CISC6000 Deep Learning
Object Detection & Segmentation

Dr. Yijun Zhao

Fordham University

So far: Image Classification

This image is CC0 public domain

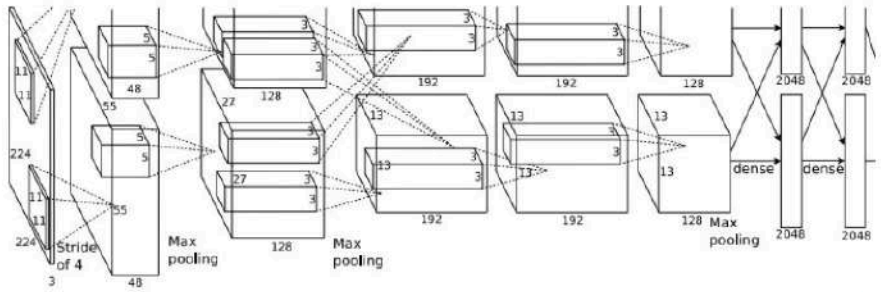


Figure copyright Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, 2012. Reproduced with permission.

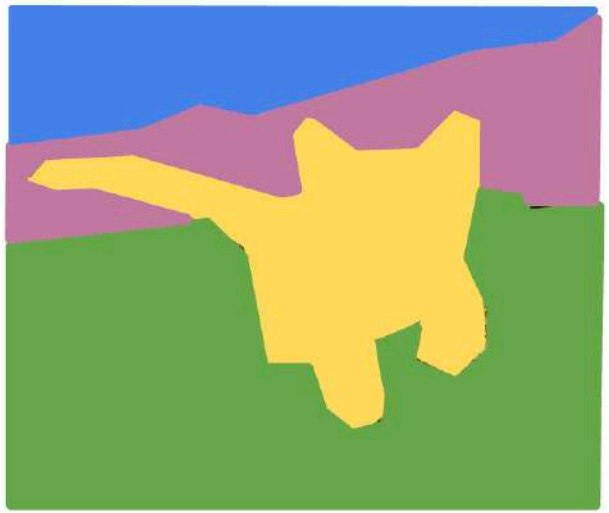
Vector:
4096

Fully-Connected:
4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Other Computer Vision Tasks

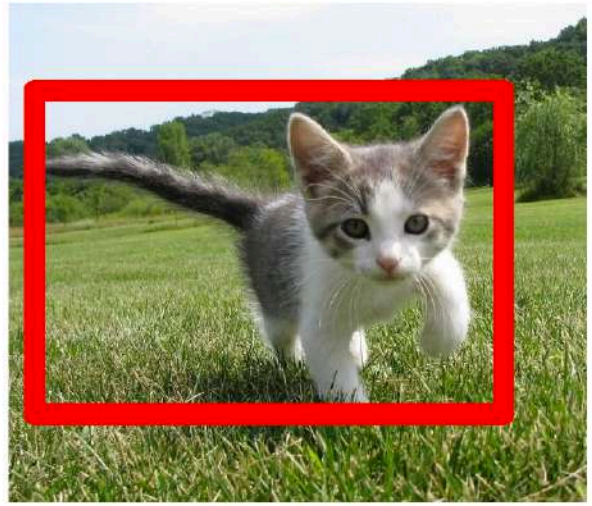
Semantic Segmentation



GRASS, CAT, TREE, SKY

No objects, just pixels

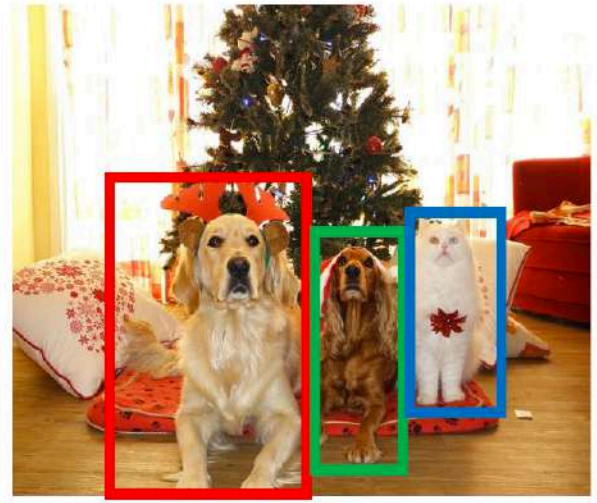
Classification + Localization



CAT

Single Object

Object Detection



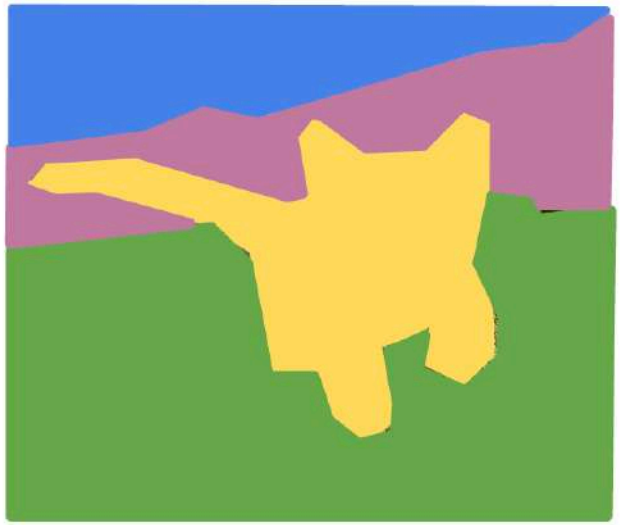
DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

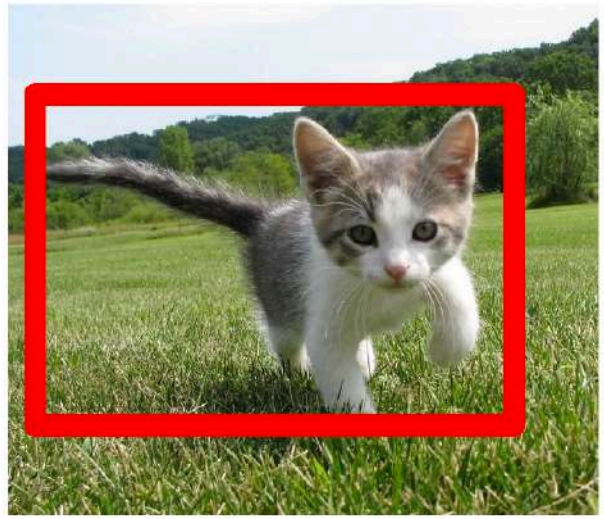
Multiple Object

Semantic Segmentation



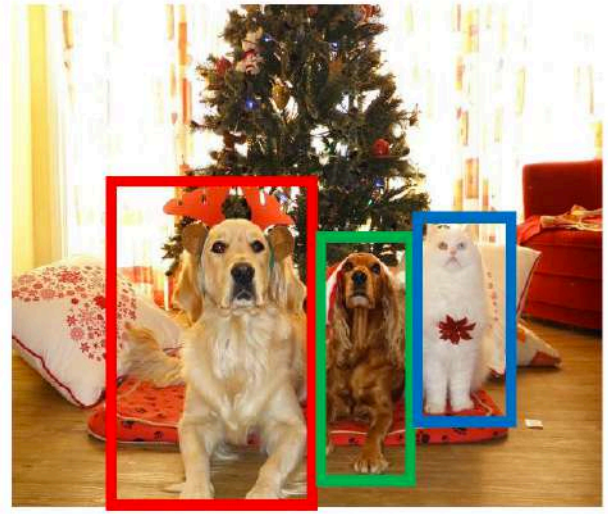
GRASS, CAT,
TREE, SKY

No objects, just pixels



CAT

Single Object



DOG, DOG, CAT

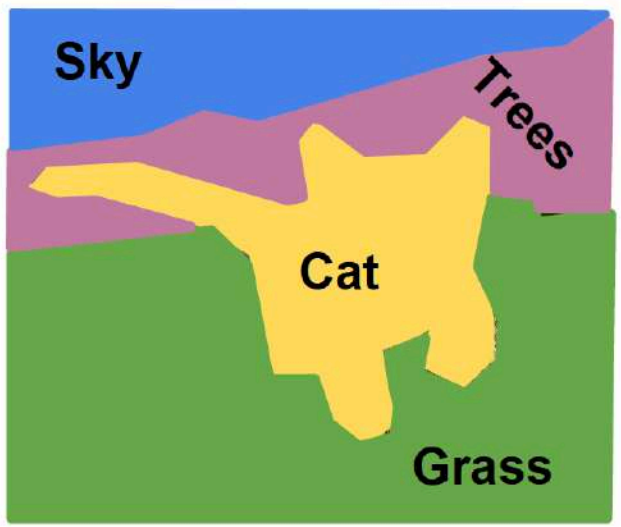
DOG, DOG, CAT

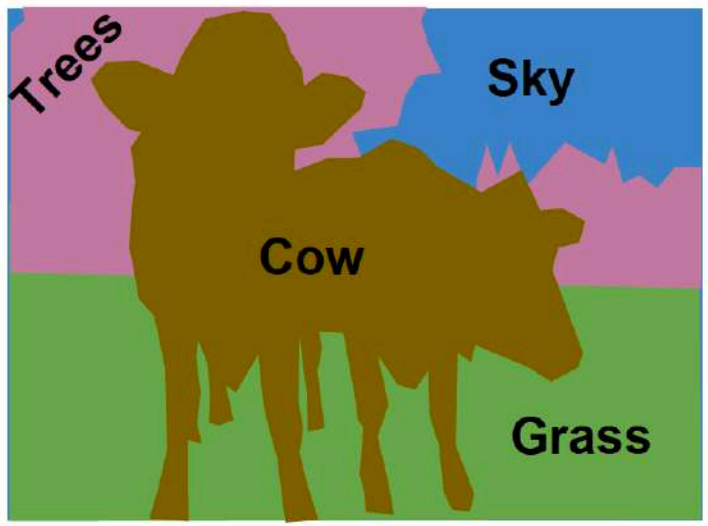
Multiple Object

Semantic Segmentation

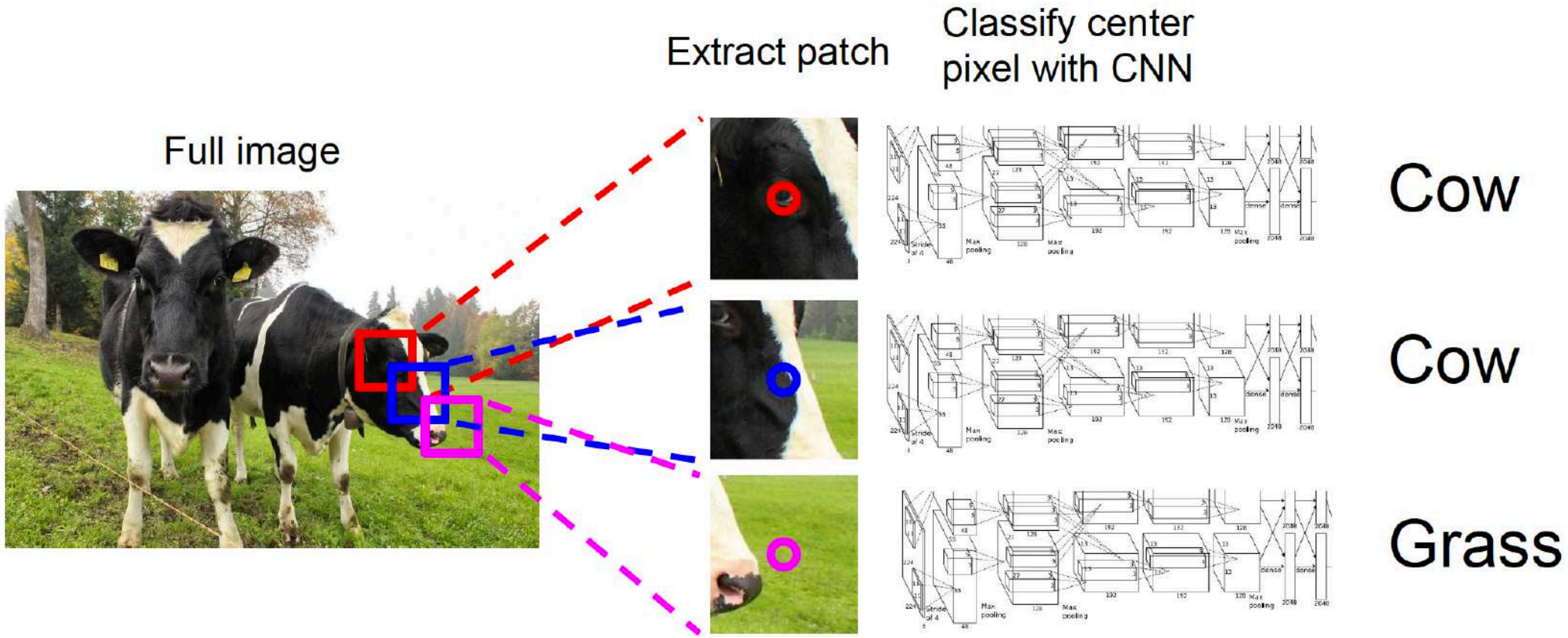
Label each pixel in the image with a category label

Don’t differentiate instances, only care about pixels





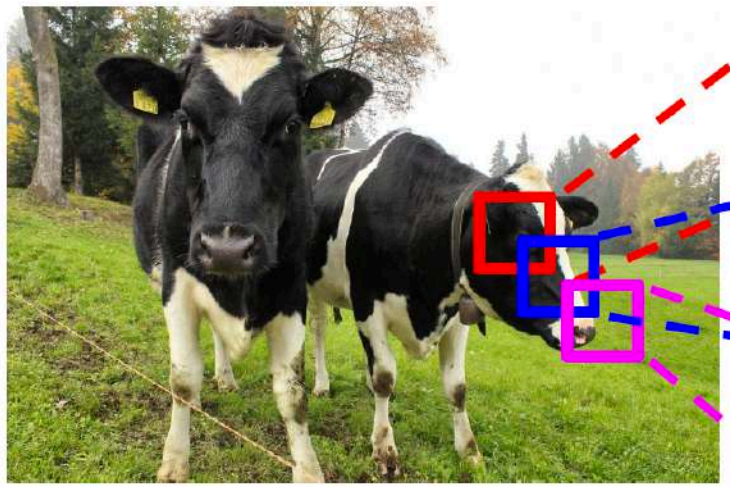
Semantic Segmentation Idea: Sliding Window



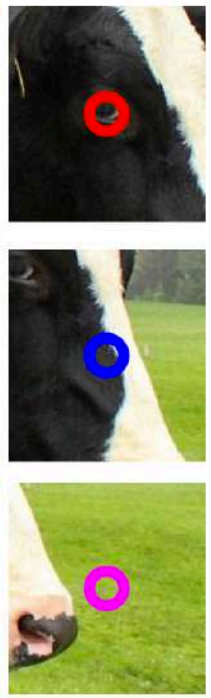
Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

Semantic Segmentation Idea: Sliding Window

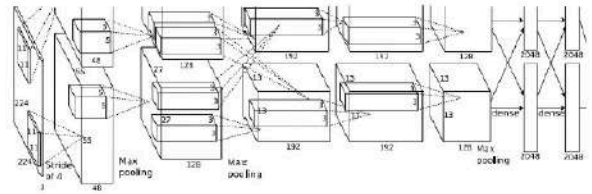
Full image



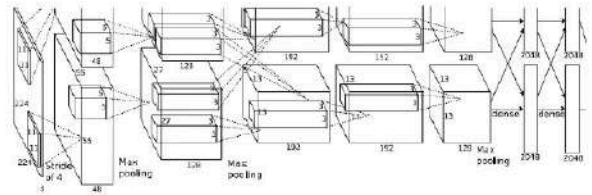
Extract patch



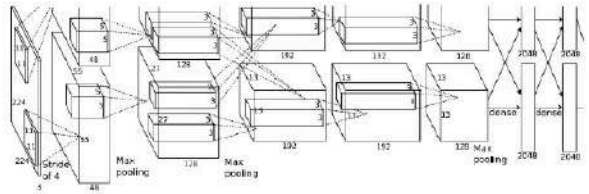
Classify center pixel with CNN



Cow



Cow



Grass

Problem: Very inefficient! Not reusing shared features between overlapping patches

Farabet et al, "Learning Hierarchical Features for Scene Labeling," TPAMI 2013
Pinheiro and Collobert, "Recurrent Convolutional Neural Networks for Scene Labeling", ICML 2014

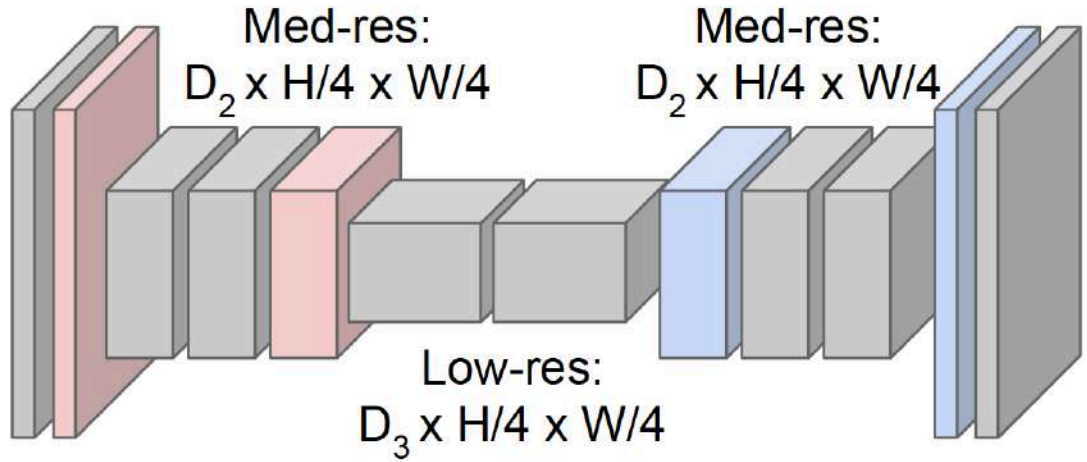
Semantic Segmentation Idea: Fully Convolutional

Downsampling:
Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling:
Unpooling or strided transpose convolution

Input:
3 x H x W



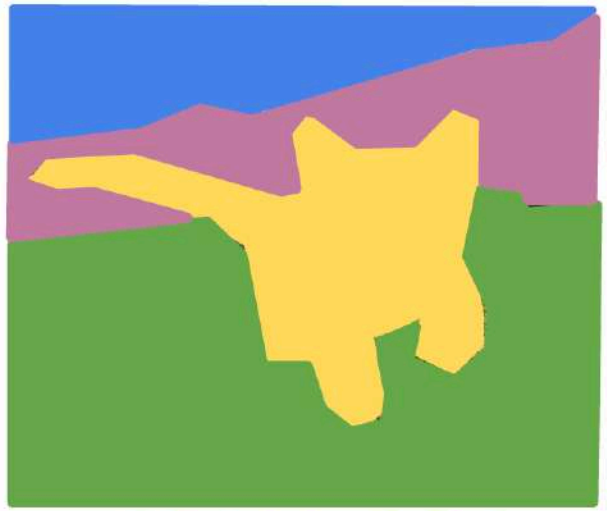
High-res:
D_1 x H/2 x W/2

High-res:
D_1 x H/2 x W/2

Predictions:
H x W

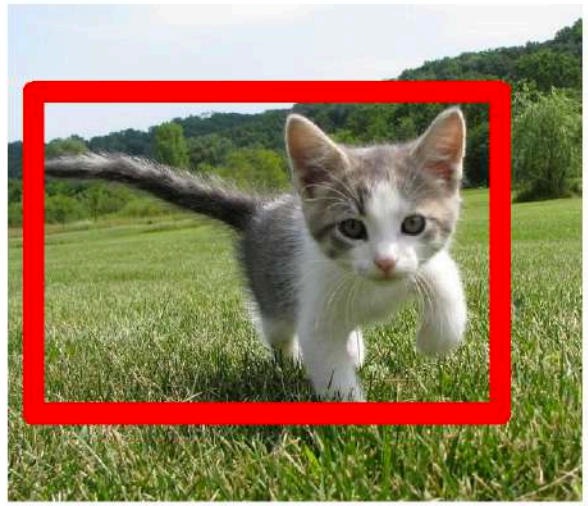
Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

Classification + Localization



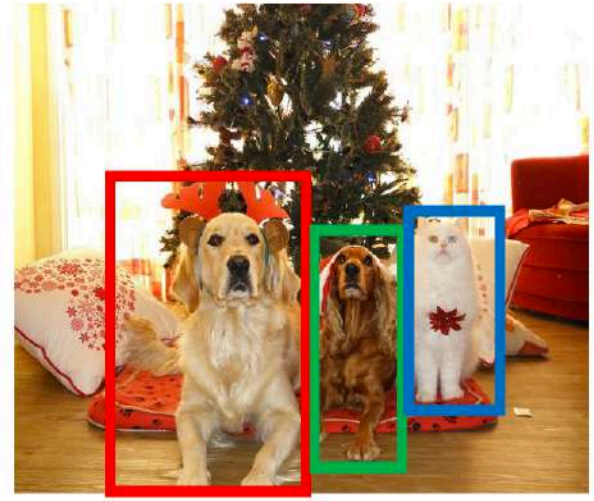
GRASS, CAT, TREE, SKY

No objects, just pixels



CAT

Single Object



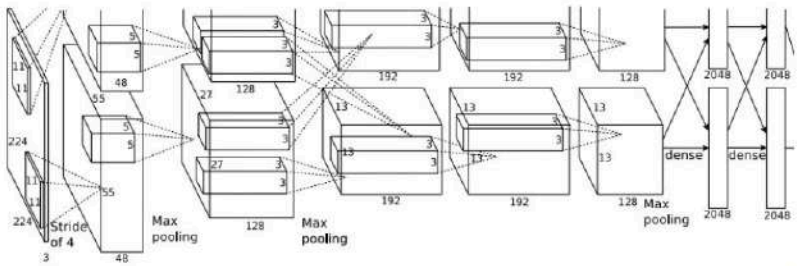
DOG, DOG, CAT

DOG, DOG, CAT

Multiple Object

Classification + Localization

This image is CC0 public domain



Vector: 4096

Fully Connected: 4096 to 1000

Class Scores
Cat: 0.9
Dog: 0.05
Car: 0.01
...

Softmax Loss

Correct label: Cat

Fully Connected: 4096 to 4

Box Coordinates (x, y, w, h)

L2 Loss

Correct box: (x', y', w', h')

Multitask Loss

+ → **Loss**

Treat localization as a regression problem!

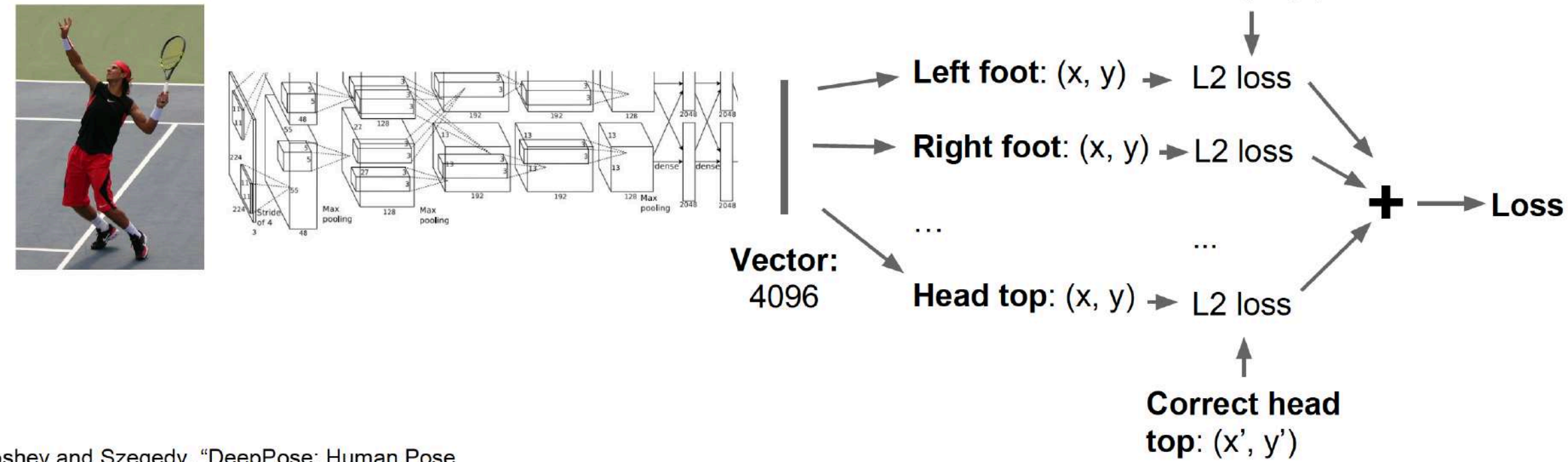
Aside: Human Pose Estimation

Represent pose as a set of 14 joint positions:

Left / right foot
Left / right knee
Left / right hip
Left / right shoulder
Left / right elbow
Left / right hand
Neck
Head top

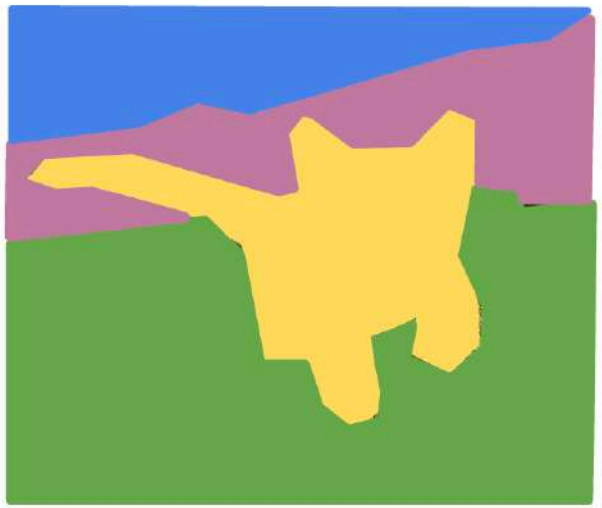
Johnson and Everingham, "Clustered Pose and Nonlinear Appearance Models for Human Pose Estimation", BMVC 2010

Aside: Human Pose Estimation



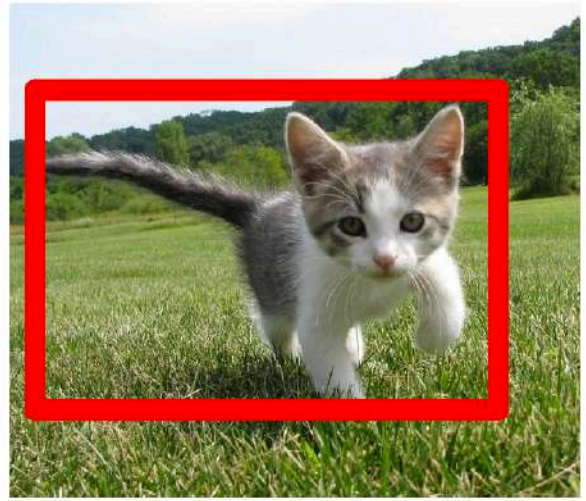
Toshev and Szegedy, "DeepPose: Human Pose Estimation via Deep Neural Networks", CVPR 2014

Object Detection



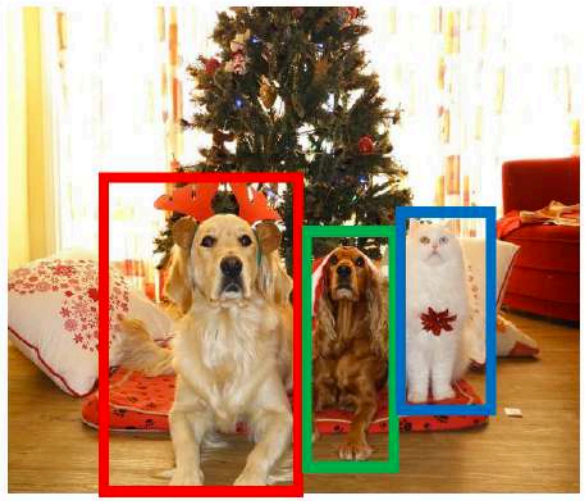
GRASS, CAT, TREE, SKY

No objects, just pixels



CAT

Single Object



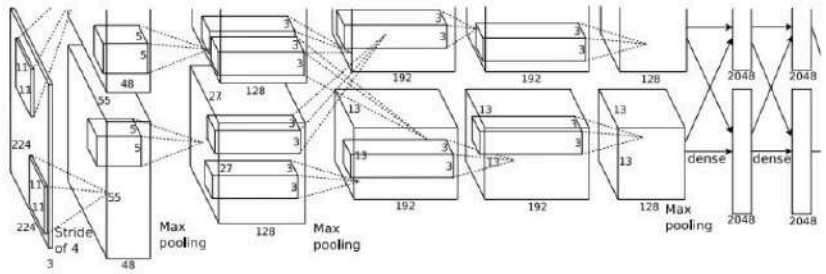
DOG, DOG, CAT

DOG, DOG, CAT

Multiple Object

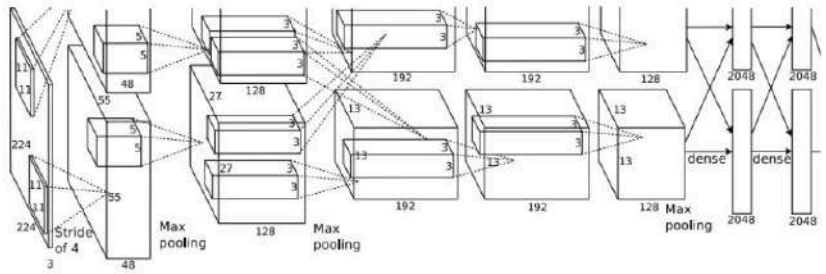
Object Detection as Regression?

Each image needs a different number of outputs!



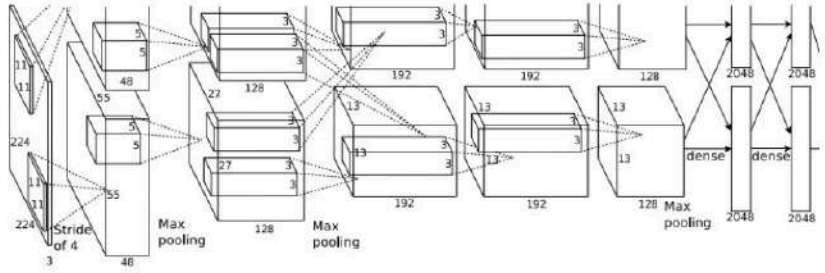
CAT: (x, y, w, h)

4 numbers



DOG: (x, y, w, h)
DOG: (x, y, w, h)
CAT: (x, y, w, h)

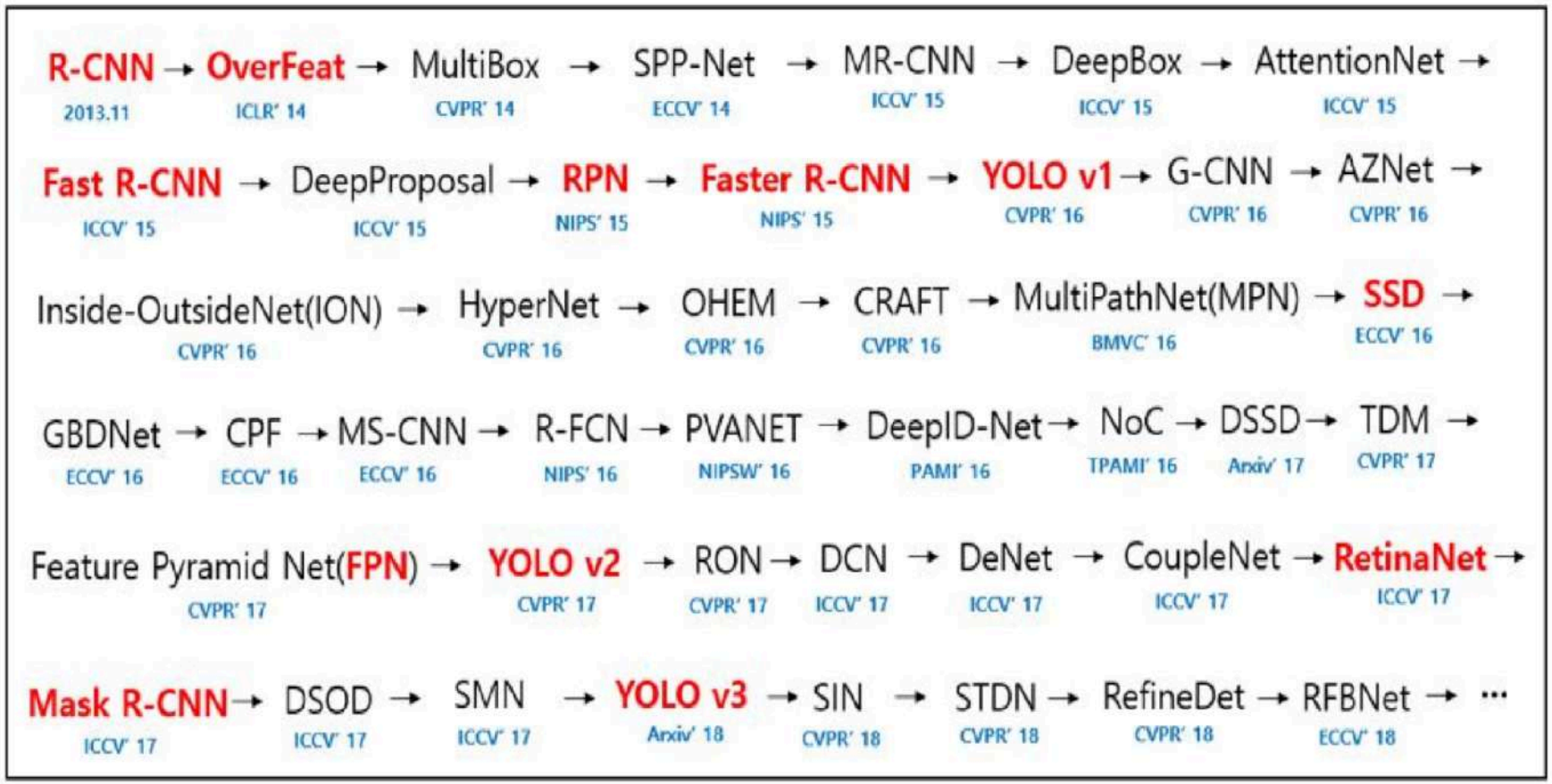
16 numbers



DUCK: (x, y, w, h)
DUCK: (x, y, w, h)
....

Many numbers!

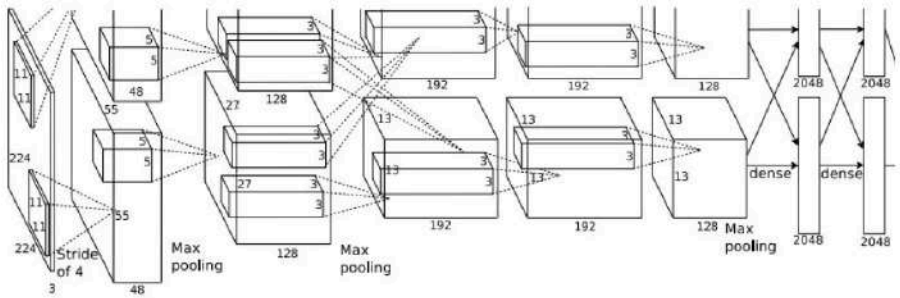
Object Detector World



- ROI
- Region Proposal
- Feature map
- ROI pooling
- RPN
- Anchors
- Multi-scale feature maps
- FPN
- ... ...

Object Detection as Classification: Sliding Window

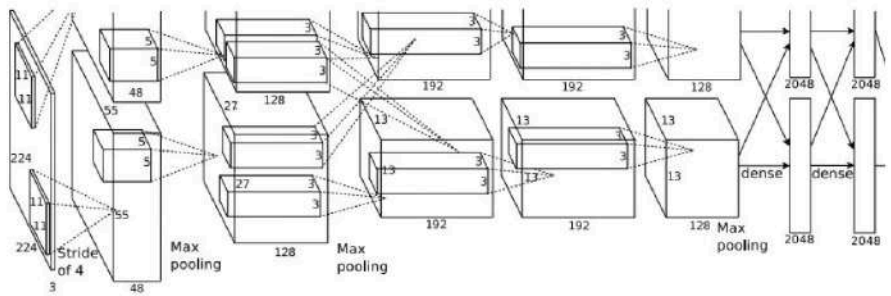
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? NO
Cat? NO
Background? YES

Object Detection as Classification: Sliding Window

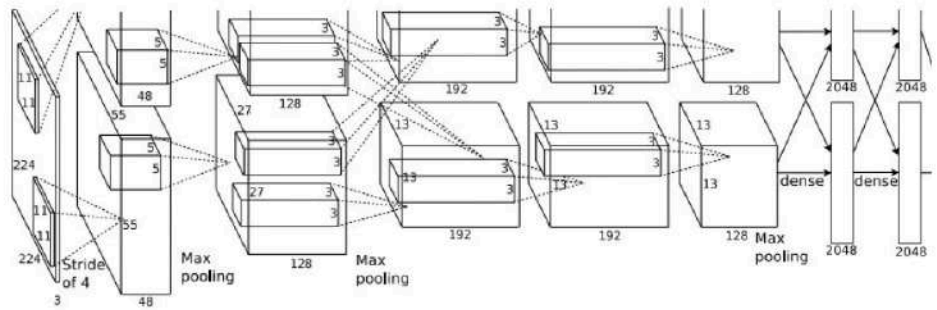
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Object Detection as Classification: Sliding Window

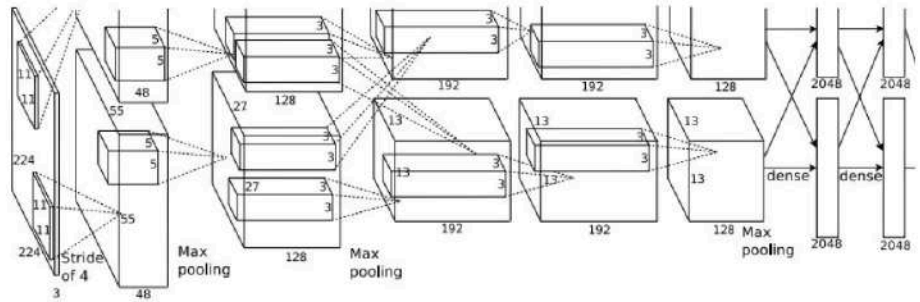
Apply a CNN to many different crops of the image, CNN classifies each crop as object or background



Dog? YES
Cat? NO
Background? NO

Object Detection as Classification: Sliding Window

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



Dog? NO
Cat? YES
Background? NO

Object Detection as Classification: Sliding Window

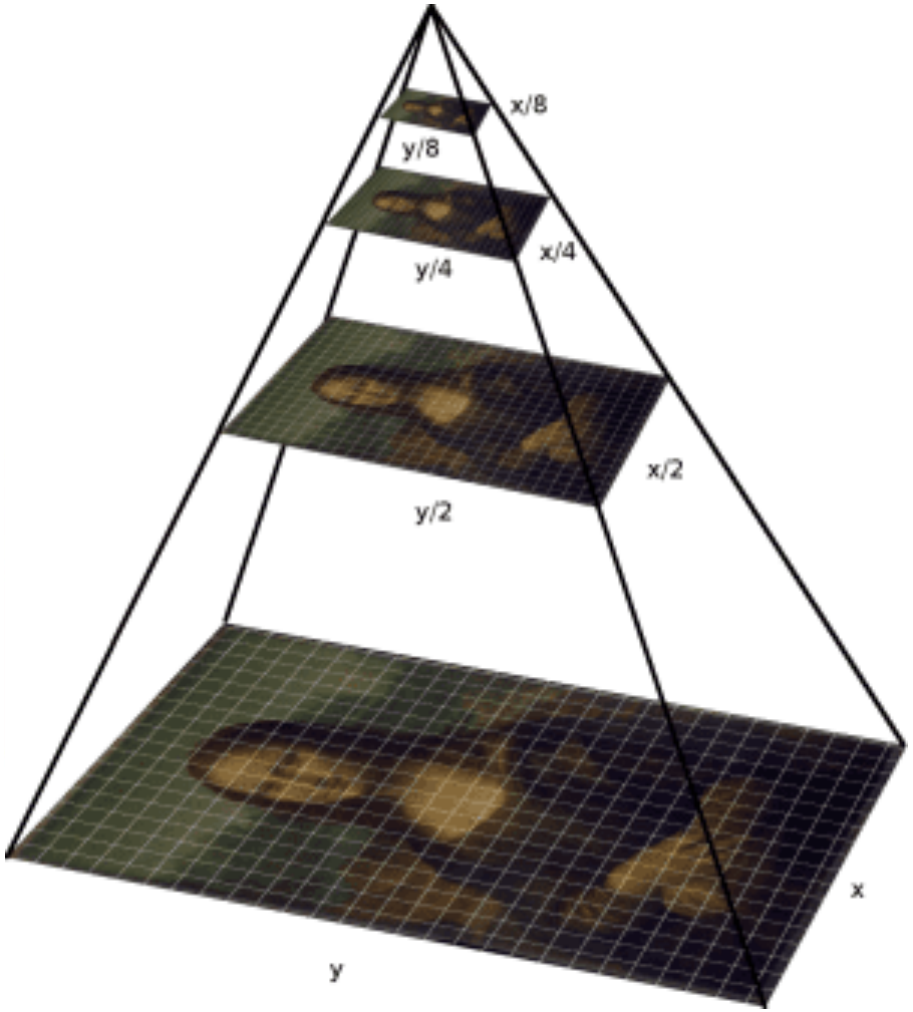
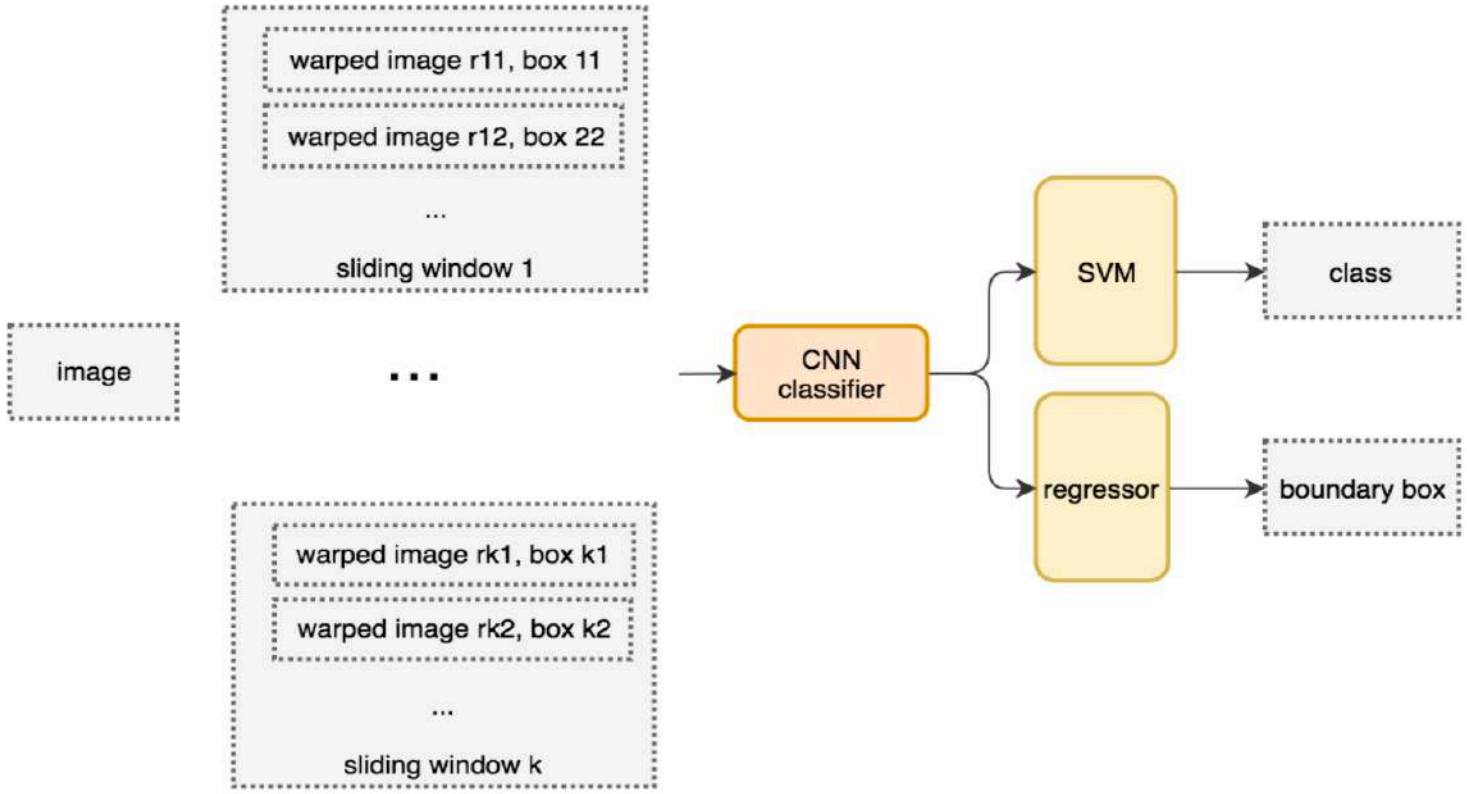


image pyramids create a multi-scale representation of an input image, allowing us to detect objects at multiple scales/sizes:

Sliding Windows – Architecture view



Problem:

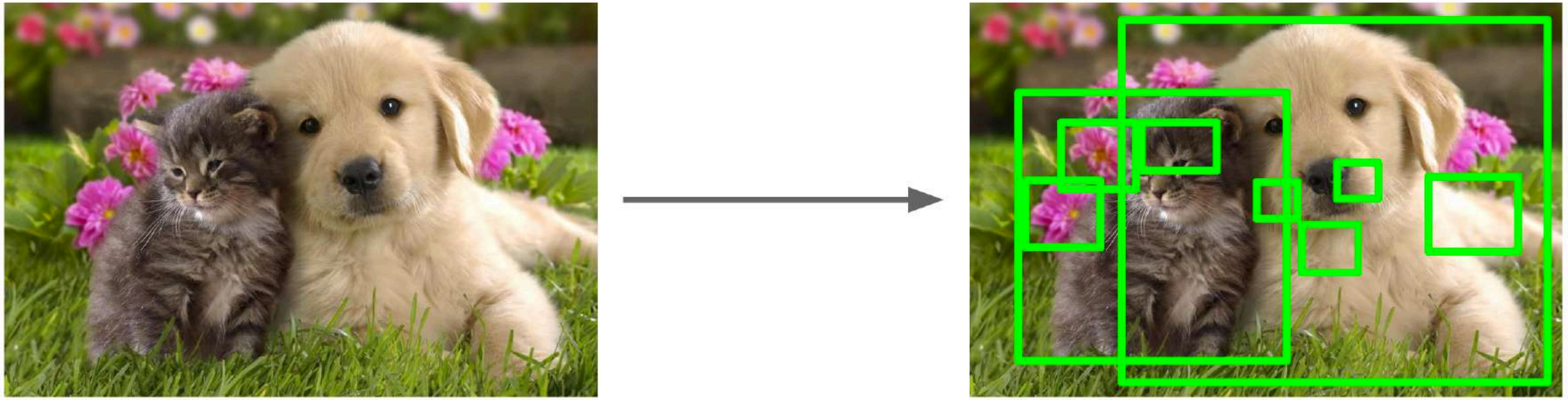
1. Too many windows
2. Windows without interested objects
3. Hard to find appropriate window size

To improve performance:

Reduce the number of *windows*.

Region Proposals

- 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 objectness 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

Selective Search — One Region Proposal Algorithm

- Introduced by Uijlings et al. in their 2013 paper

[HTML] Selective search for object recognition

JRR Uijlings, KEA Van De Sande, T Gevers… - International journal of …, 2013 - Springer

This paper addresses the problem of generating possible object locations for use in object recognition. We introduce selective search which combines the strength of both an exhaustive search and segmentation. Like segmentation, we use the image structure to …

☆ Save 99 Cite Cited by 5889 Related articles All 38 versions Import into BibTeX

- First over-segmenting an image using a superpixel algorithm

Efficient graph-based image segmentation

PF Felzenszwalb, DP Huttenlocher - International journal of computer …, 2004 - Springer

This paper addresses the problem of segmenting an image into regions. We define a predicate for measuring the evidence for a boundary between two regions using a graph-based representation of the image. We then develop an efficient segmentation algorithm …

☆ Save 99 Cite Cited by 7372 Related articles All 47 versions »

Selective Search — One Region Proposal Algorithm

- Next, seeks to merge together the superpixels to find regions of an image that *could* contain an object using a linear combination of five key similarity measures:

Color similarity: Computing a 25-bin histogram for each channel of an image, concatenating them together, and obtaining a final descriptor. Color similarity of any two regions is measured by the histogram intersection distance.

Texture similarity: Extracts Gaussian derivatives at 8 orientations per channel (assuming a 3-channel image). These orientations are used to compute a 10-bin histogram per channel, generating a final texture descriptor. Texture similarity between any two regions is measured by histogram intersection.

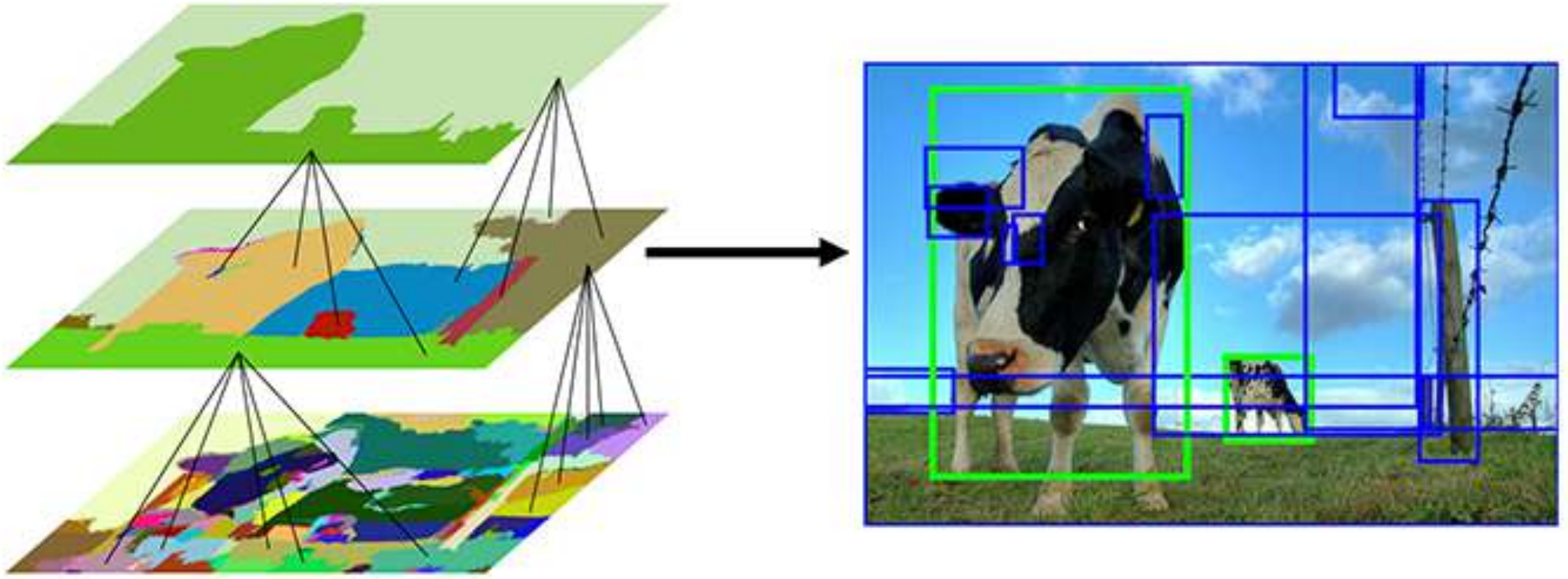
Size similarity: Prefers that smaller regions be grouped *earlier* rather than *later.*

Shape similarity/compatibility: Two merged regions should be *compatible* with each other. Two regions are considered “compatible” if they “fit” into each other. Shapes that do not touch should not be merged.

A final meta-similarity measure: a linear combination of the color similarity, texture similarity, size similarity, and shape similarity/compatibility.

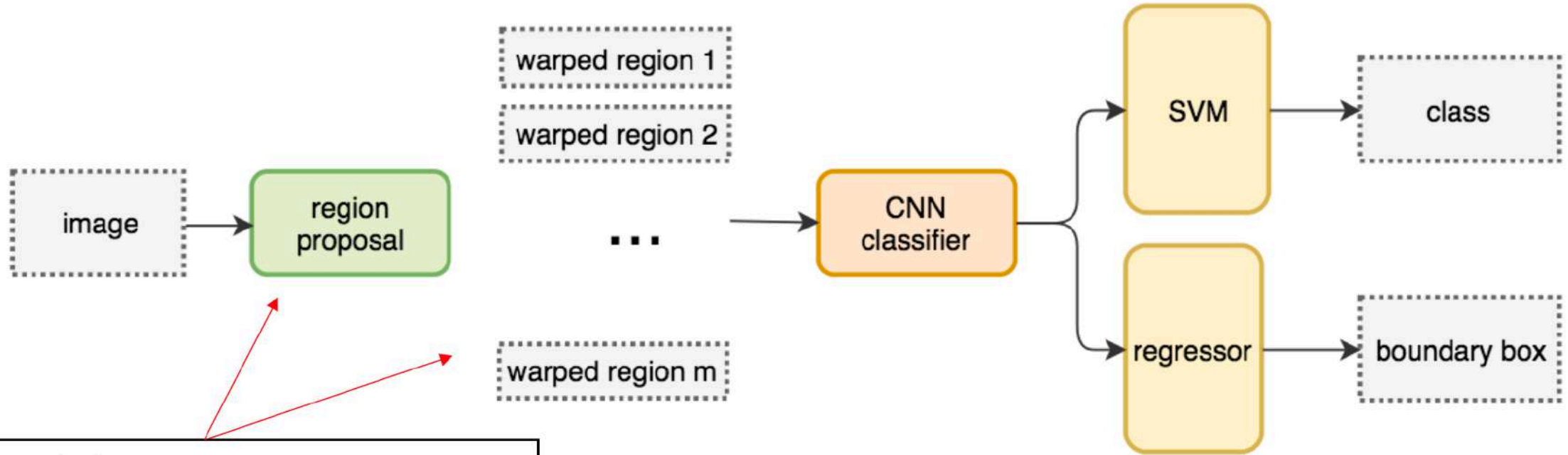
Selective Search — One Region Proposal Algorithm

- Region Proposals generated by the Selective Search Algorithm



Selective Search — One Region Proposal Algorithm

- Region Proposals generated by the Selective Search Algorithm



- No *windows* anymore;
- Get Region of Interest (ROI) via some external **region proposal method**, and extract features from ROIs instead of images

R-CNN

Input image

Girshick et al, “Rich feature hierarchies for accurate object detection and semantic segmentation”, CVPR 2014.

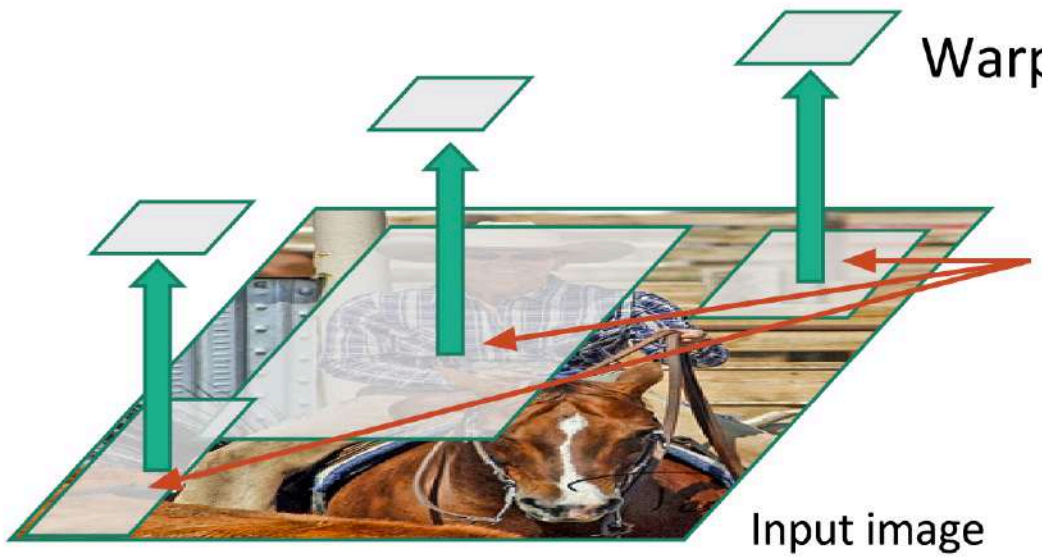
R-CNN

Input image

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

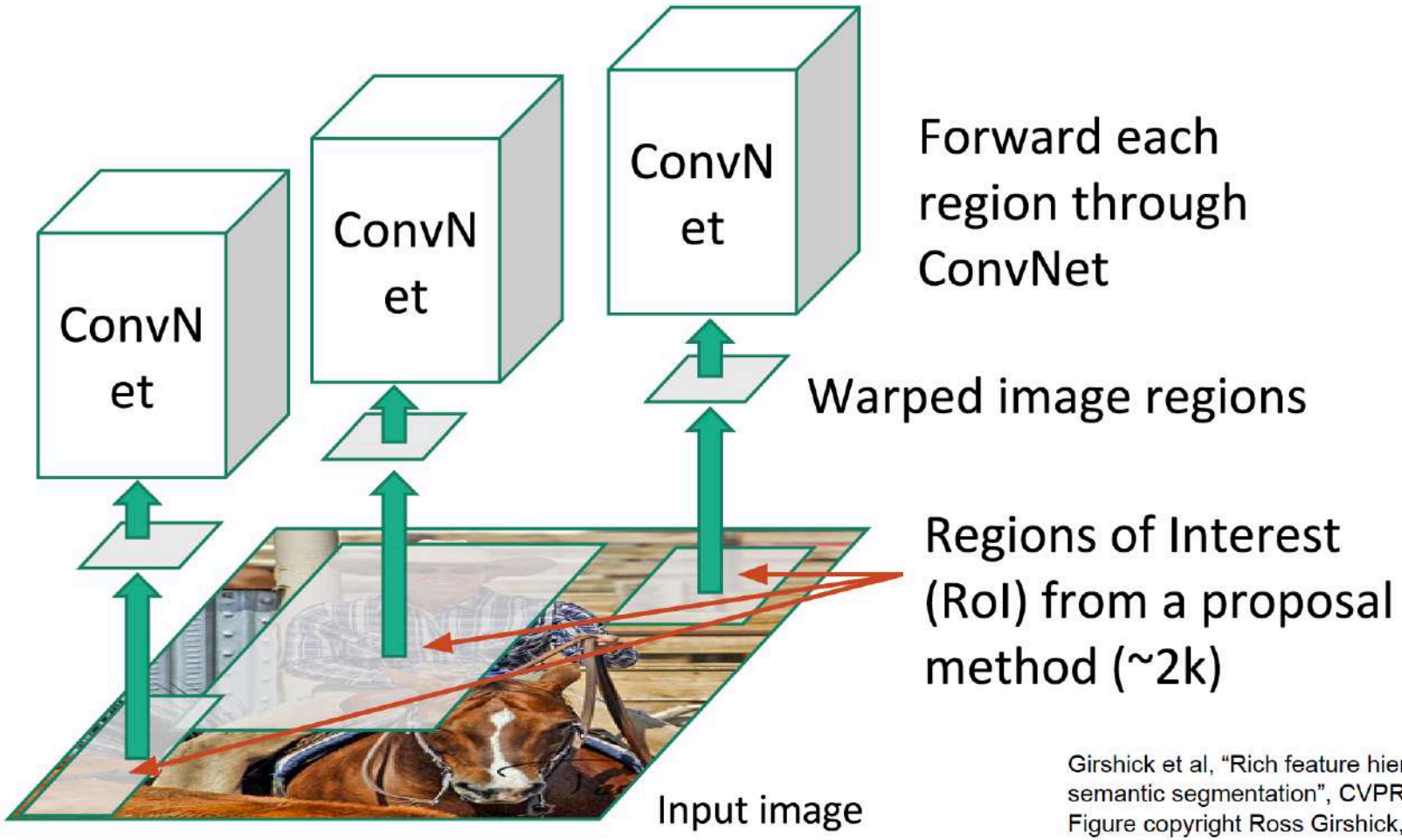
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

R-CNN



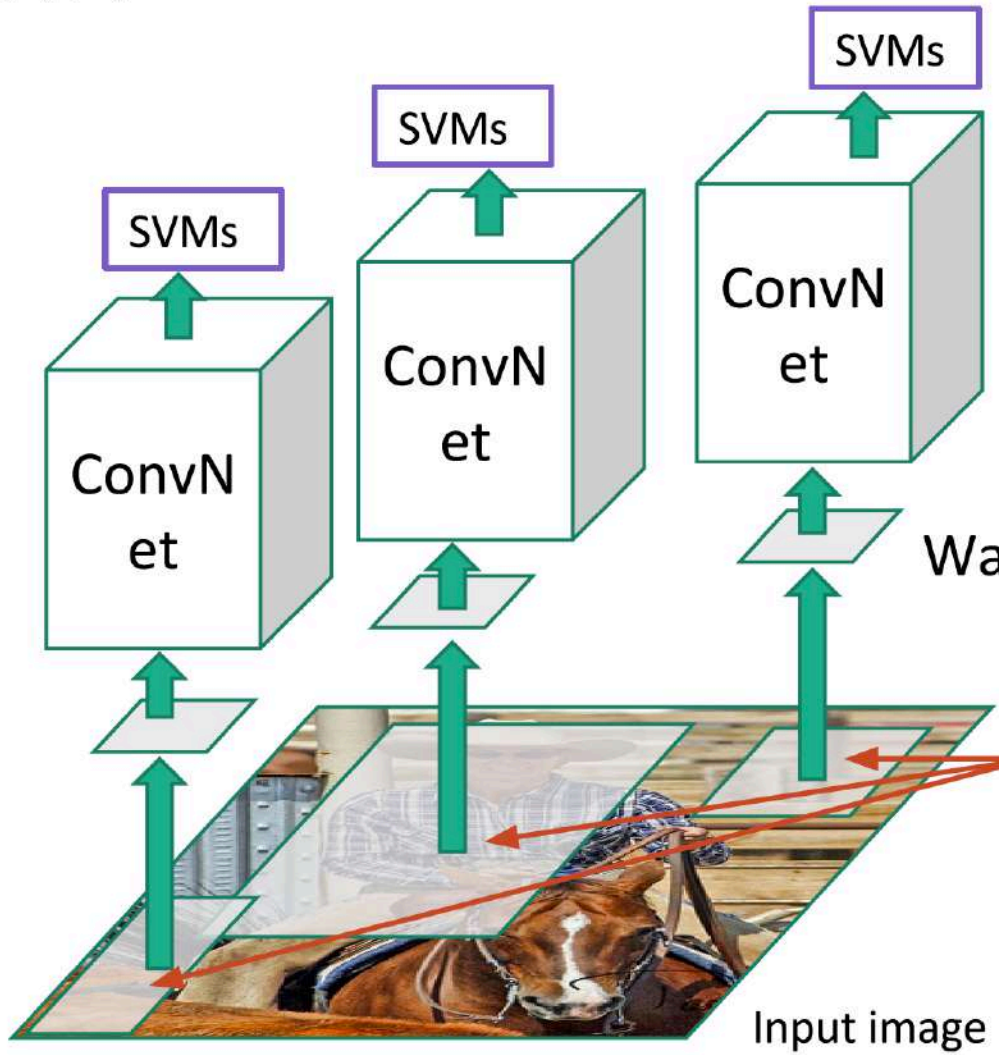
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

R-CNN



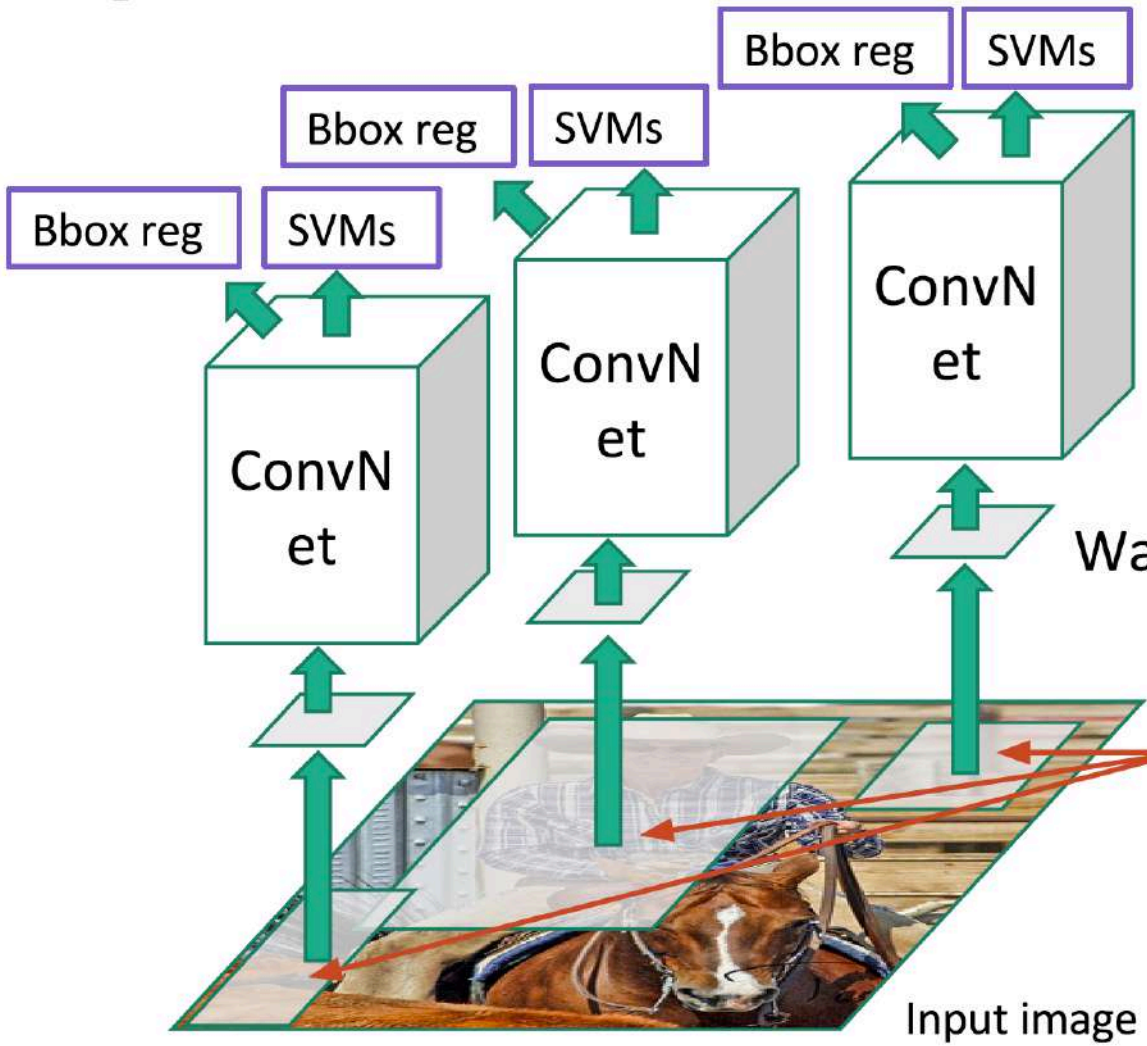
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

R-CNN



Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

R-CNN



Linear Regression for bounding box offsets

Classify regions with SVMs

Forward each region through ConvNet

Warped image regions

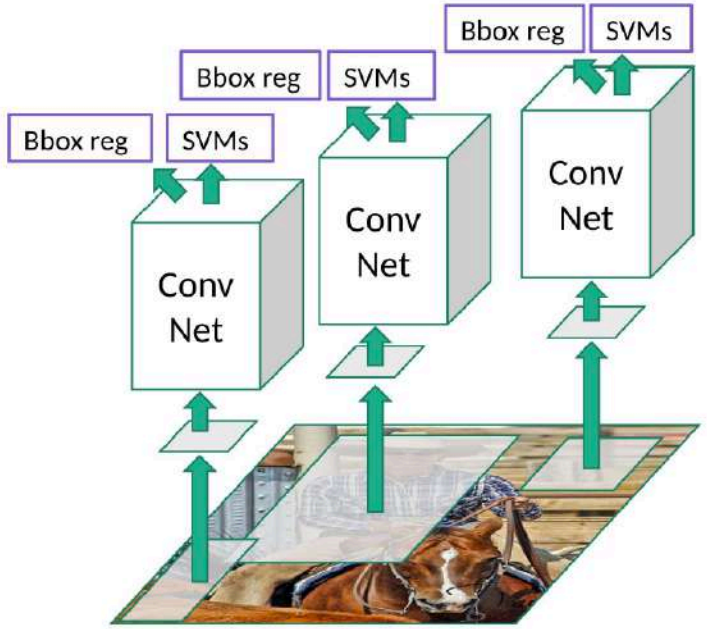
Regions of Interest (RoI) from a proposal method (~2k)

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.

R-CNN: Problems

- Painfully slow (84 hours)
- Takes a lot of disk space
- Inference (detection) is slow
  - 47s/image with VGG16
  - Some improvements made by SPP-net



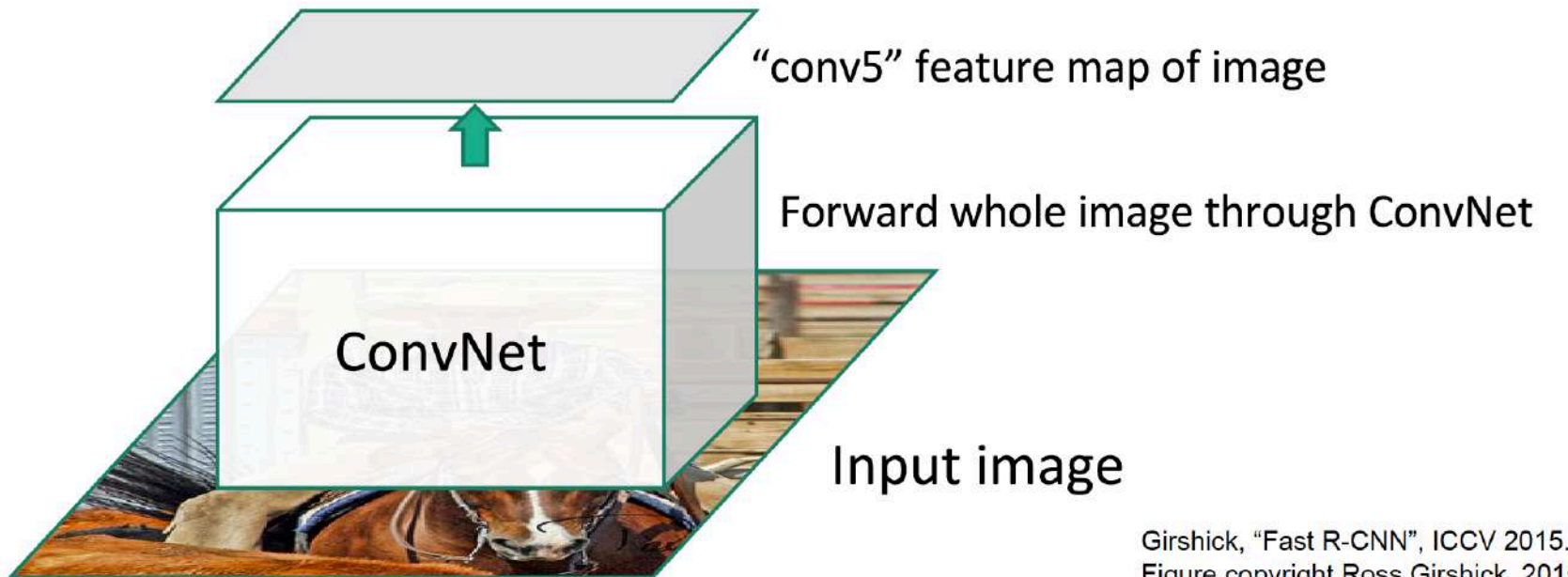
Girshick et al, "Rich feature hierarchies for accurate object detection and semantic segmentation", CVPR 2014.

Fast R-CNN

Input image

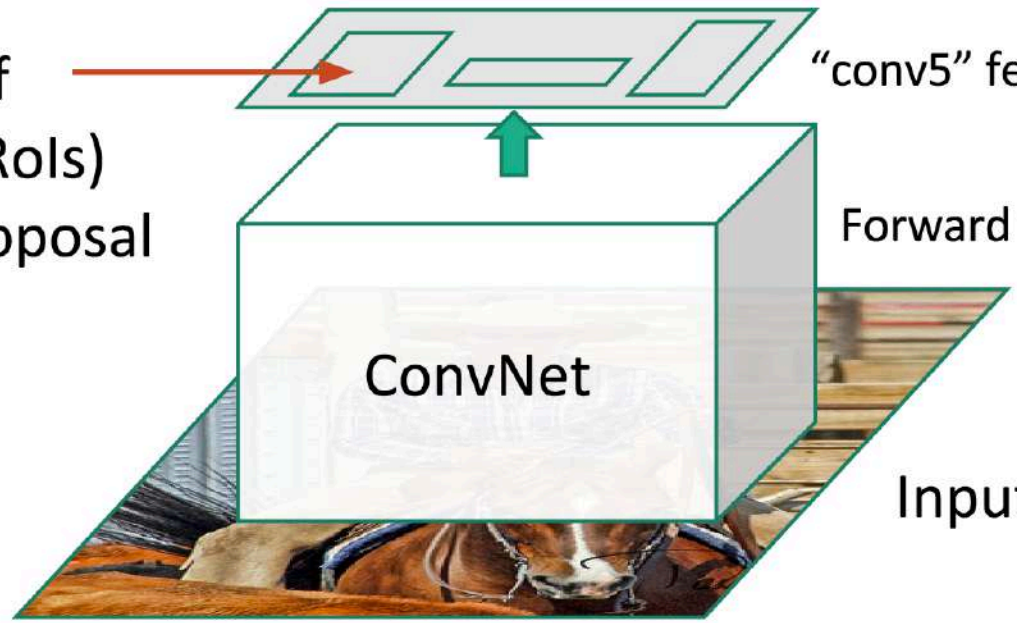
Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



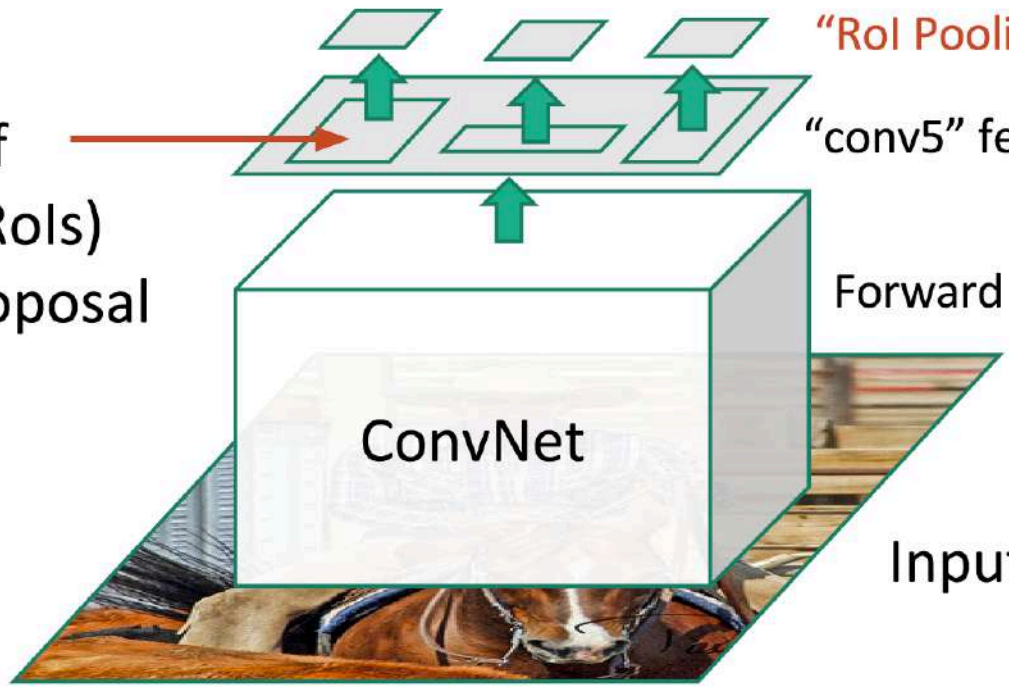
Girshick, "Fast R-CNN", ICCV 2015.

Fast R-CNN



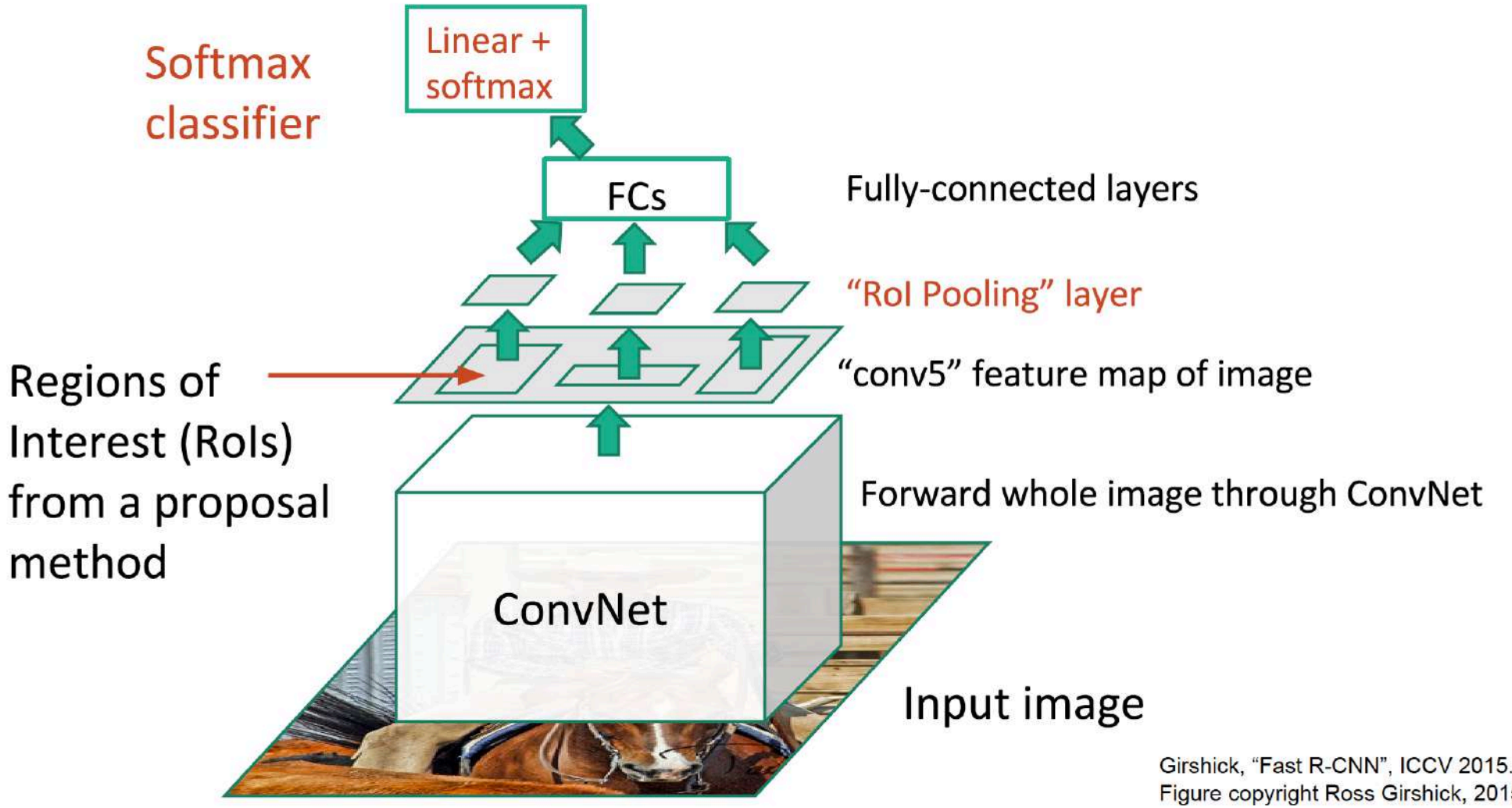
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast R-CNN



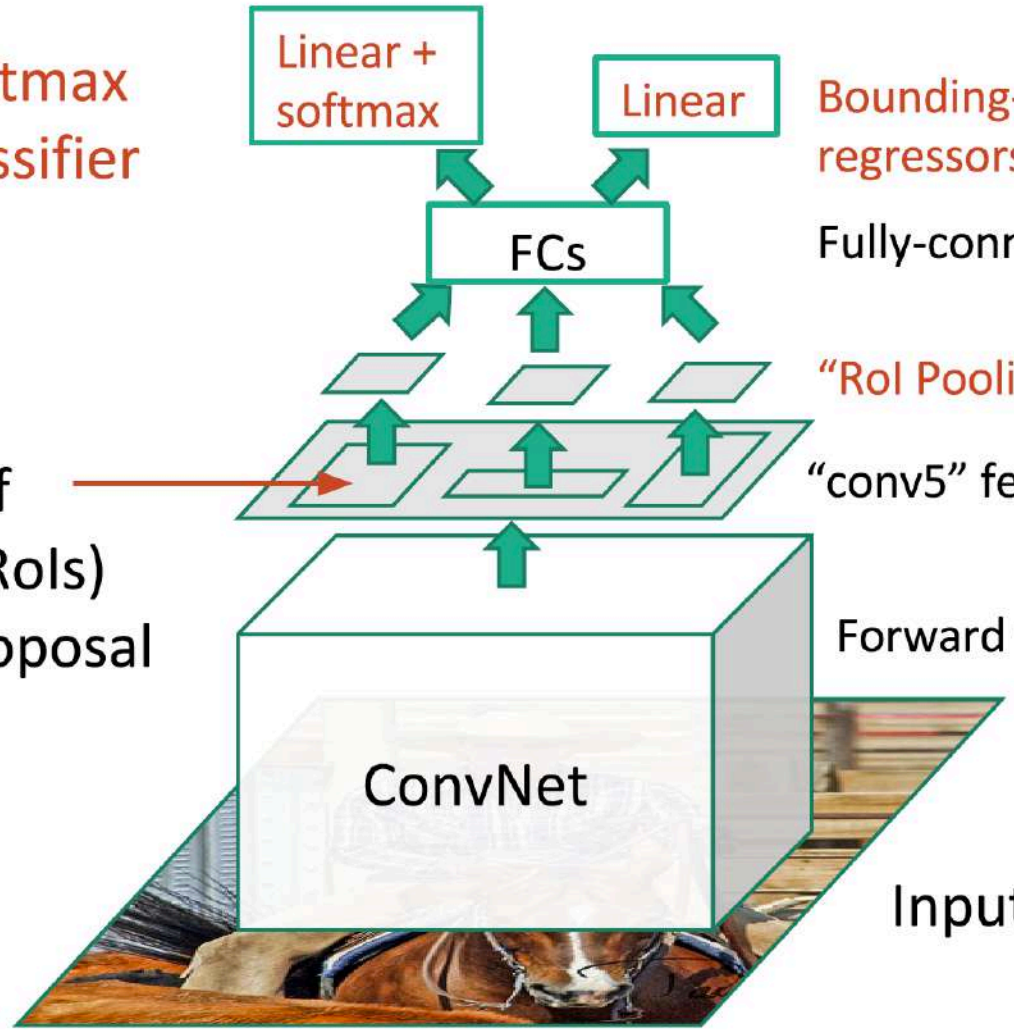
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast R-CNN



Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Fast R-CNN



Softmax classifier

Bounding-box regressors

Fully-connected layers

"RoI Pooling" layer

"conv5" feature map of image

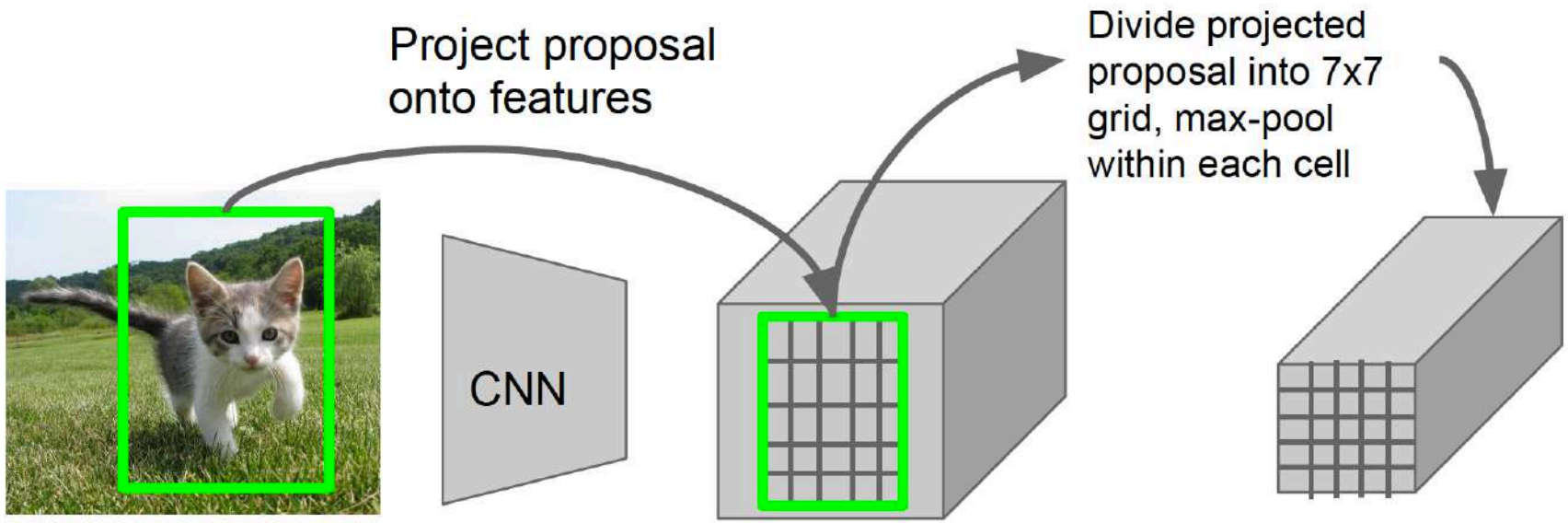
Regions of Interest (RoIs) from a proposal method

Forward whole image through ConvNet

Input image

Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

Faster R-CNN: RoI Pooling



Hi-res input image:
3 x 640 x 480
with region
proposal

Hi-res conv features:
512 x 20 x 15;

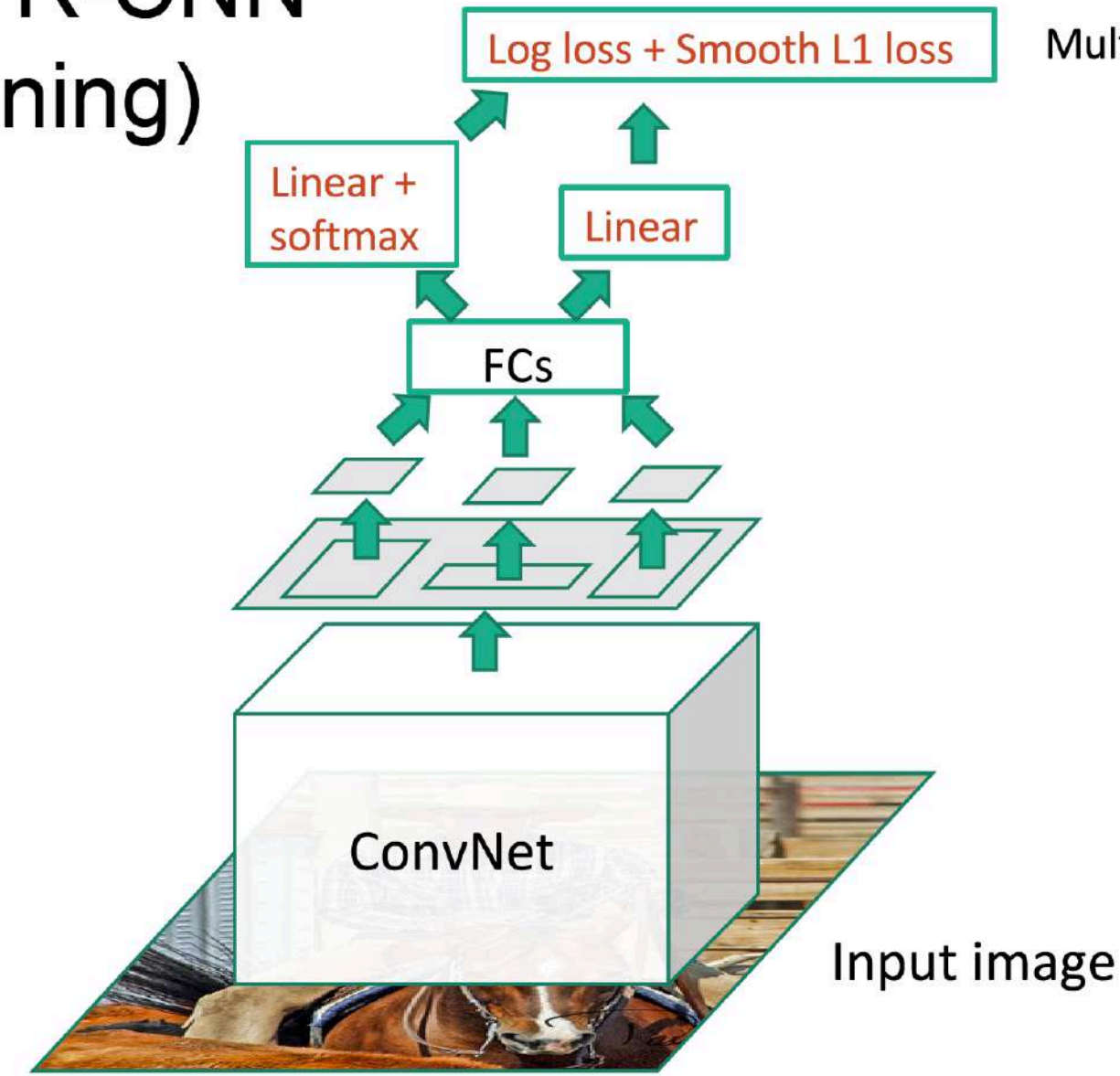
Projected region
proposal is e.g.
512 x 18 x 8
(varies per proposal)

RoI conv features:
512 x 7 x 7
for region proposal

Fully-connected layers expect
low-res conv features:
512 x 7 x 7

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

Fast R-CNN (Training)

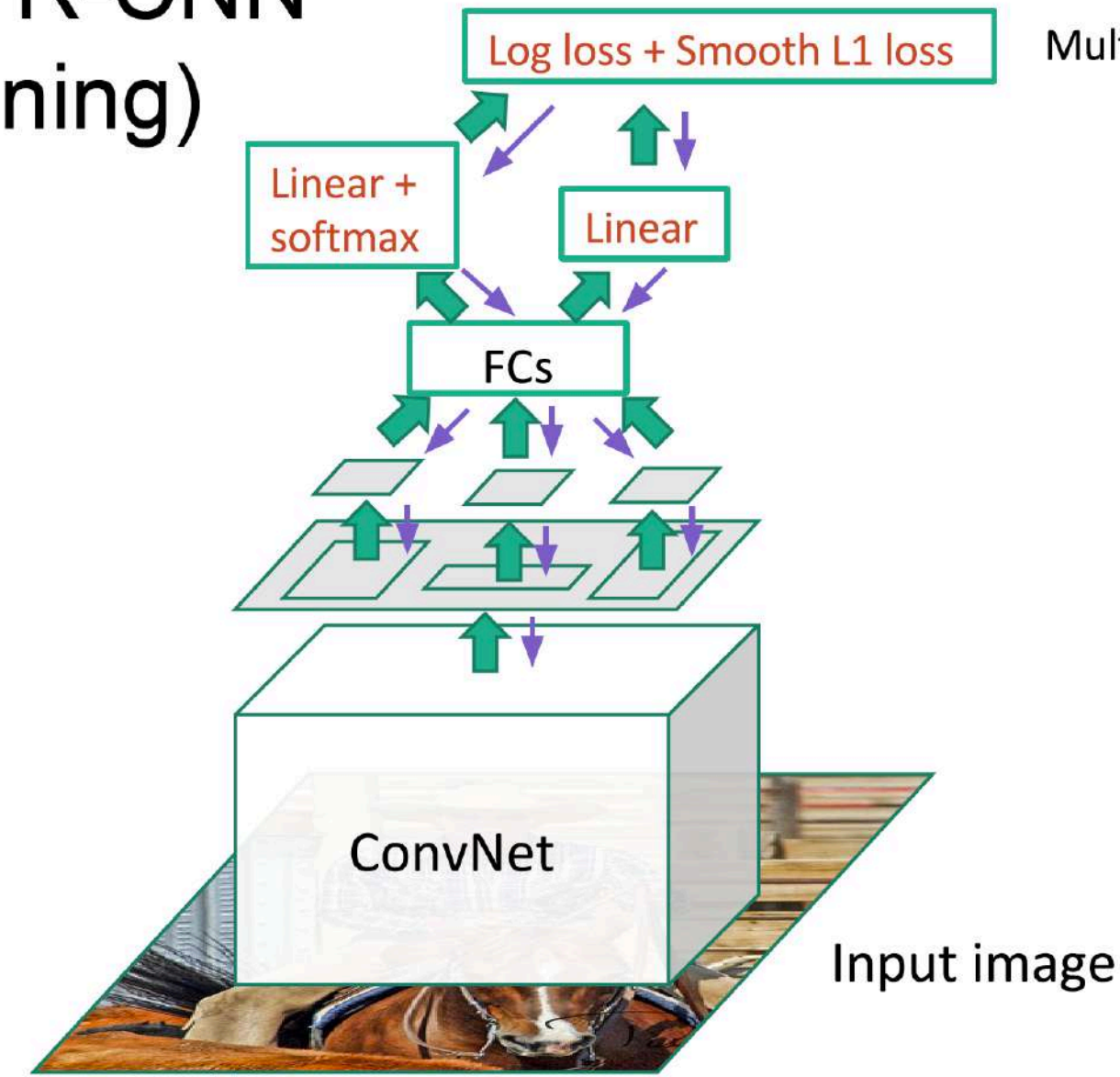


Multi-task loss

Input image

Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

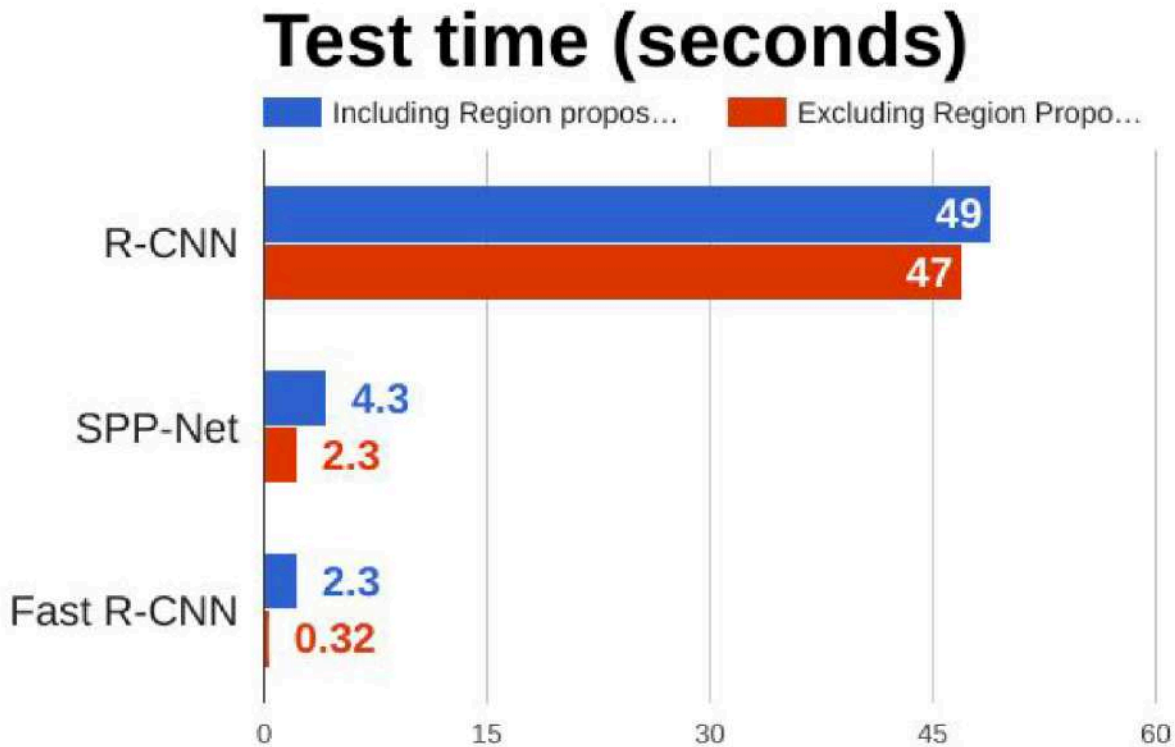
Fast R-CNN (Training)



Multi-task loss

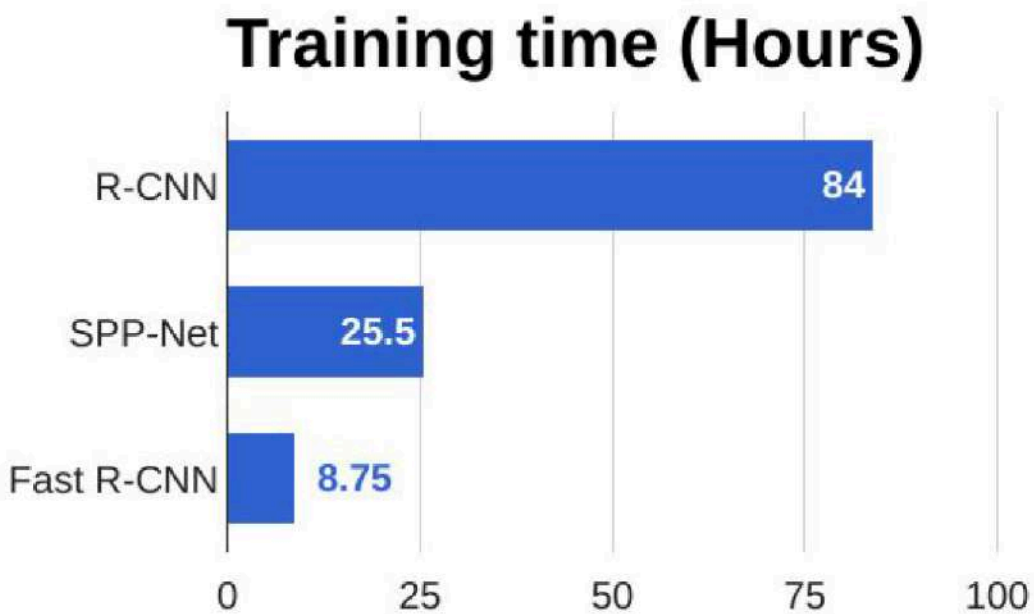
Girshick, "Fast R-CNN", ICCV 2015.
Figure copyright Ross Girshick, 2015; source. Reproduced with permission.

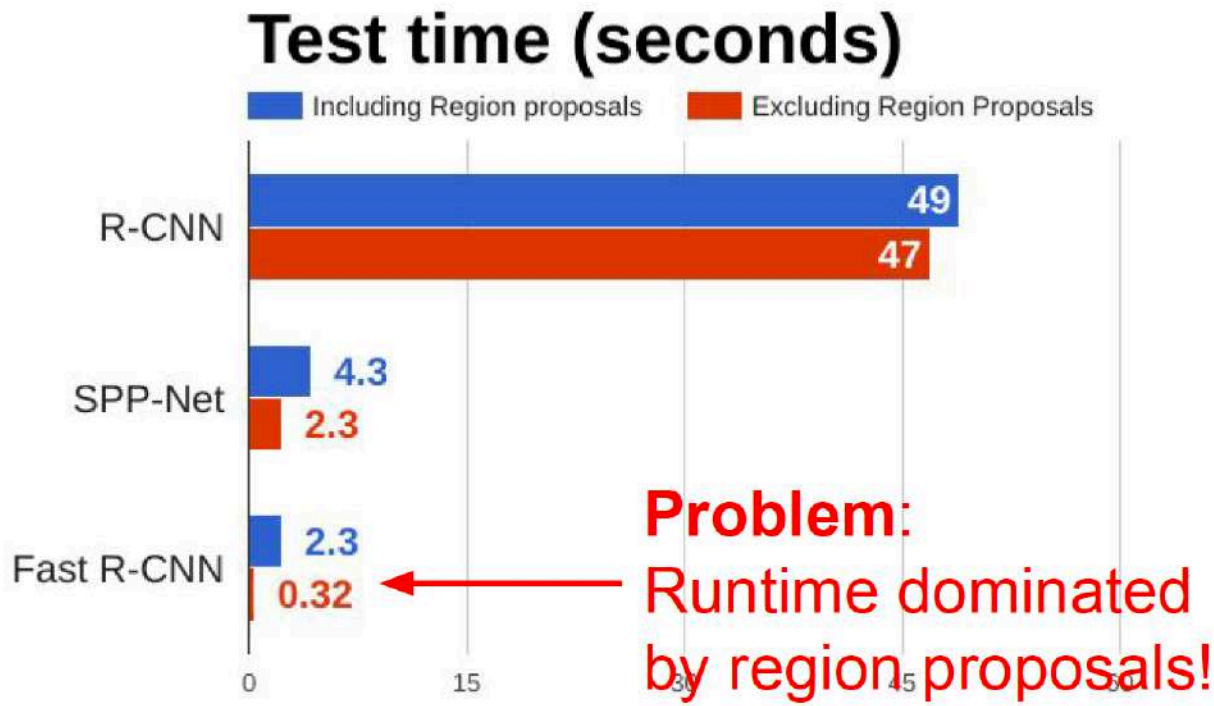
R-CNN vs SPP vs Fast R-CNN



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

R-CNN vs SPP vs Fast R-CNN





Problem:
Runtime dominated by region proposals!

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

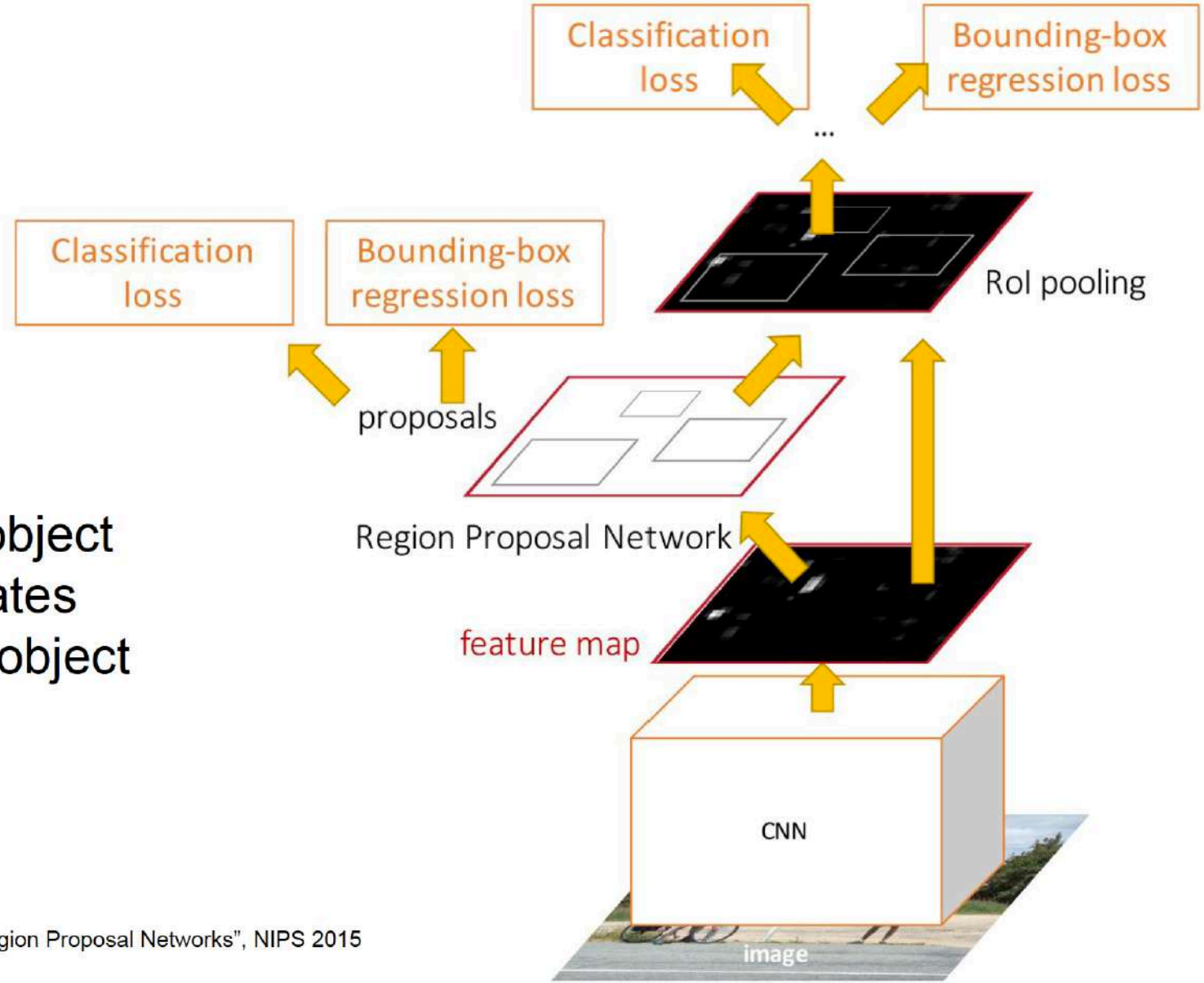
Faster R-CNN:

Make CNN do proposals!

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

Jointly train with 4 losses:

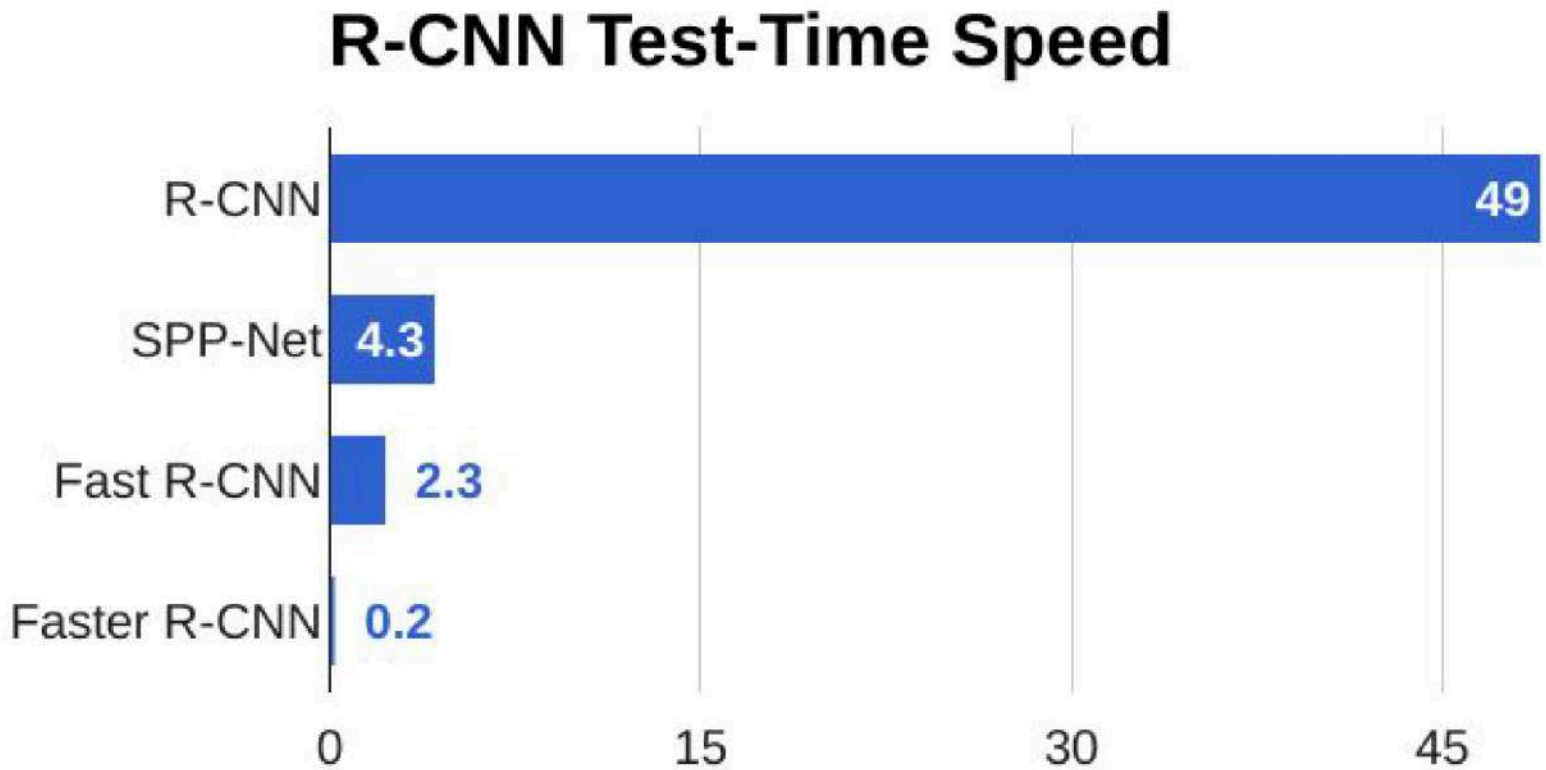
1. RPN classify object / not object
2. RPN regress box coordinates
3. Final classification score (object classes)
4. Final box coordinates



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

Faster R-CNN:

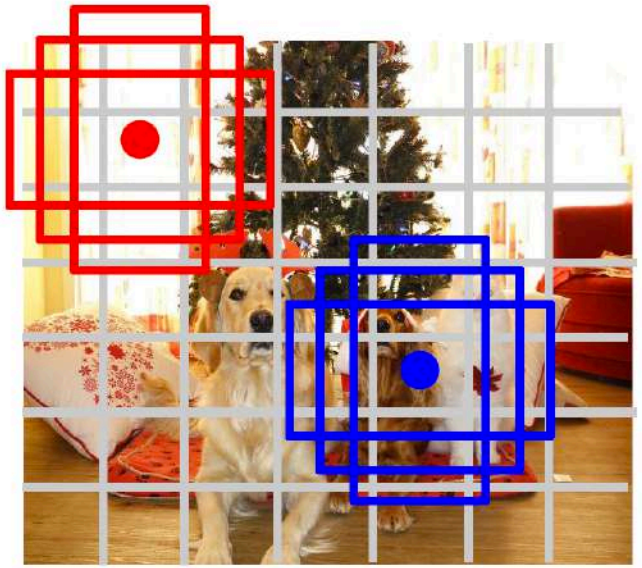
Make CNN do proposals!



Detection without Proposals: YOLO / SSD

Go from input image to tensor of scores with one big convolutional network!

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, confidence)
- Predict scores for each of C classes (including background as a class)

Output:
7 x 7 x (5 * B + C)

Redmon et al, "You Only Look Once: Unified, Real-Time Object Detection", CVPR 2016
Liu et al, "SSD: Single-Shot MultiBox Detector", ECCV 2016

Over 5 versions (3 official) and cited over **20,000 times.**

Joseph Redmon @pjreddie · Feb 20, 2020

"We shouldn't have to think about the societal impact of our work because it's hard and other people can do it for us" is a really bad argument.

> 
> **Roger Grosse** @RogerGrosse
>
> Replying to @kevin_zakka and @hardmaru
>
> To be clear, I don't think this is a positive step. Societal impacts of AI is a tough field, and there are researchers and organizations that study it professionally. Most authors do not have expertise in the area and won't do good enough scholarship to say something meaningful.

Joseph Redmon
@pjreddie

I stopped doing CV research because I saw the impact my work was having. I loved the work but the military applications and privacy concerns eventually became impossible to ignore.

Object Detection: Lots of variables ...

Base Network
VGG16
ResNet-101
Inception V2
Inception V3
Inception
ResNet
MobileNet

Object Detection architecture
Faster R-CNN
R-FCN
SSD

Image Size
 # Region Proposals
...

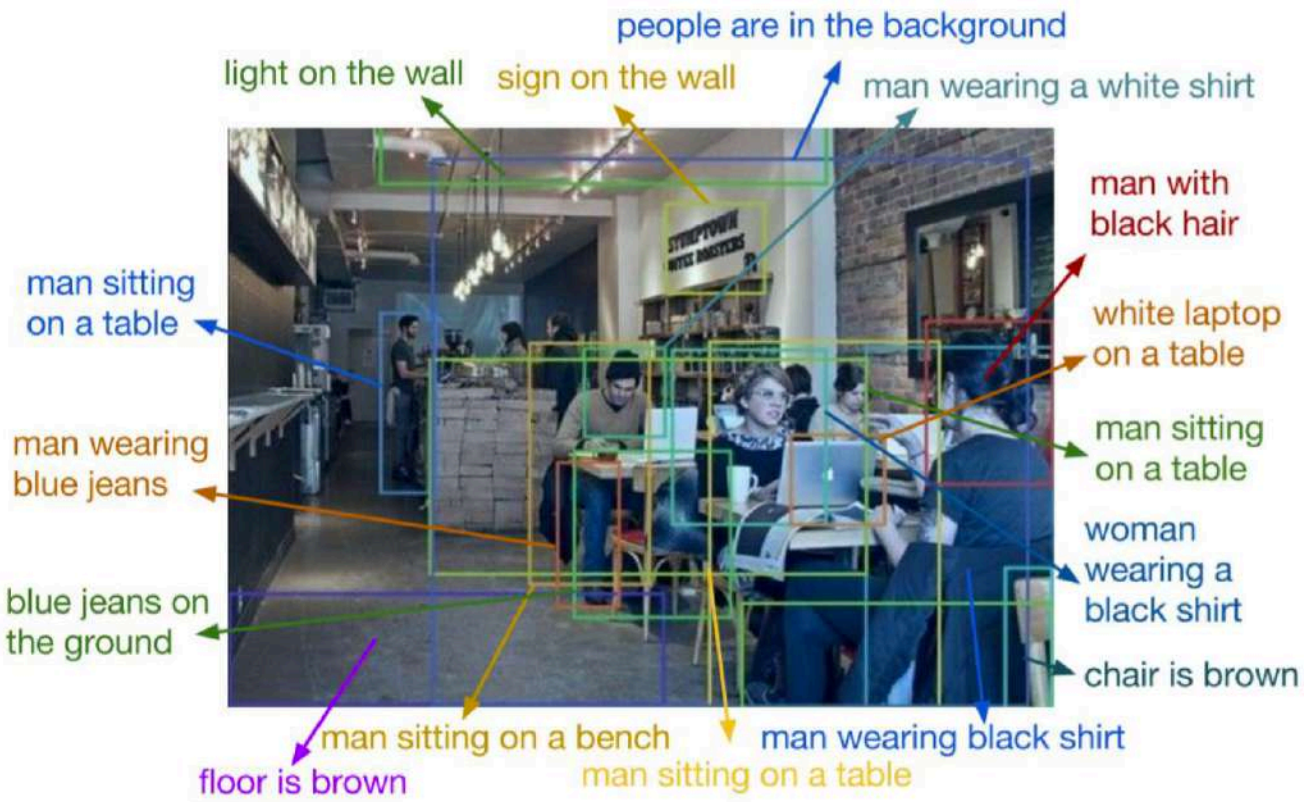
Takeaways
Faster R-CNN is slower but more accurate

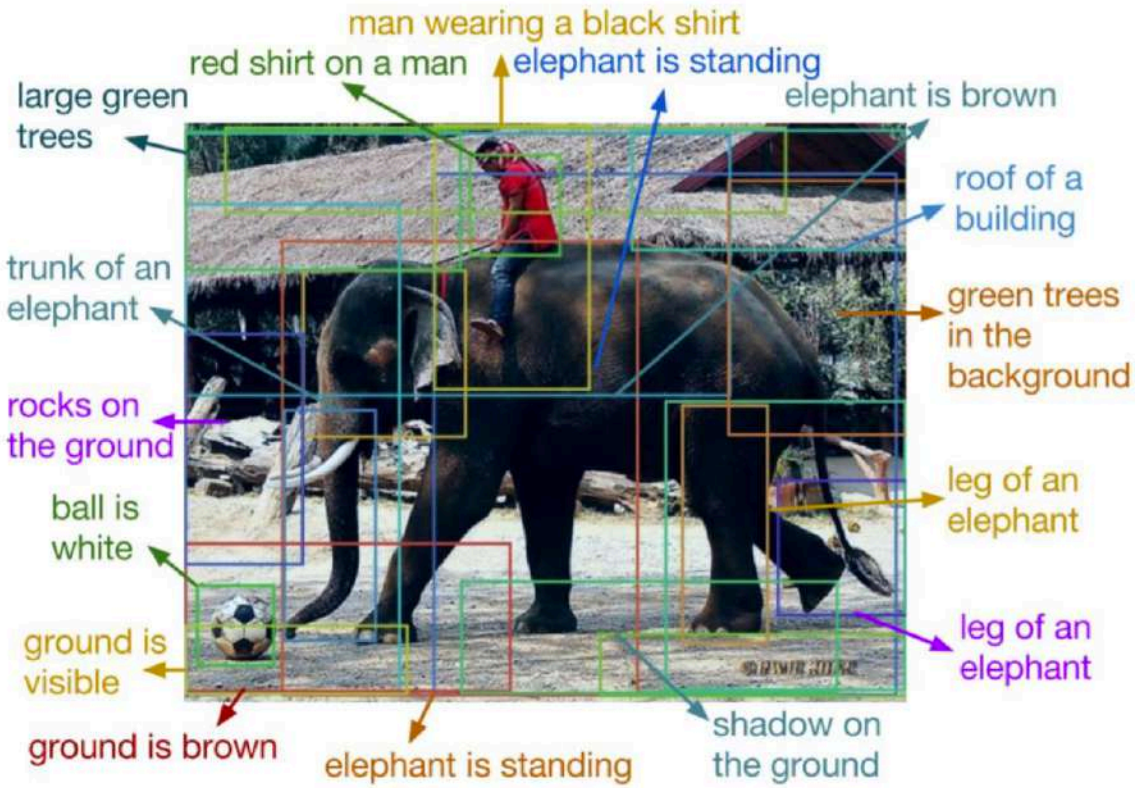
SSD is much faster but not as accurate

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

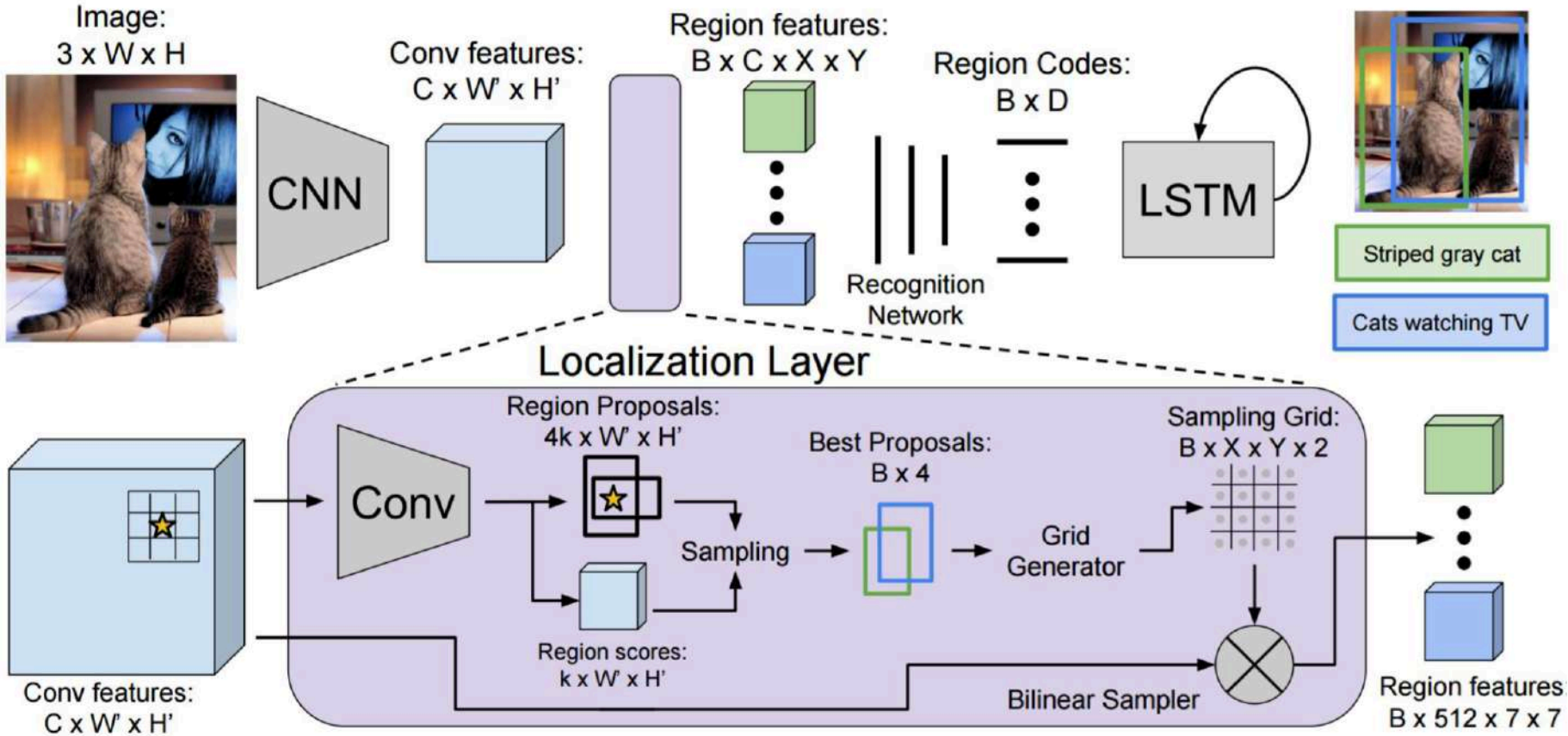
Aside: Object Detection + Captioning = Dense Captioning





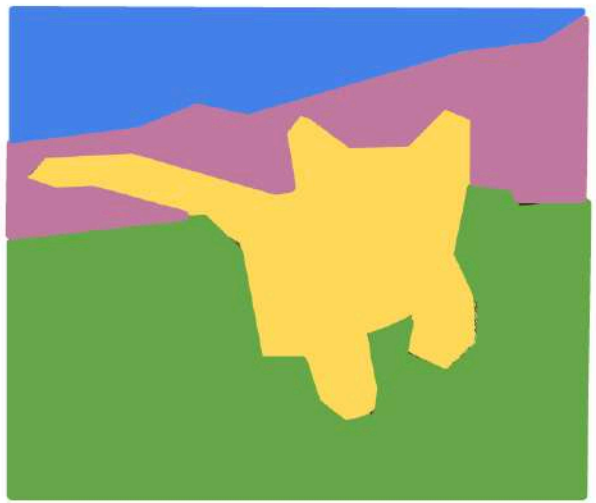
Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016

Aside: Object Detection + Captioning = Dense Captioning



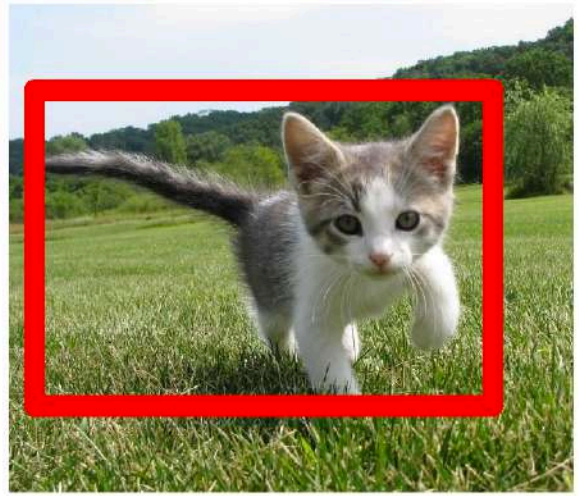
Johnson, Karpathy, and Fei-Fei, "DenseCap: Fully Convolutional Localization Networks for Dense Captioning", CVPR 2016

Instance Segmentation



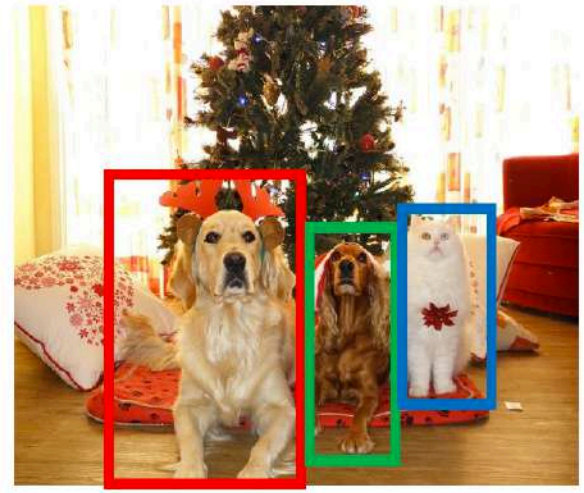
GRASS, CAT, TREE, SKY

No objects, just pixels



CAT

Single Object



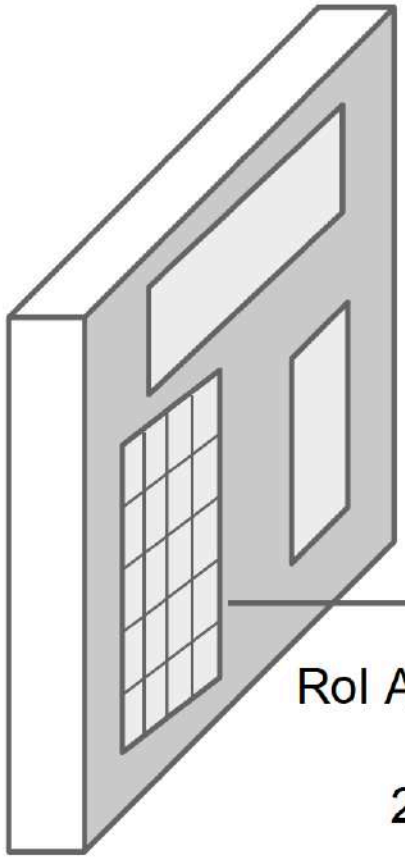
DOG, DOG, CAT

DOG, DOG, CAT

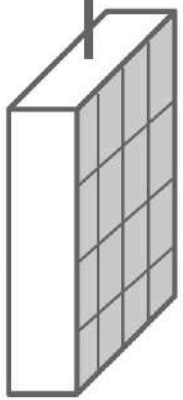
Multiple Object

Mask R-CNN

CNN

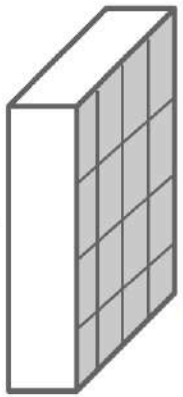


RoI Align



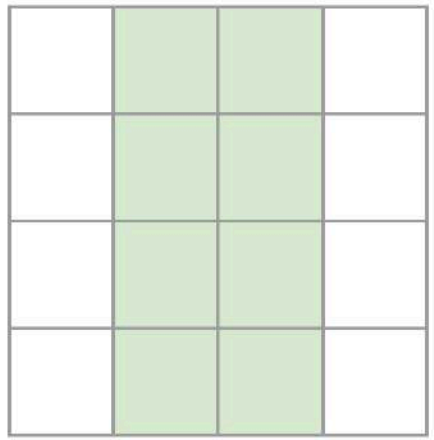
256 x 14 x 14

Conv



256 x 14 x 14

Conv



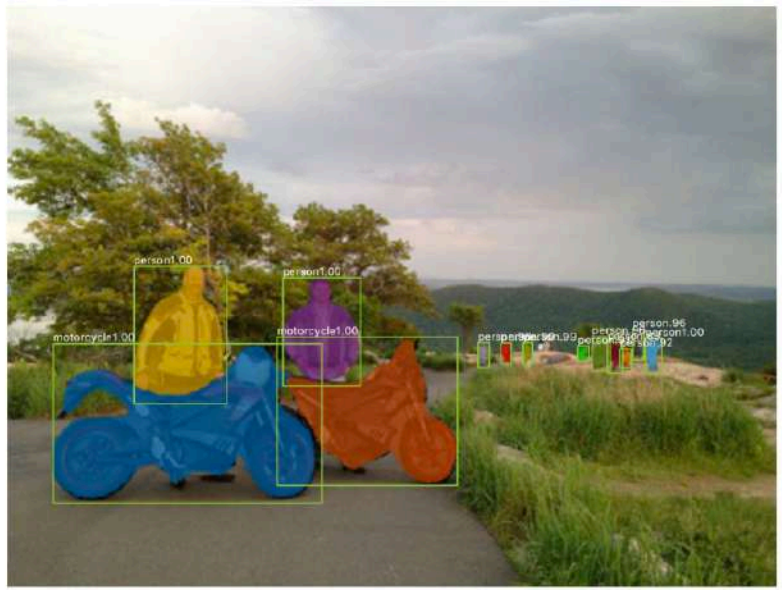
Classification Scores: C
Box coordinates (per class): 4 * C

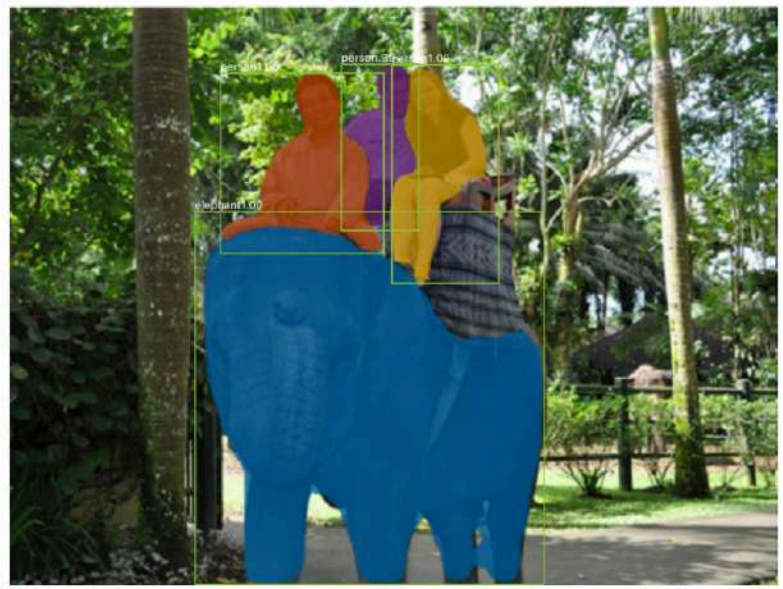
Predict a mask for each of C classes

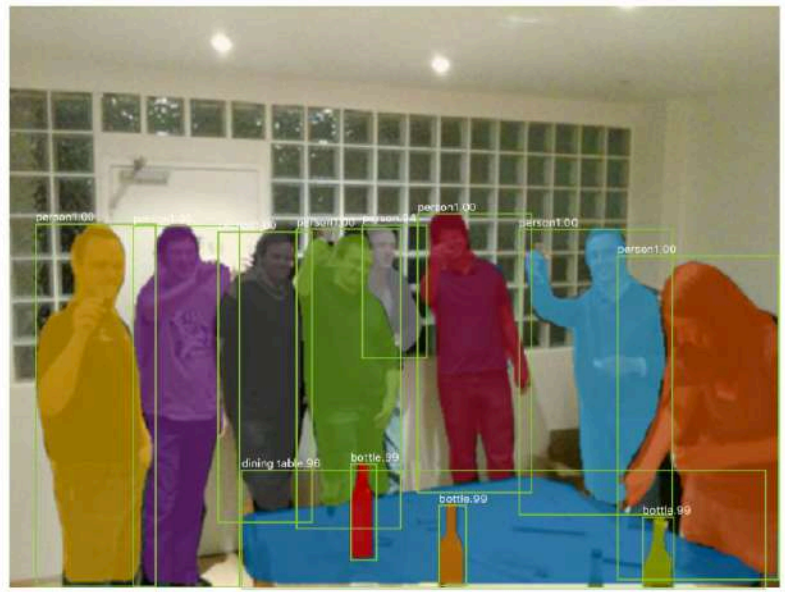
C x 14 x 14

He et al, “Mask R-CNN”, arXiv 2017

Mask R-CNN: Very Good Results!



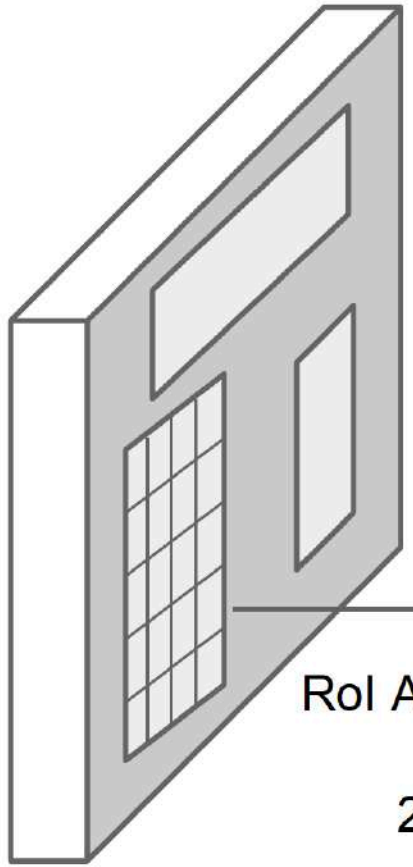




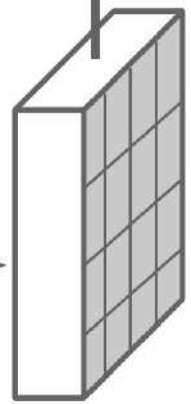
He et al, “Mask R-CNN”, arXiv 2017

Mask R-CNN
Also does pose

CNN

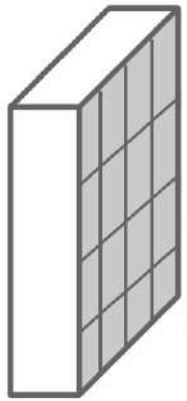


RoI Align



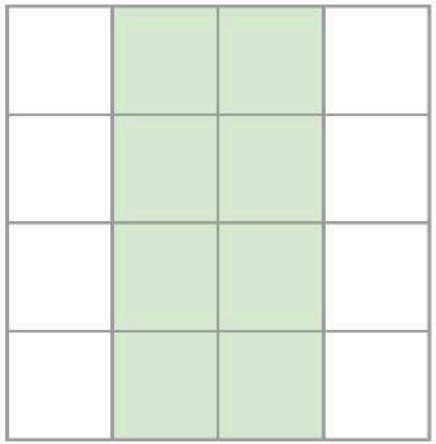
256 x 14 x 14

Conv



256 x 14 x 14

Conv



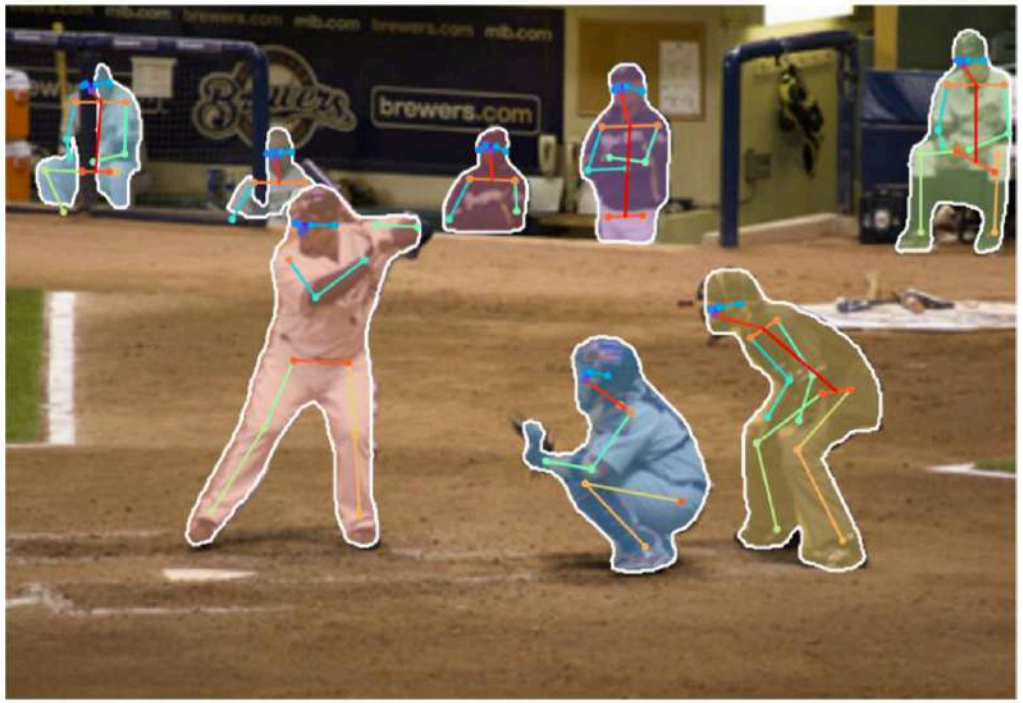
Classification Scores: C
Box coordinates (per class): 4 * C
Joint coordinates

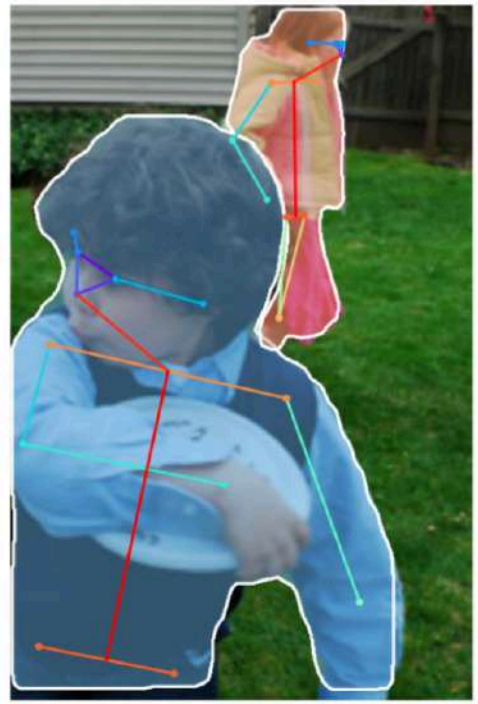
Predict a mask for each of C classes

C x 14 x 14

He et al, “Mask R-CNN”, arXiv 2017

Mask R-CNN
Also does pose





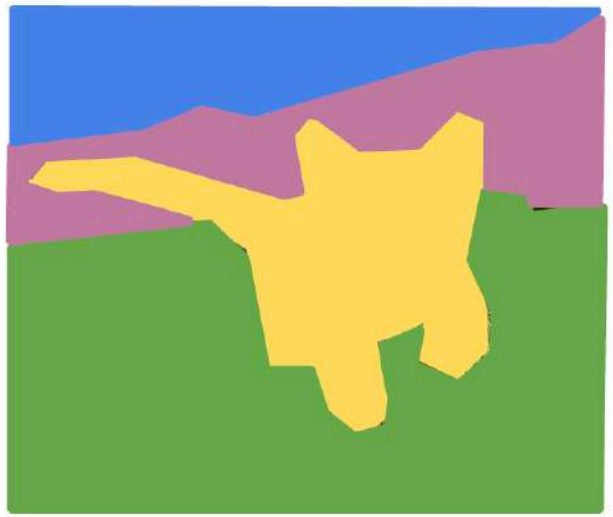
He et al, “Mask R-CNN”, arXiv 2017

R-CNN Family of Networks

- R-CNN used selective search to select proposal in a image. Each proposal is sent through the deep learning model and a 2048 vector is extracted.
- Fast R-CNN applied selective search on feature map instead of image, thus eliminating the need of sending each proposal through the entire network.
- Faster R-CNN removed selective search and used a deep convolution network called RPN (region proposal network) to generate proposals thus allowing to train a end to end neural network in a single stage.
- Mask R-CNN is an extension of Faster R-CNN with an additional module to generate high quality segmentation masks for each image.

Recap:

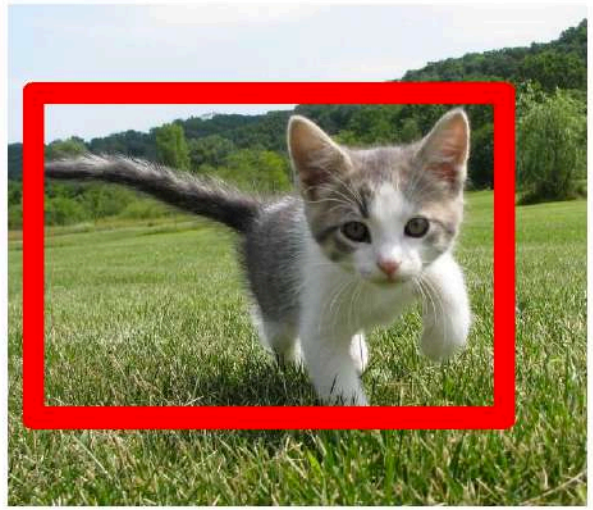
Semantic Segmentation



GRASS, CAT, TREE, SKY

No objects, just pixels

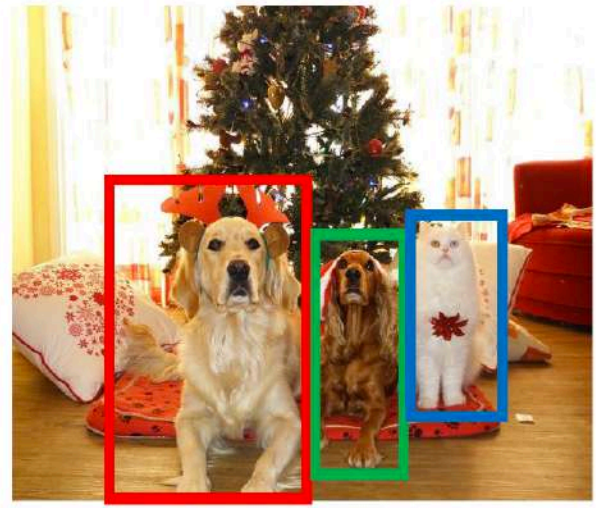
Classification + Localization



CAT

Single Object

Object Detection



DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

Multiple Object