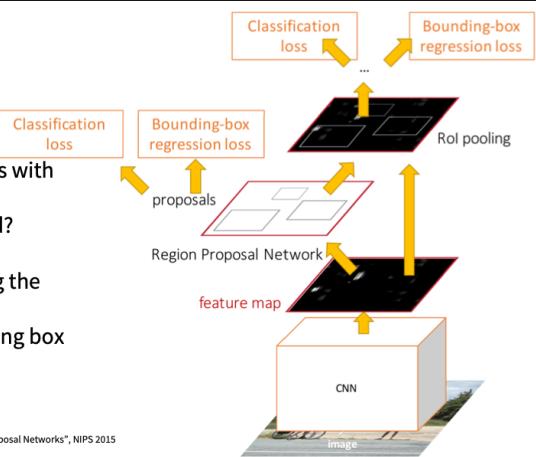
April 30, 2024

Faster R-CNN:

Make CNN do proposals!

Glossing over many details:

- Ignore overlapping proposals with non-max suppression
- How are anchors determined?
- How do we sample positive / negative samples for training the RPN?
- How to parameterize bounding box regression?



Ren et al., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", NIPS 2015
Figure copyright 2015, Ross Girshick; reproduced with permission

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 89

April 30, 2024

Faster R-CNN:

Make CNN do proposals!

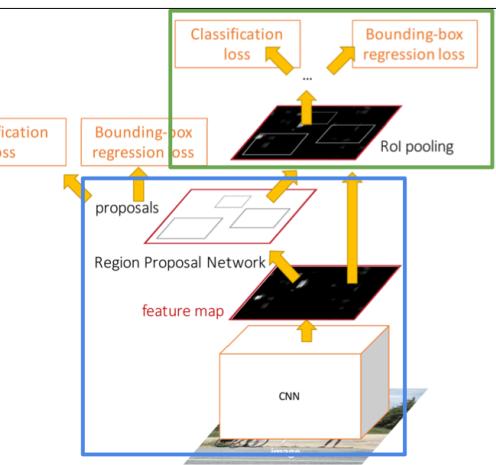
Faster R-CNN is a Two-stage object detector

First stage: Run once per image

- Backbone network
- Region proposal network

Second stage: Run once per region

- Crop features: RoI pool / align
- Predict object class
- Prediction bbox offset



Fei-Fei Li, Ehsan Adeli

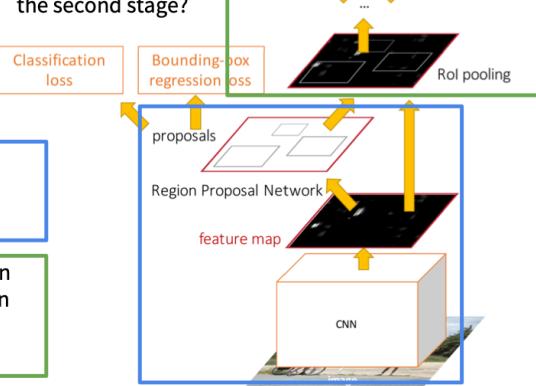
Lecture 11 - 90

April 30, 2024

Faster R-CNN:

Make CNN do proposals!

Do we really need the second stage?



Faster R-CNN is a Two-stage object detector

First stage: Run once per image

- Backbone network
- Region proposal network

Second stage: Run once per region

- Crop features: RoI pool / align
- Predict object class
- Prediction bbox offset

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 91

April 30, 2024

Single-Stage Object Detectors: YOLO / SSD / RetinaNet



Input image
3 x H x W



Divide image into grid
7 x 7

Image a set of base boxes
centered at each grid cell
Here B = 3

Within each grid cell:

- Regress from each of the B base boxes to a final box with 5 numbers: $(dx, dy, dh, dw, \text{confidence})$
- Predict scores for each of C classes (including background as a class)
- Looks a lot like RPN, but category-specific!

Output:
 $7 \times 7 \times (5 * B + C)$

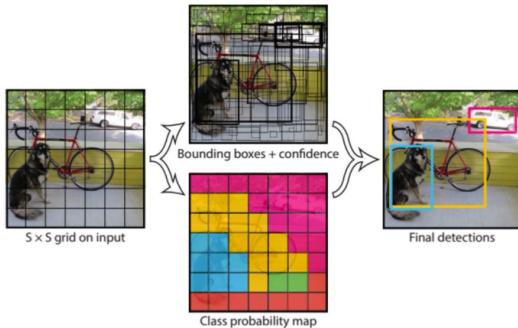
Fei-Fei Li, Ehsan Adeli

Lecture 11 - 92

April 30, 2024

Redmon et al., "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al., "SSD: Single-Shot MultiBox Detector", ECCV 2016
Liu et al., "Focal Loss for Dense Object Detection", ICCV 2017

YOLO- real-time object detection



Redmon et al. "You only look once: unified, real-time object detection (2015)."

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 93

April 30, 2024

Object Detection: Lots of variables ...

Backbone Network	"Meta-Architecture"	Takeaways
VGG16	Two-stage: Faster R-CNN	Faster R-CNN is slower but more accurate
ResNet-101	Single-stage: YOLO / SSD	
Inception V2	Hybrid: R-FCN	
Inception V3	Image Size	SSD is much faster but not as accurate
Inception ResNet	# Region Proposals	
MobileNet	...	Bigger / Deeper backbones work better

Huang et al, "Speed/accuracy trade-offs for modern convolutional object detectors", CVPR 2017

R-FCN: Dai et al., "R-FCN: Object Detection via Region-based Fully Convolutional Networks", NIPS 2016
Inception V2: Ioffe and Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ICML 2015
Inception V3: Szegedy et al., "Rethinking the Inception Architecture for Computer Vision", arXiv 2016
Inception ResNet: Szegedy et al., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", arXiv 2016
MobileNet: Howard et al., "Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv 2017

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 94

April 30, 2024

Object Detection: Lots of variables ...

Backbone Network	"Meta-Architecture"	Takeaways
VGG16	Two-stage: Faster R-CNN	Faster R-CNN is slower but more accurate
ResNet-101	Single-stage: YOLO / SSD	
Inception V2	Hybrid: R-FCN	
Inception V3	Image Size	SSD is much faster but not as accurate
Inception ResNet	# Region Proposals	
MobileNet	...	Bigger / Deeper backbones work better

Huang et al, "Speed/accuracy trade-offs for modern convolutional object detectors", CVPR 2017
Zou et al, "Object Detection in 20 Years: A Survey", arXiv 2019

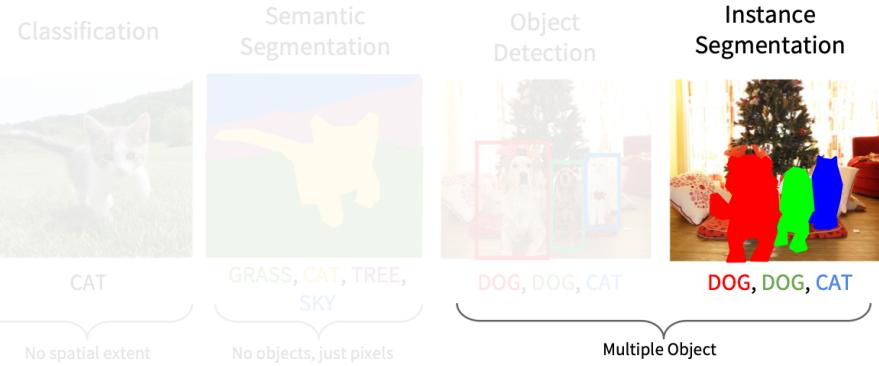
R-FCN: Dai et al., "R-FCN: Object Detection via Region-based Fully Convolutional Networks", NIPS 2016
Inception V2: Ioffe and Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift", ICML 2015
Inception V3: Szegedy et al., "Rethinking the Inception Architecture for Computer Vision", arXiv 2016
Inception ResNet: Szegedy et al., "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning", arXiv 2016
MobileNet: Howard et al., "Efficient Convolutional Neural Networks for Mobile Vision Applications", arXiv 2017

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 95

April 30, 2024

Instance Segmentation

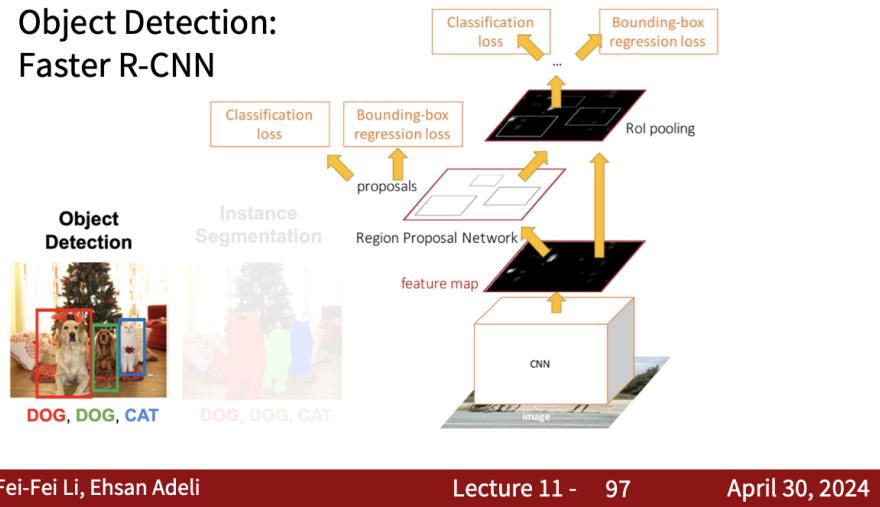


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 96

April 30, 2024

Object Detection: Faster R-CNN

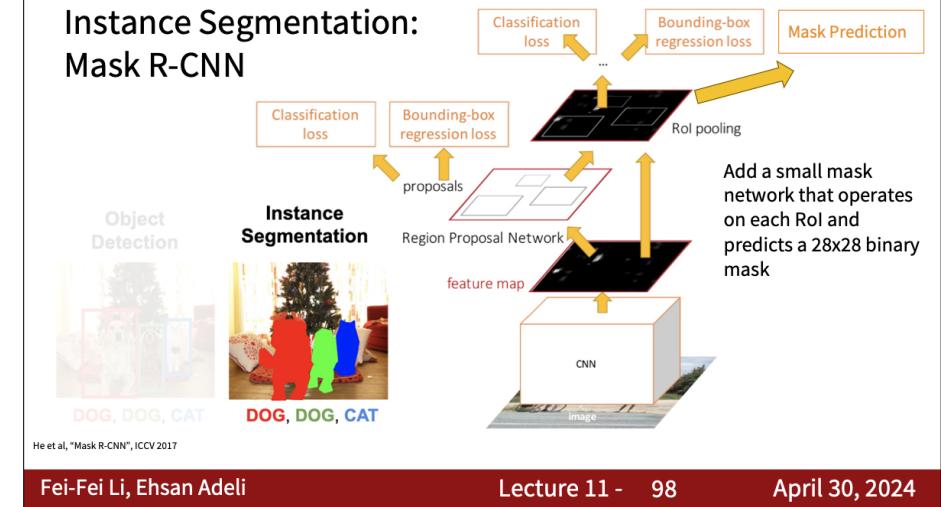


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 97

April 30, 2024

Instance Segmentation: Mask R-CNN

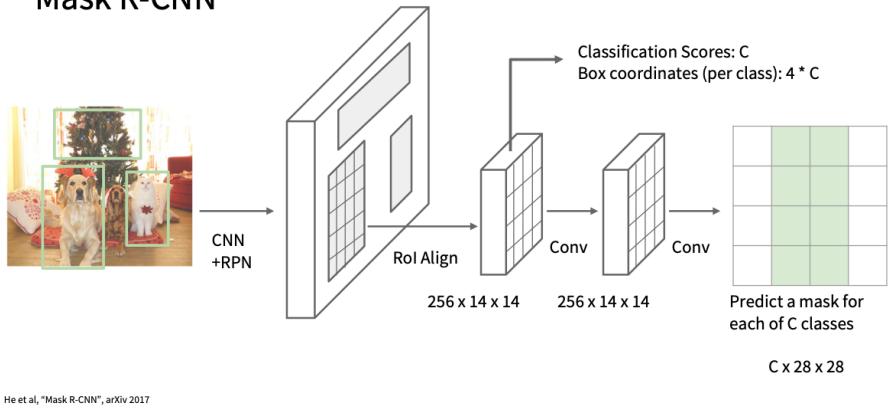


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 98

April 30, 2024

Mask R-CNN



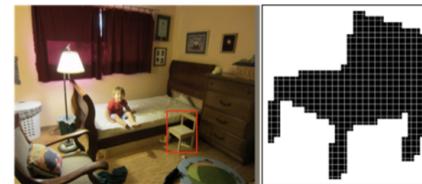
He et al., "Mask R-CNN", arXiv 2017

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 99

April 30, 2024

Mask R-CNN: Example Mask Training Targets

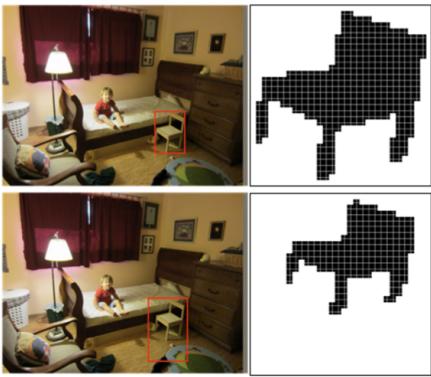


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 100

April 30, 2024

Mask R-CNN: Example Mask Training Targets

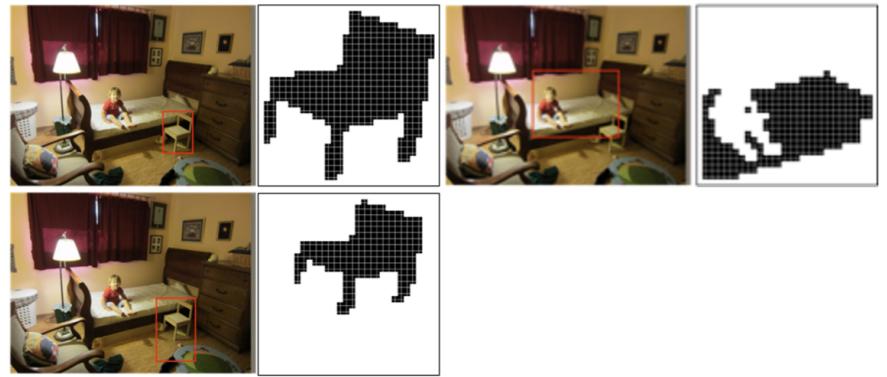


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 101

April 30, 2024

Mask R-CNN: Example Mask Training Targets

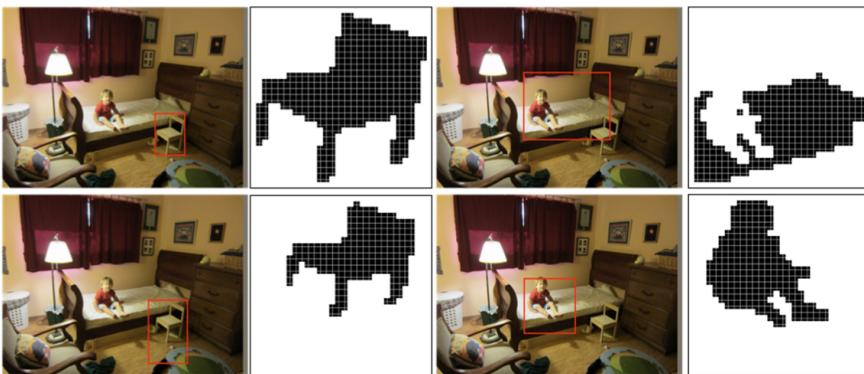


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 102

April 30, 2024

Mask R-CNN: Example Mask Training Targets

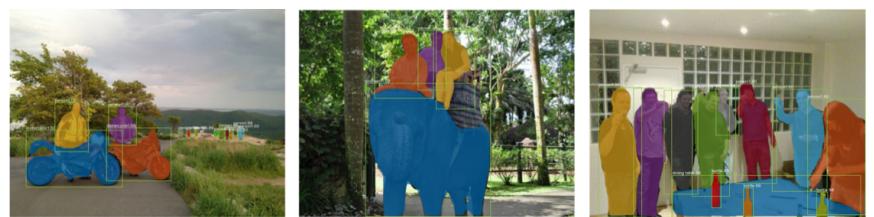


Fei-Fei Li, Ehsan Adeli

Lecture 11 - 103

April 30, 2024

Mask R-CNN: Very Good Results!



He et al., "Mask R-CNN", ICCV 2017

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 104

April 30, 2024

Mask R-CNN

Also does pose



He et al., "Mask R-CNN", ICCV 2017

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 105

April 30, 2024

Open Source Frameworks

Lots of good implementations on GitHub!

TensorFlow Detection API:

https://github.com/tensorflow/models/tree/master/research/object_detection
Faster RCNN, SSD, RFCN, Mask R-CNN, ...

Detectron2 (PyTorch)

<https://github.com/facebookresearch/detectron2>
Mask R-CNN, RetinaNet, Faster R-CNN, RPN, Fast R-CNN, R-FCN, ...

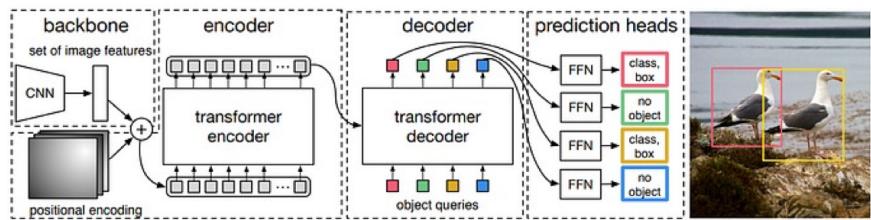
Finetune on your own dataset with pre-trained models

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 106

April 30, 2024

DETR (DEtection TRansformer)



Carion et al. "End-to-End Object Detection with Transformers" ECCV 2020

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 108

April 30, 2024

Recap: Lots of computer vision tasks!

Classification



CAT

Semantic Segmentation



GRASS, CAT, TREE, SKY

Object Detection



DOG, DOG, CAT

Instance Segmentation



DOG, DOG, CAT

This image is CC0 public domain

Fei-Fei Li, Ehsan Adeli

Lecture 11 - 109

April 30, 2024