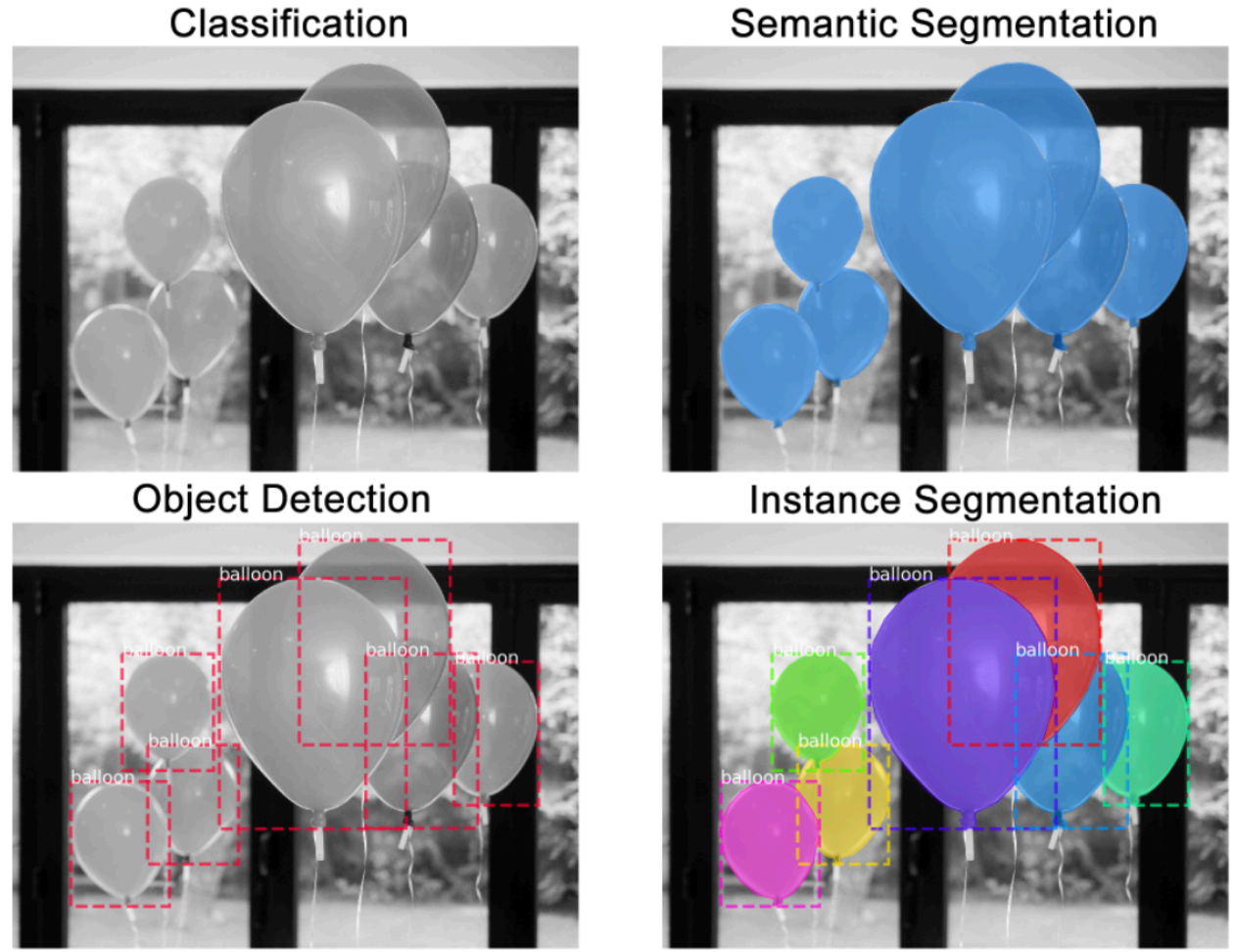
Deep Learning in Computer Vision Applications

Slides credit: Dr. Fatemeh Saleh, Australian National University

Applications of Deep Learning in Computer Vision

- **Classification:** There is a balloon in this image.
- **Semantic Segmentation:** These are all the balloon pixels.
- **Object Detection:** There are 7 balloons in this image at these locations. We're starting to account for objects that overlap.
- **Instance Segmentation**: There are 7 balloons at these locations, and these are the pixels that belong to each one.

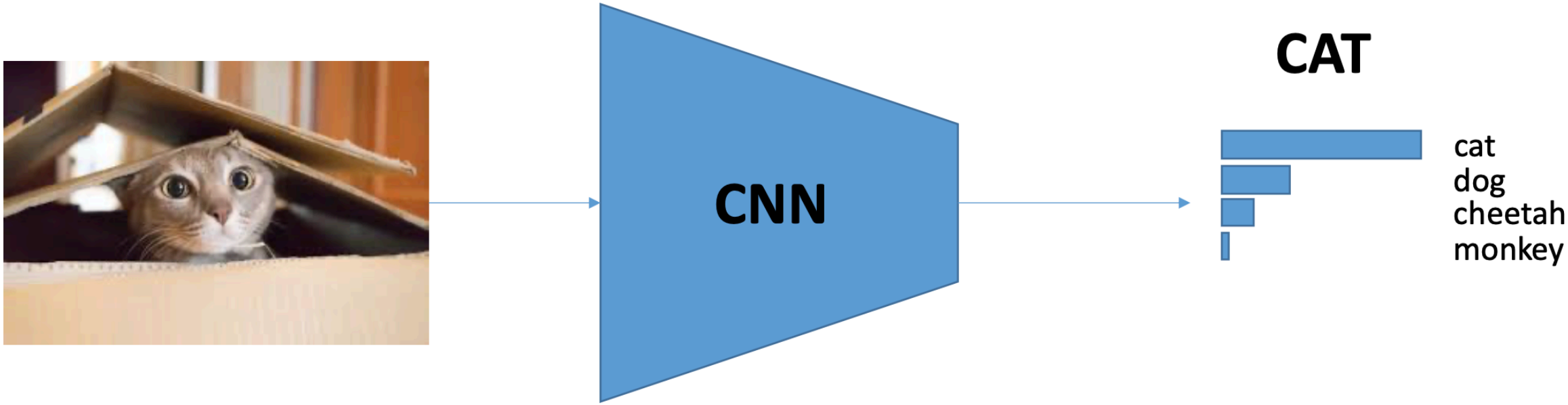


CNNs for Image Classification

- Image Classification Problem
- ImageNet Challenge
- CNN Architectures for Image Classification
  - AlexNet
  - VGG

Image Classification

- Given an image, what is the correct class label?

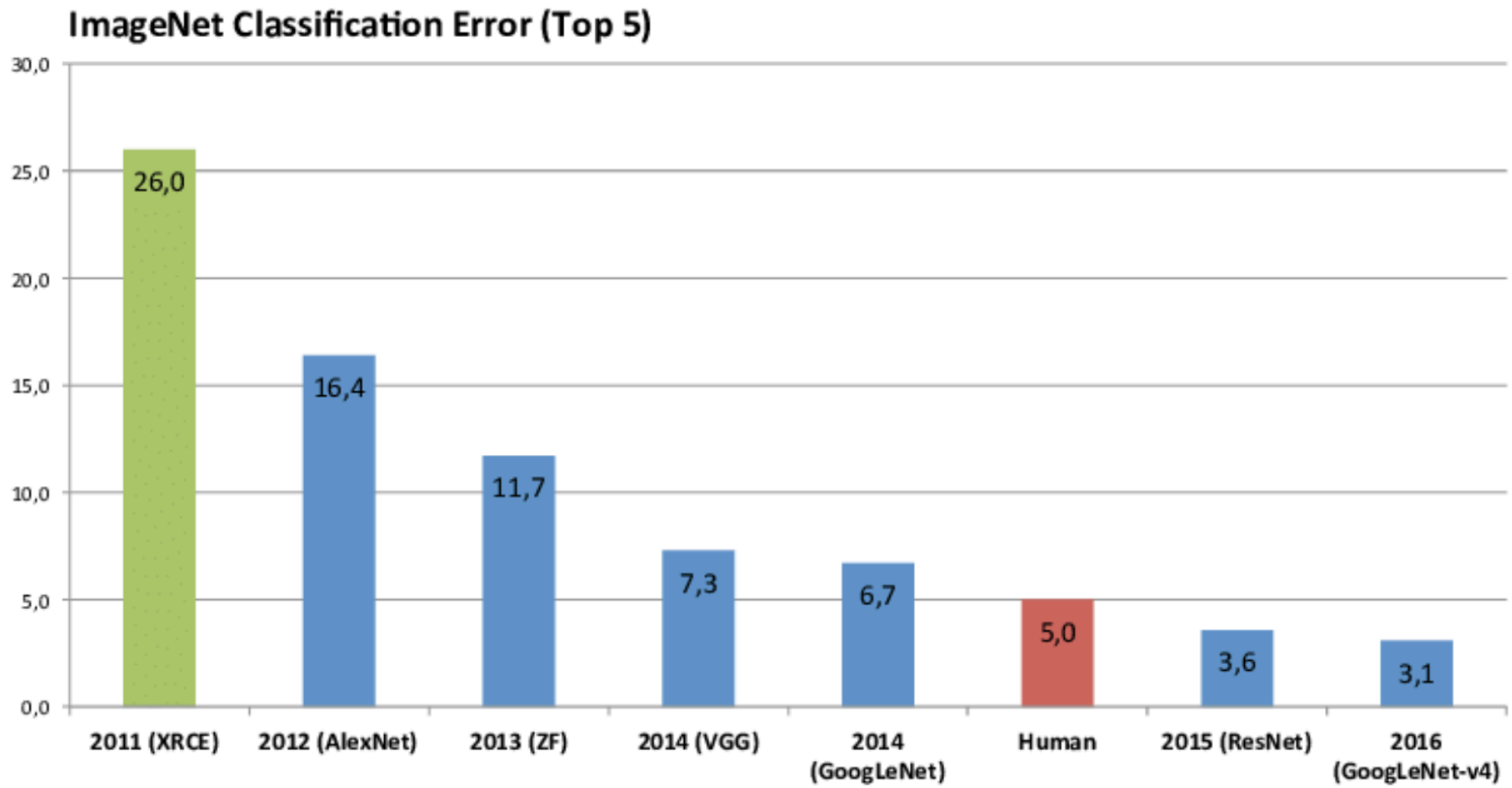


ImageNet Challenge

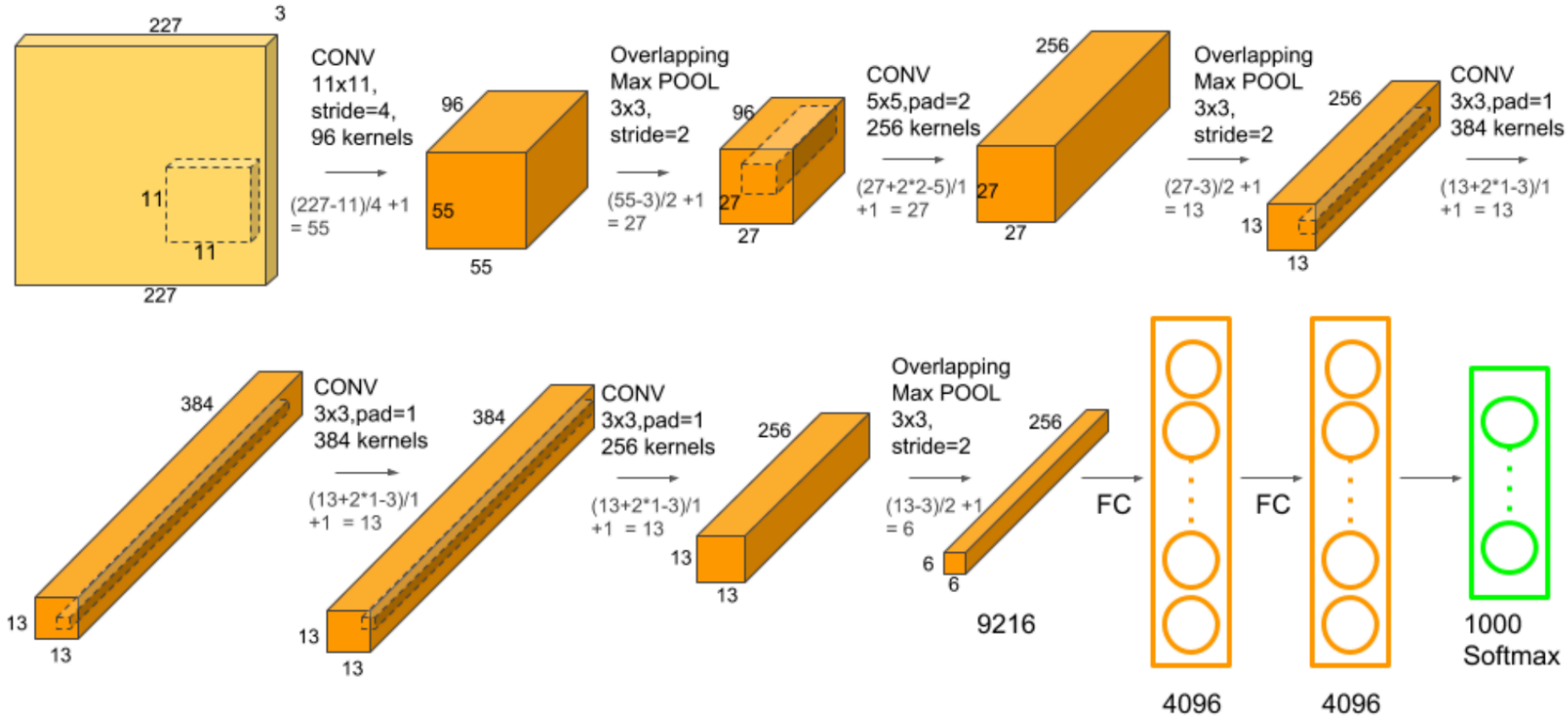
http://www.image-net.org

ImageNet Challenge

- Evaluation Metric: Top-k error



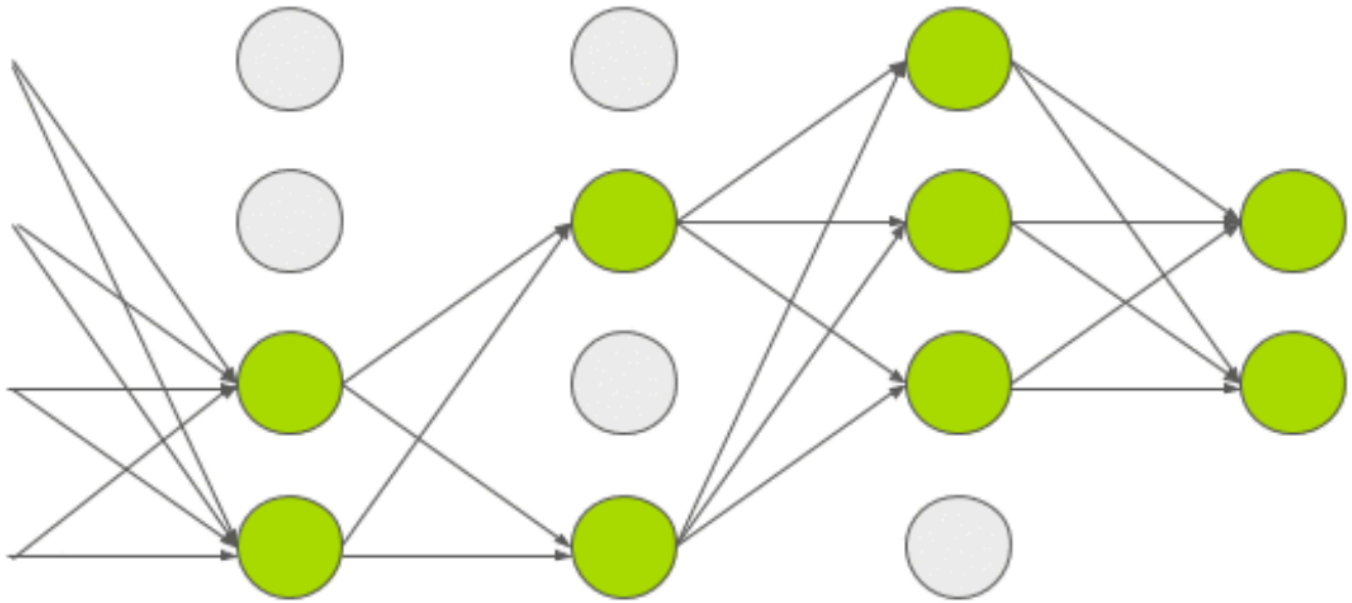
AlexNet



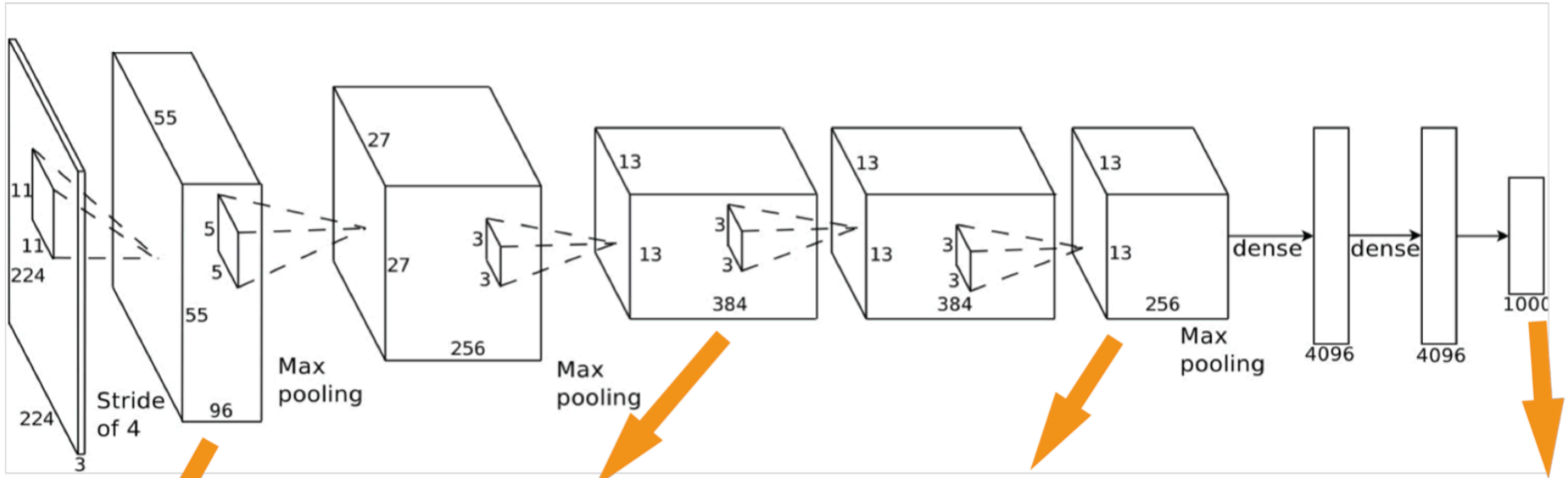
Krizhevsky, Alex et al., "Imagenet classification with deep convolutional neural networks." Advances in neural information processing systems. 2012.

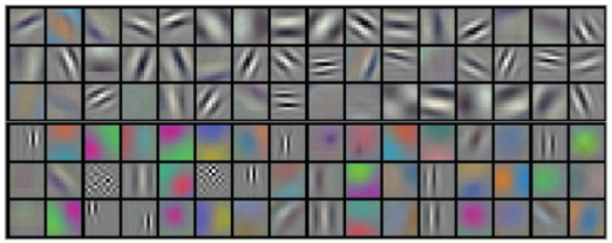
AlexNet

- Dropout Layer
  - Reducing over-fitting by generalization

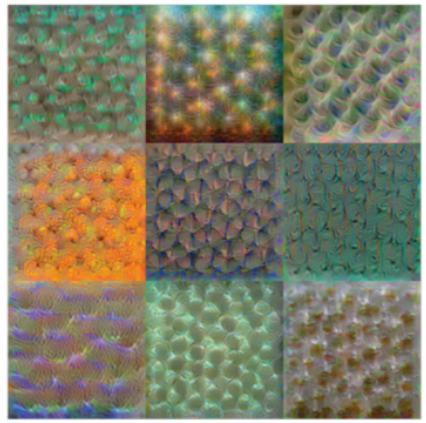


AlexNet





Conv 1: Edge+Blob

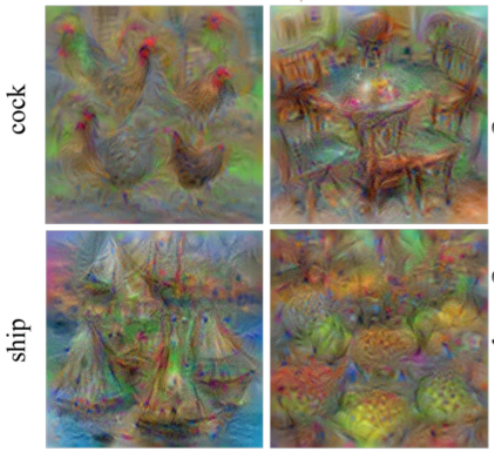


Conv 3: Texture

Numerical

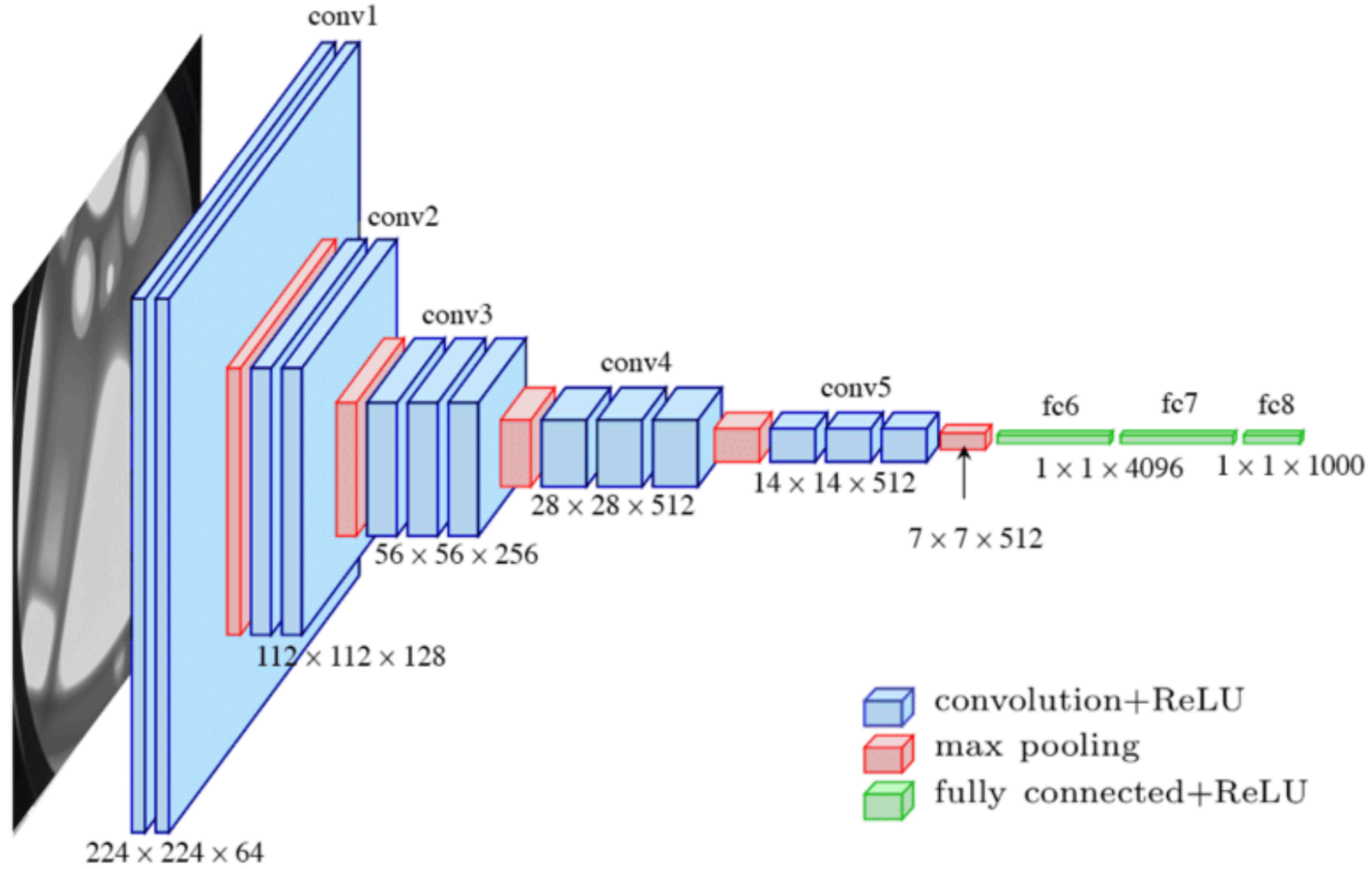
Data-driven

Conv 5: Object Parts



Fc8: Object Classes

VGG



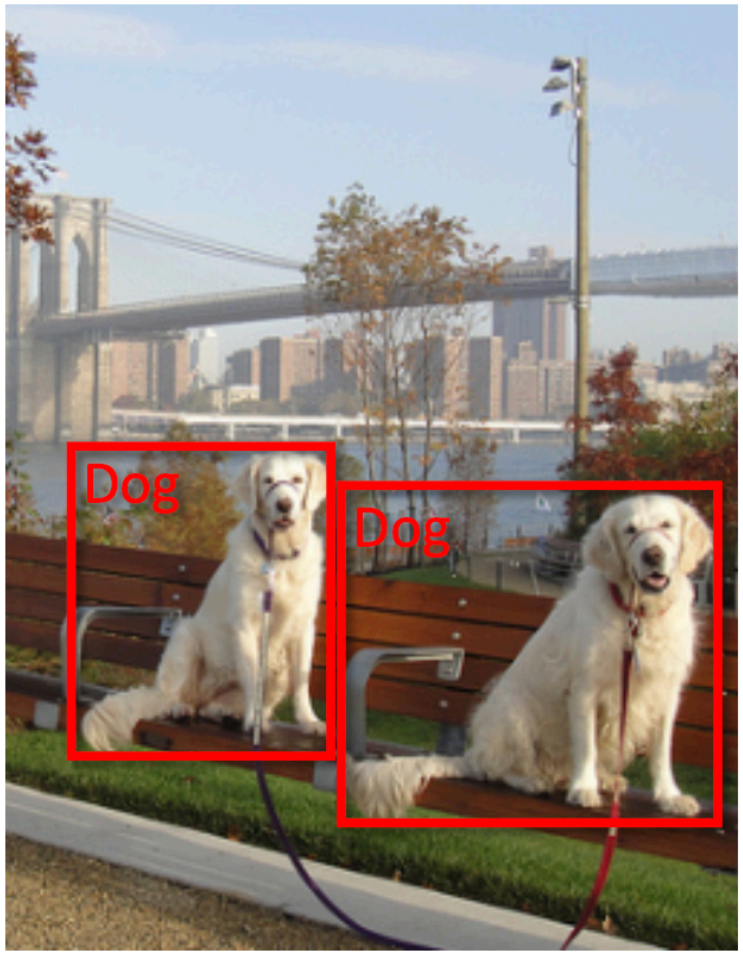
Simonyan, Karen, and Andrew Zisserman. " deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556(2014)

CNNs for Object Detection

- Object Detection Problem
- Evaluation Metrics
- CNN Architectures for Object Detection (two-stage methods)
  - RCNN
  - Fast RCNN
  - Faster RCNN
- Instance level semantic segmentation

Object Detection

- Classification vs. Detection



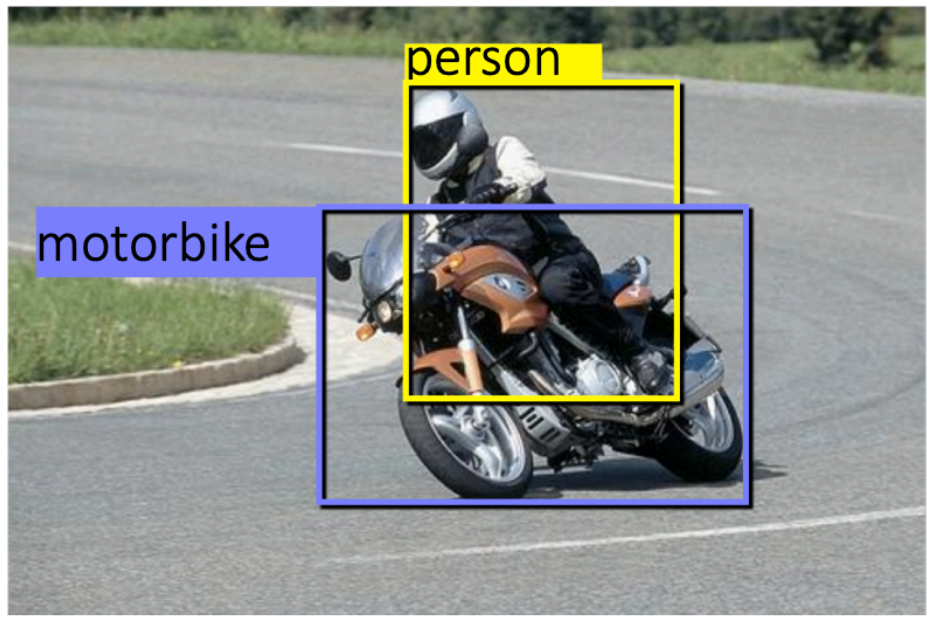
Slide credit: Ross Girshick

Object Detection

- Problem formulation

{ airplane, bird, motorbike, person, sofa }

Input



Desired output

Slide credit: Ross Girshick

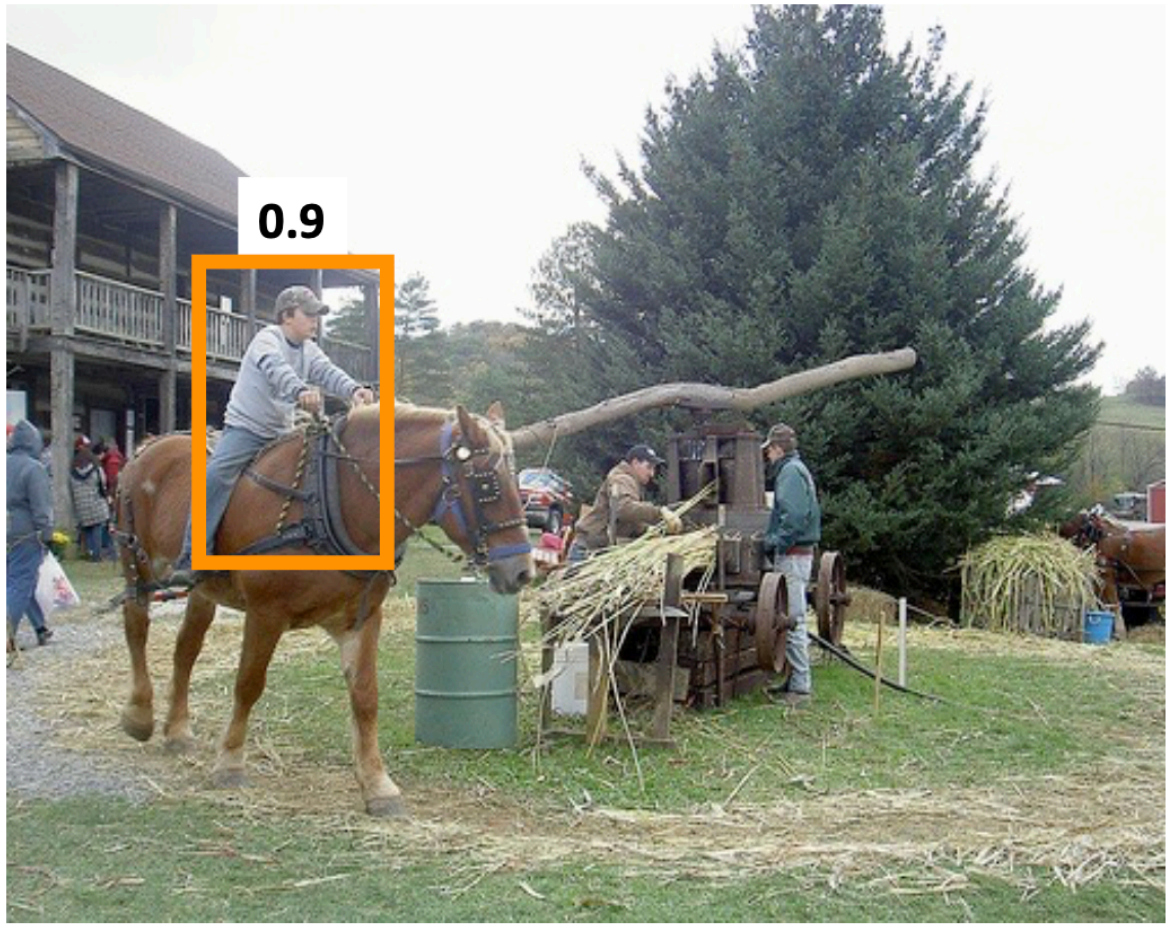
Evaluation Metrics

- Test image (previously unseen)

Evaluation Metrics

- First detection ...

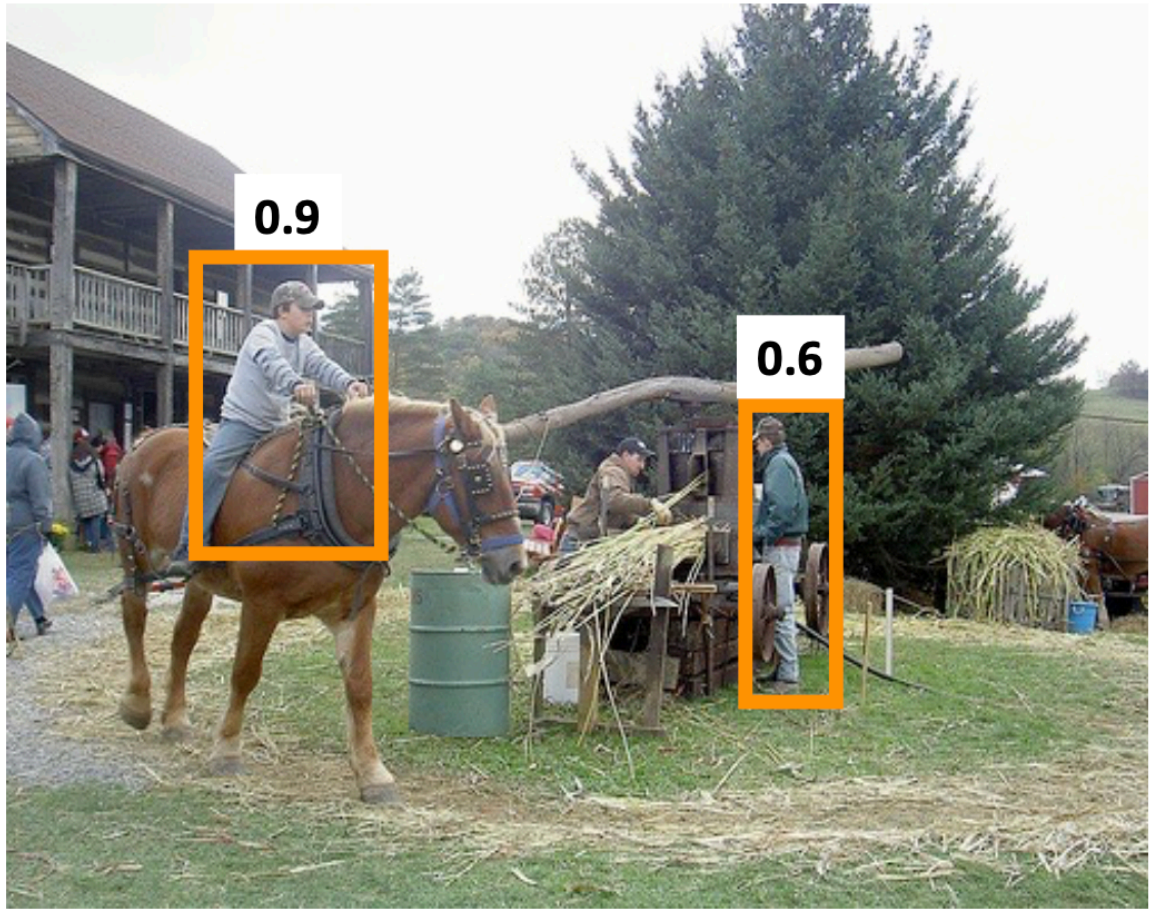
- ‘person’ detector predictions



Evaluation Metrics

- Second detection …

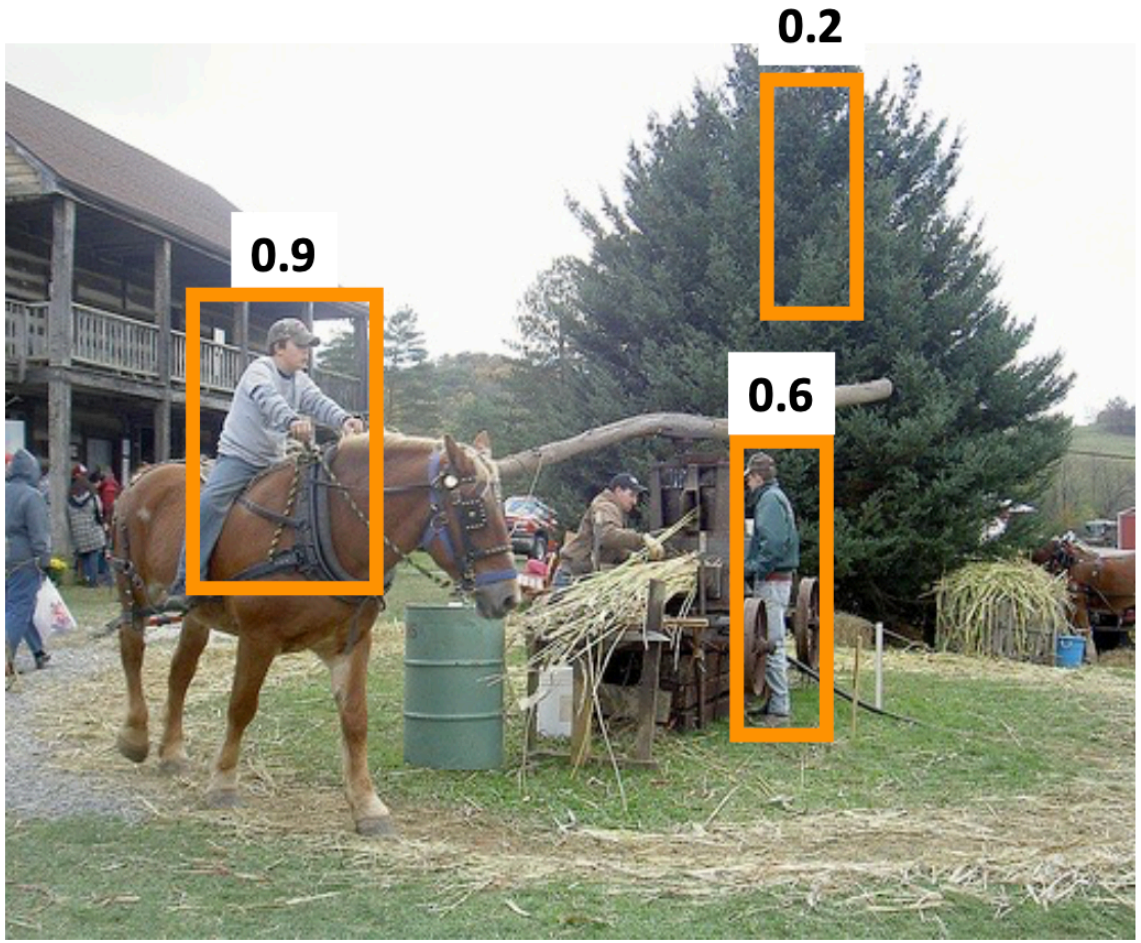
- ‘person’ detector predictions



Evaluation Metrics

- Third detection ...

☐ ‘person’ detector predictions

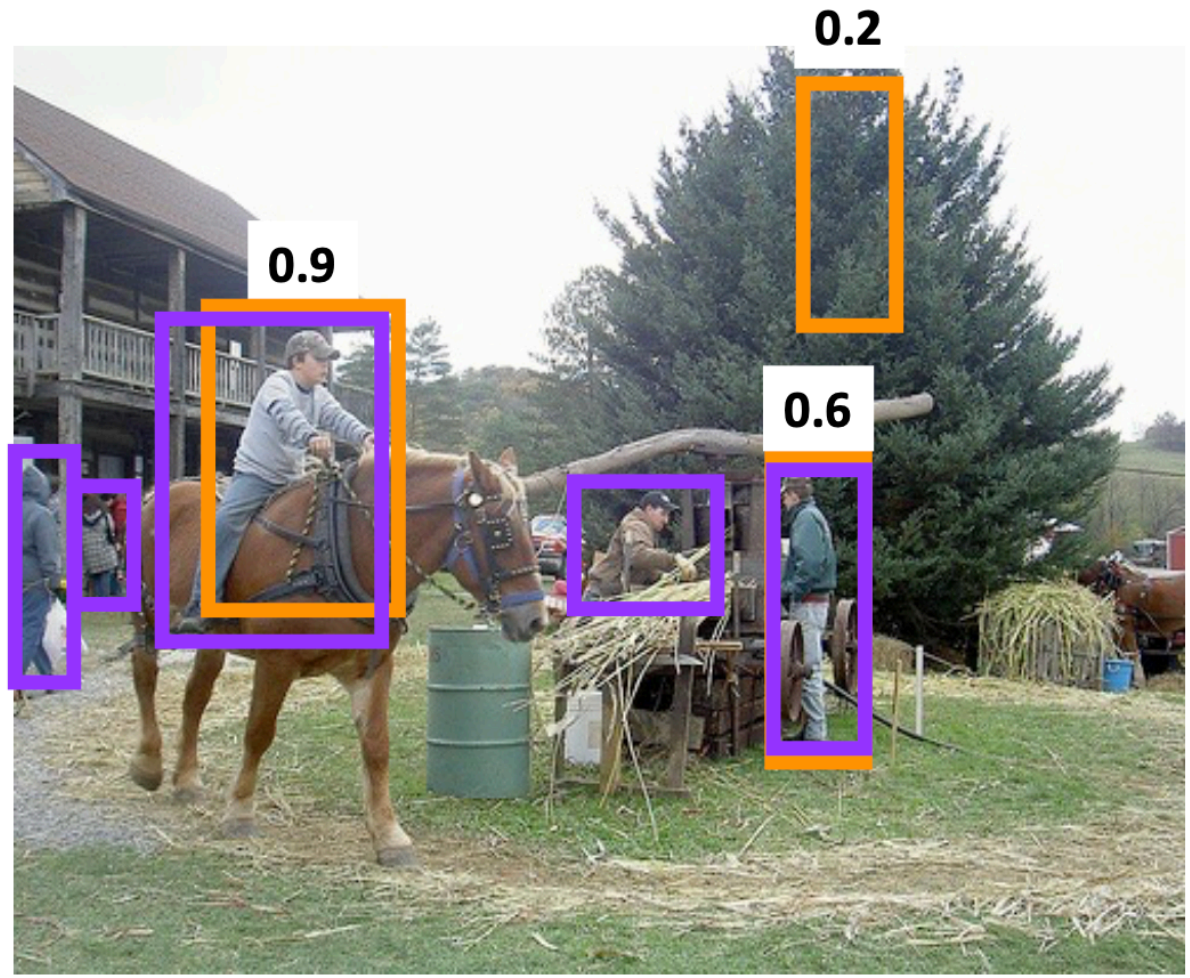


Evaluation Metrics

- Compare to ground-truth

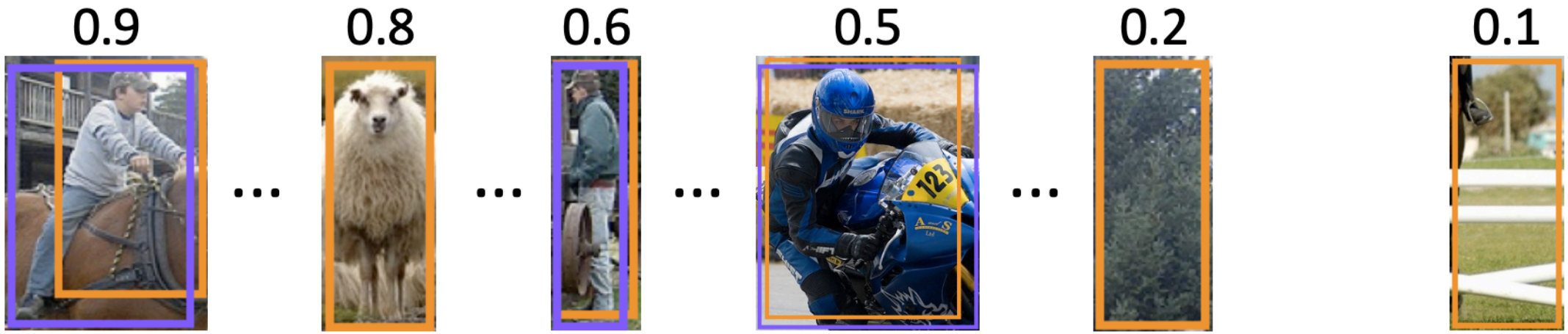
‘person’ detector predictions

Ground-truth ‘person’ boxes



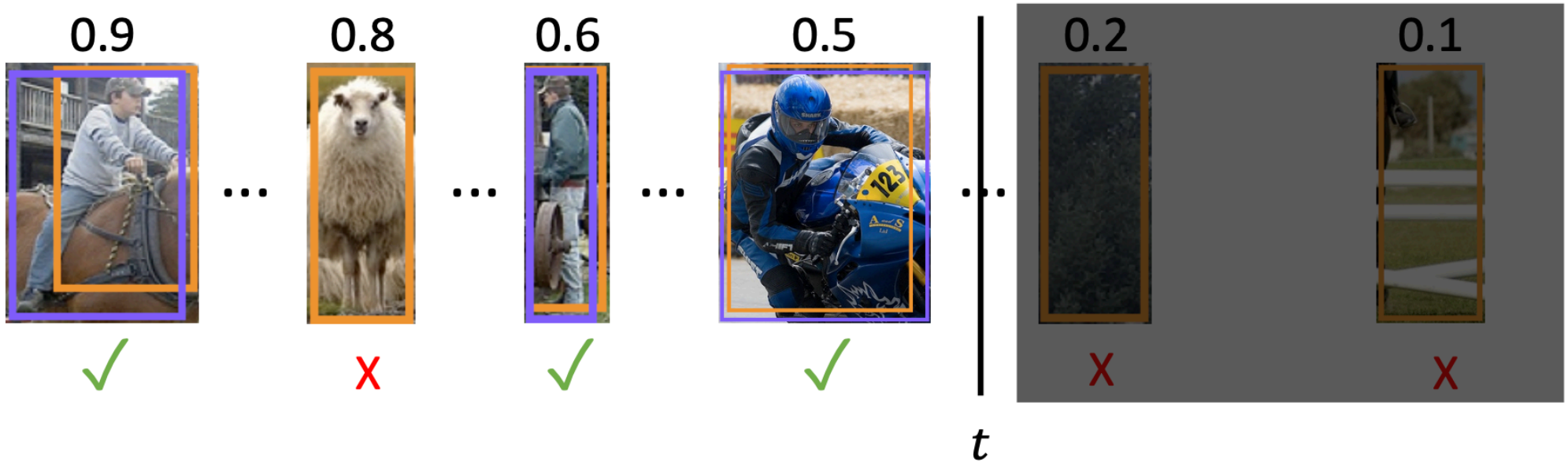
Evaluation Metrics

- Sort by confidence



$$IoU = \frac{area\ of\ overlap}{area\ of\ union}$$

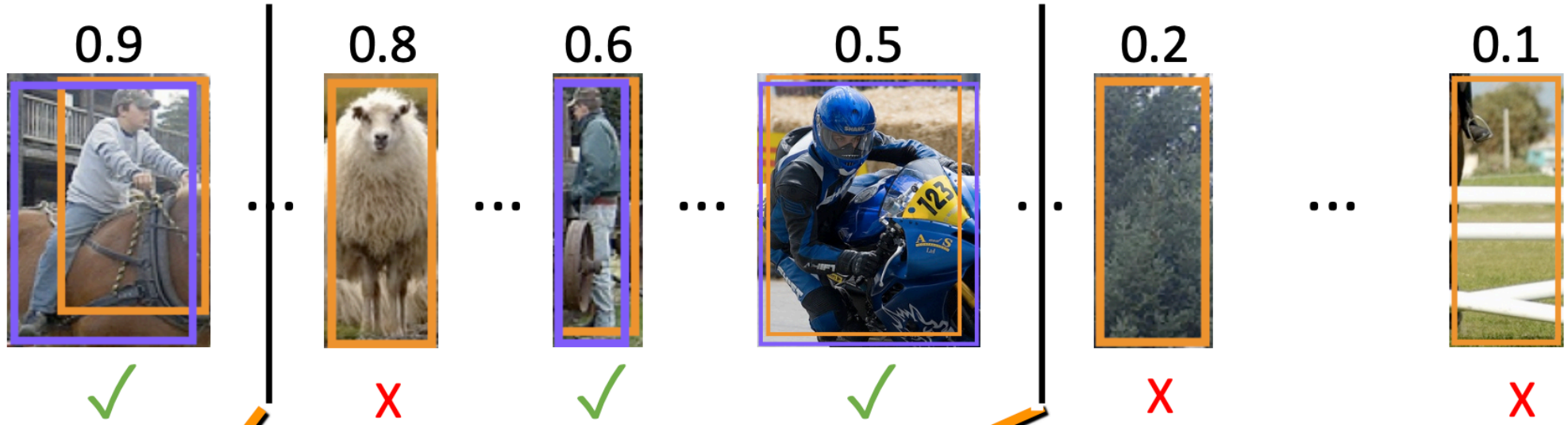
Evaluation Metrics

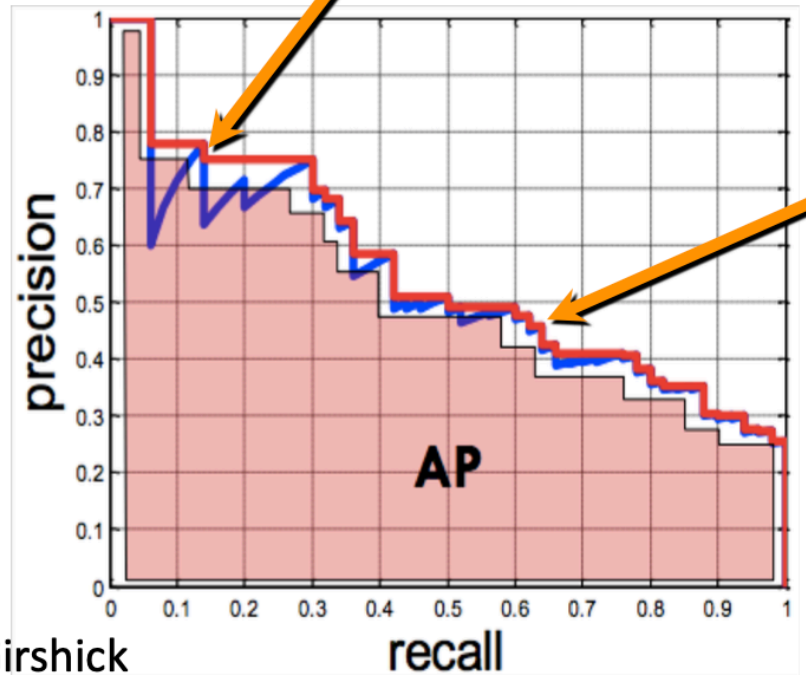


$$precision@t = \frac{\#true\ positives@t}{\#true\ positives@t + \#false\ positives@t}$$

$$recall@t = \frac{\#true\ positives@t}{\#ground\ truth\ objects}$$

Evaluation Metrics





Average Precision (AP)

0% is worst

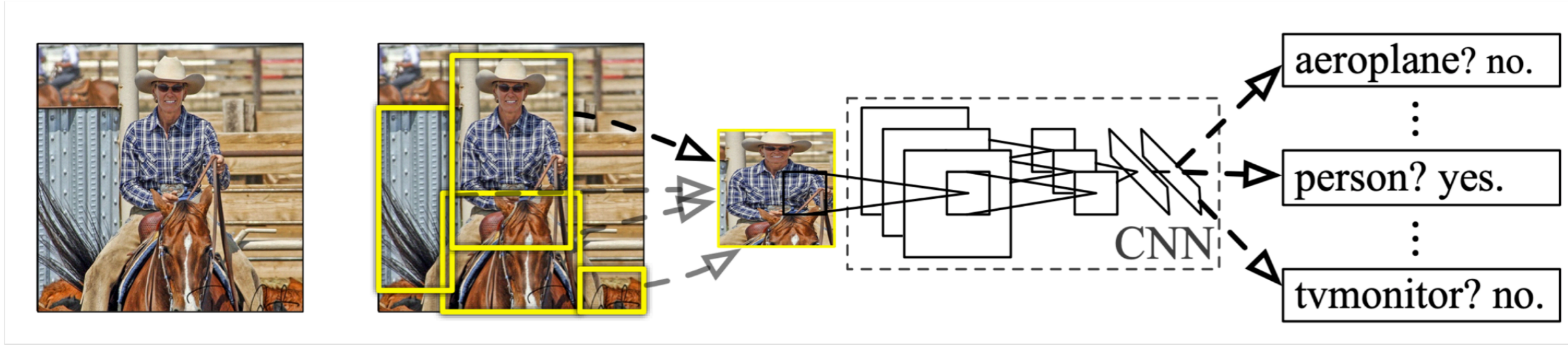
100% is best

mean AP over classes (mAP)

Slide credit: Ross Girshick

RCNN

- Regions with CNN features



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

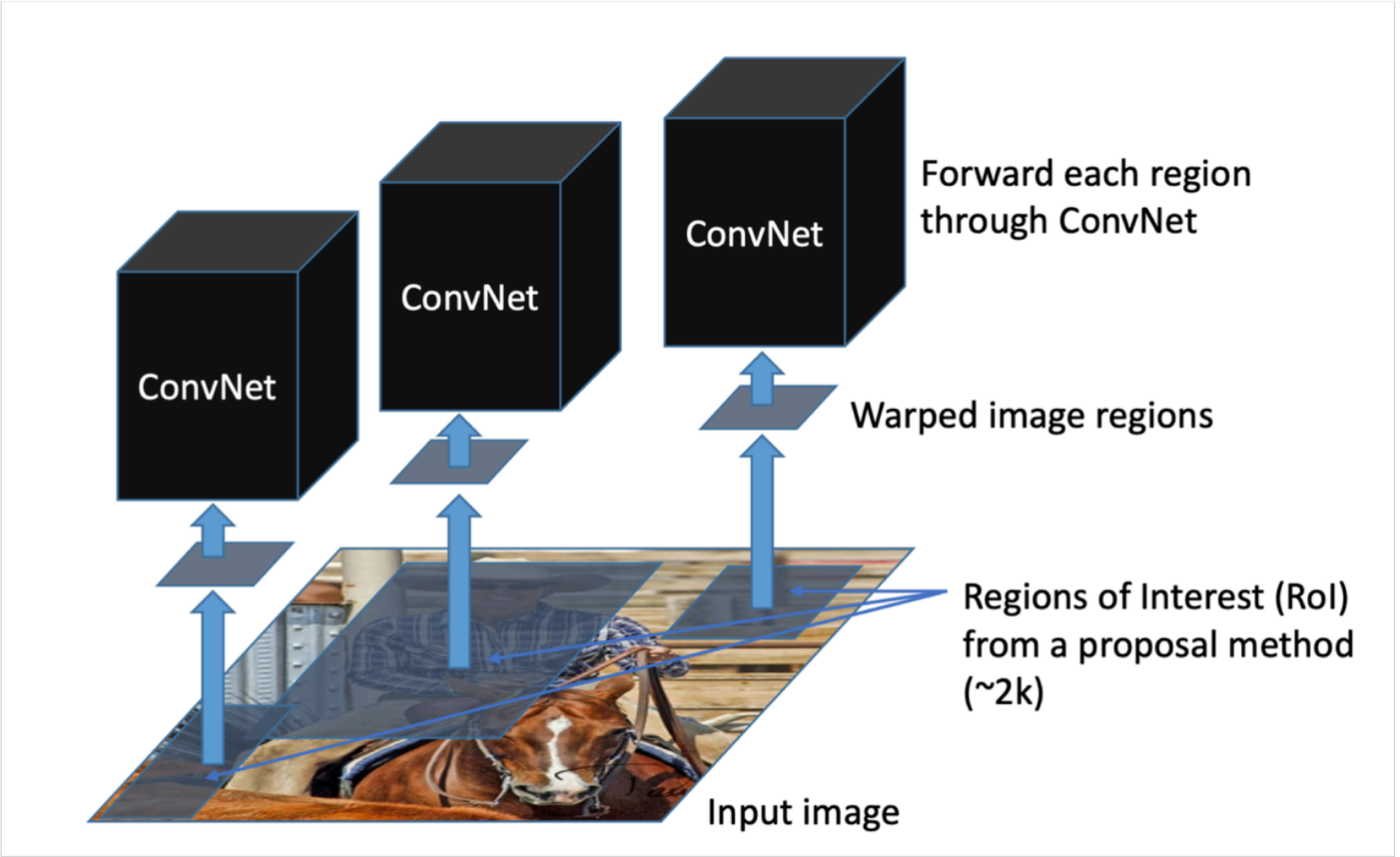
RCNN

Slide credit: Ross Girshick

RCNN

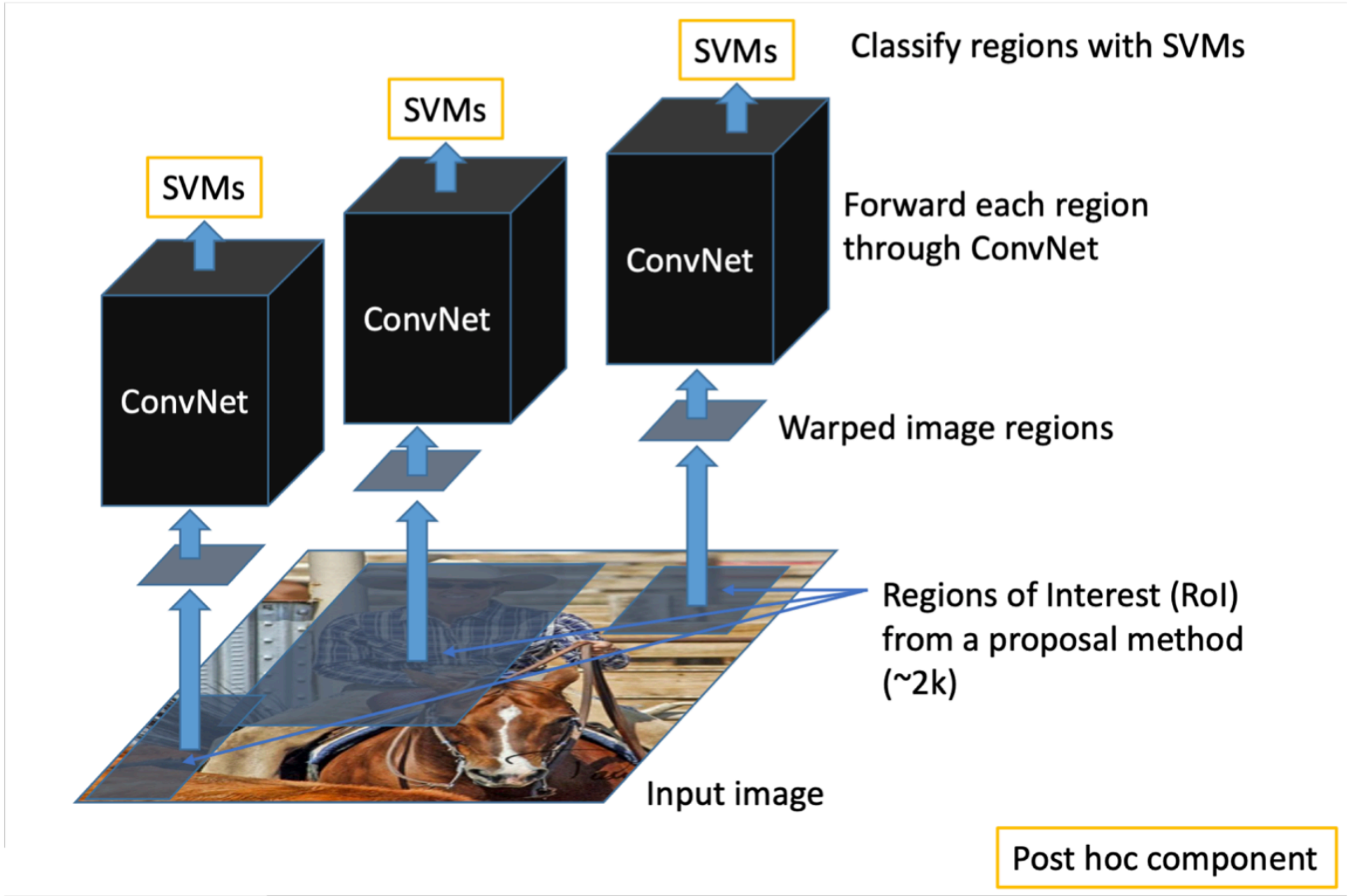
Slide credit: Ross Girshick

RCNN



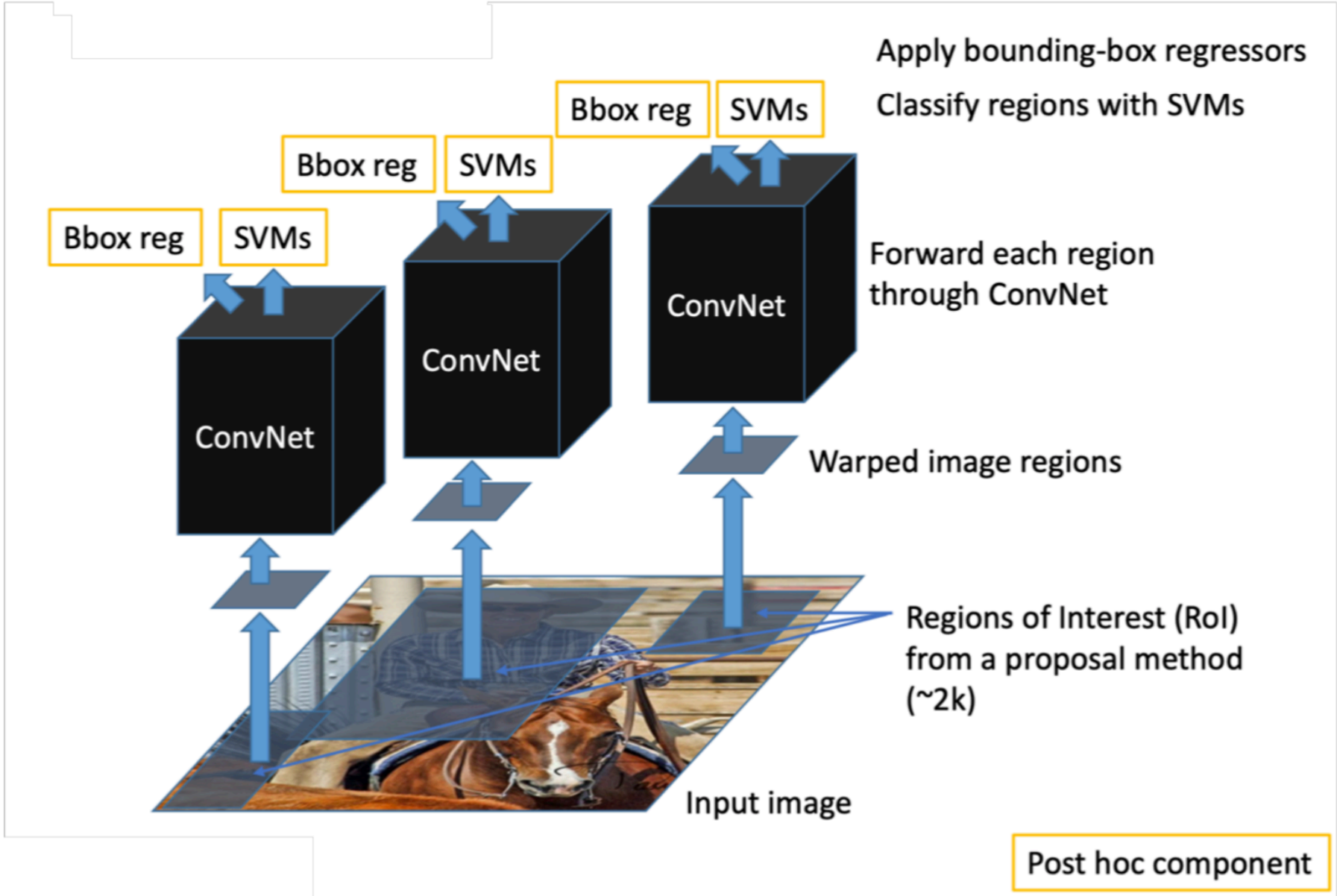
Slide credit: Ross Girshick

RCNN



Slide credit: Ross Girshick

RCNN



Slide credit: Ross Girshick

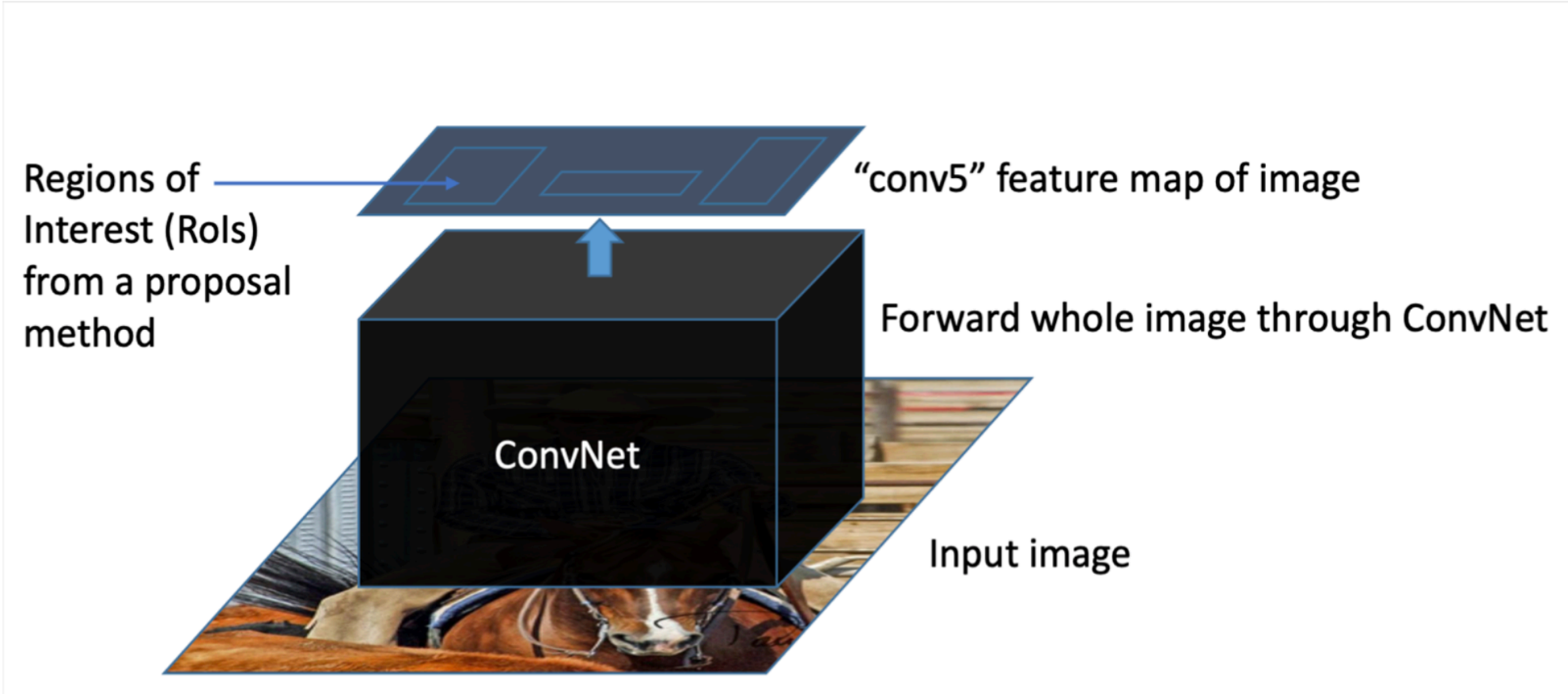
RCNN

- Problems
  - Ad hoc training objectives
    - Fine-tune network with softmax classifier
    - Train linear SVMs
    - Train bounding-box regressors
  - Training is slow (84h), takes a lot of disk space
  - Inference (detection) is slow, 47s/image with VGG16

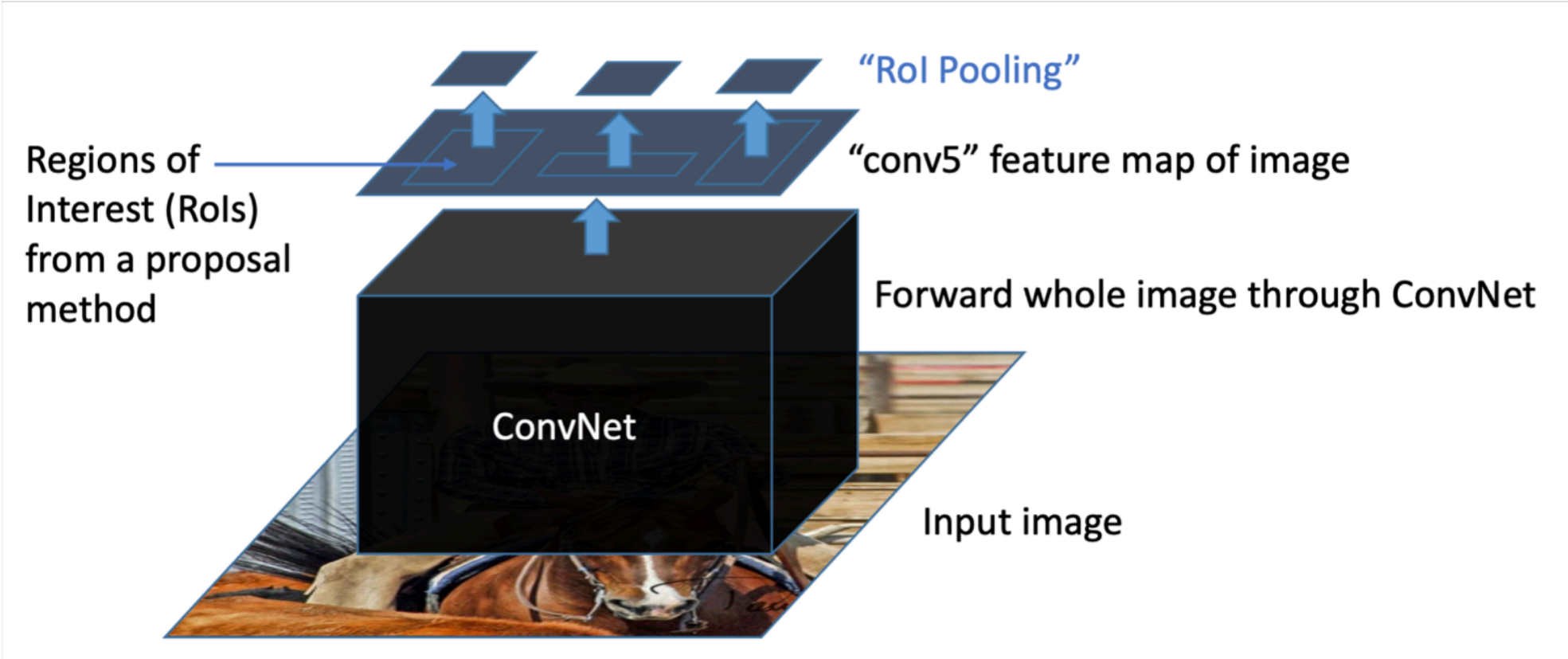
Fast RCNN

- Fast test-time
- One network, trained in one stage
- Higher mean average precision than R-CNN

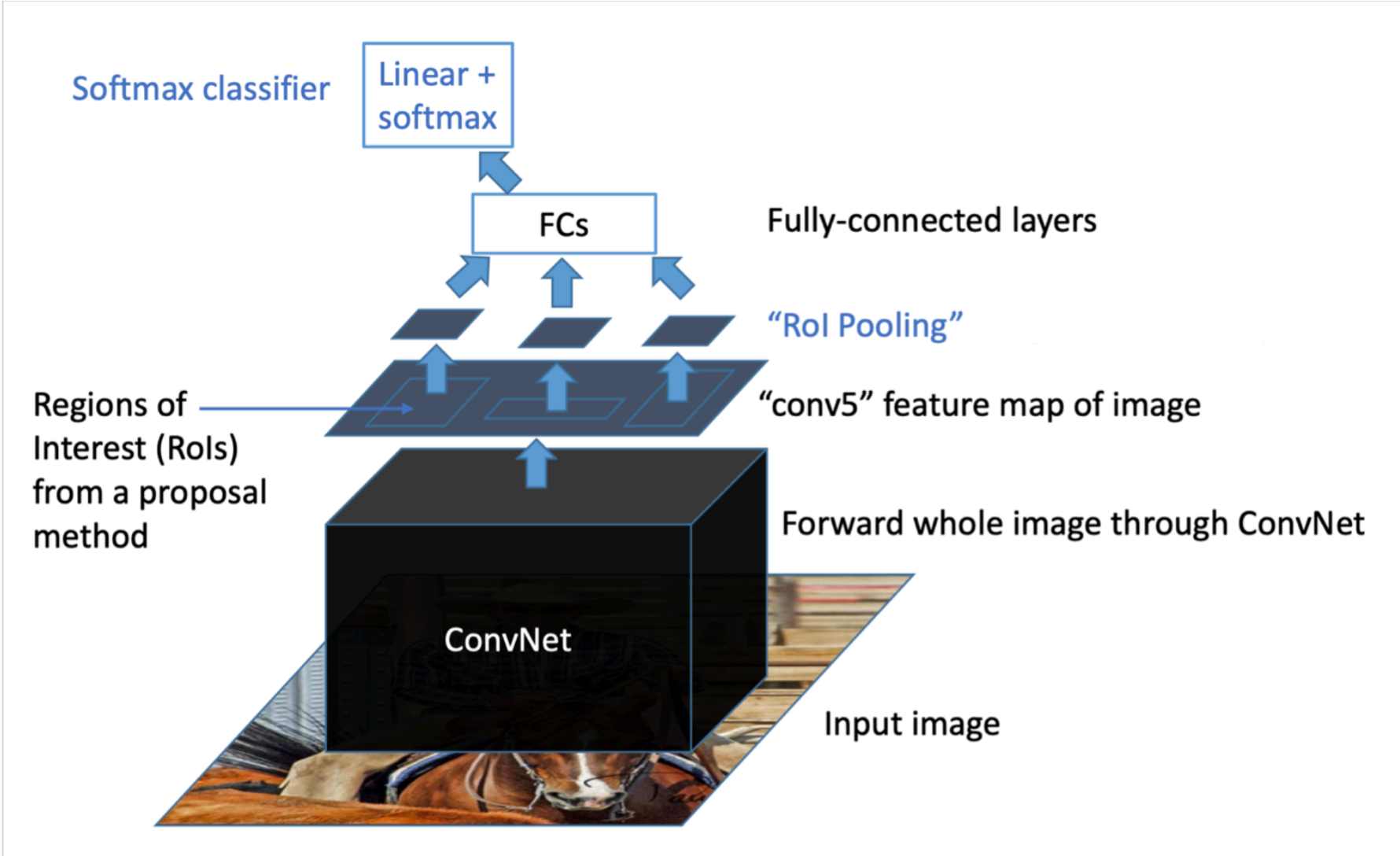
Fast RCNN (test time)



Fast RCNN (test time)

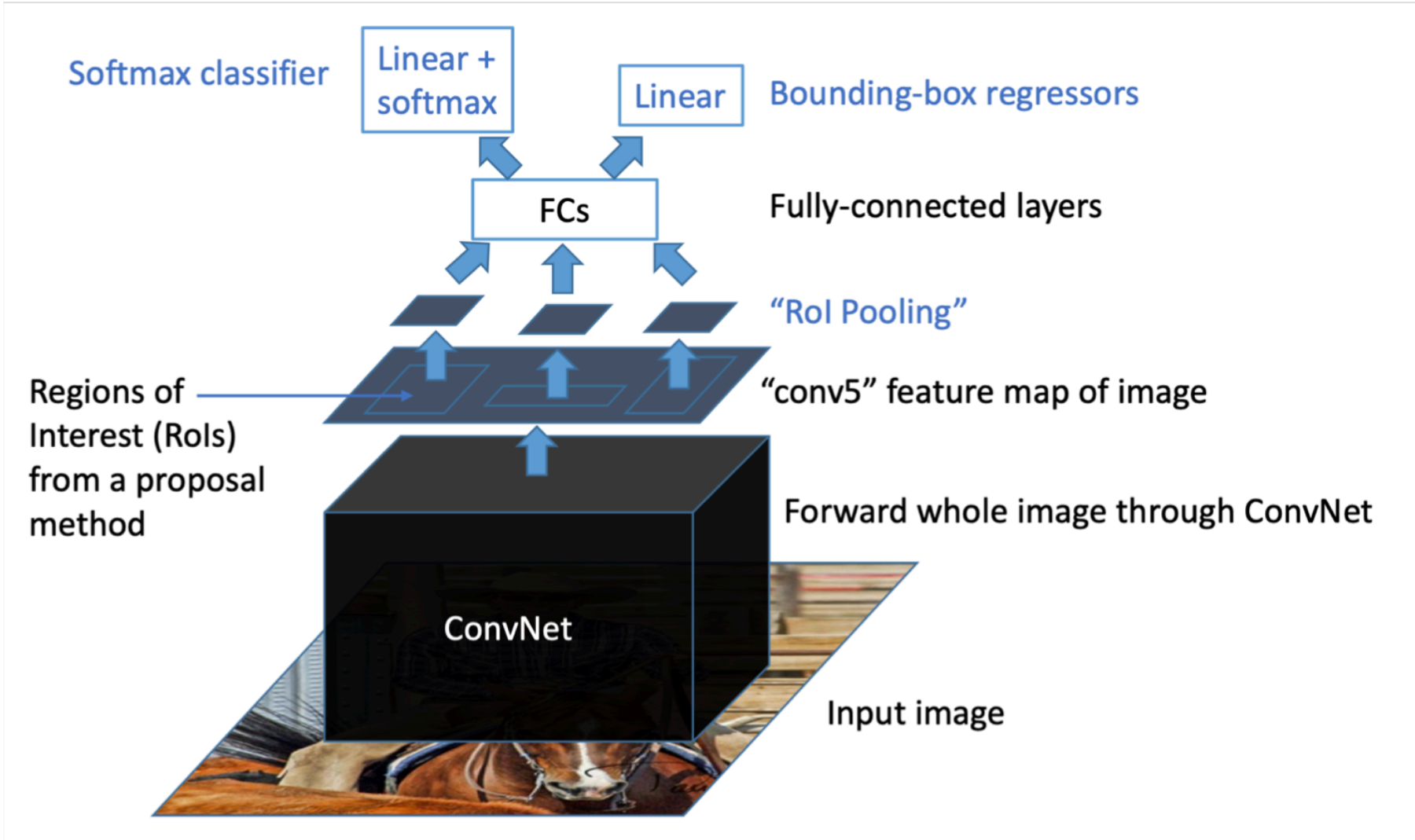


Fast RCNN (test time)



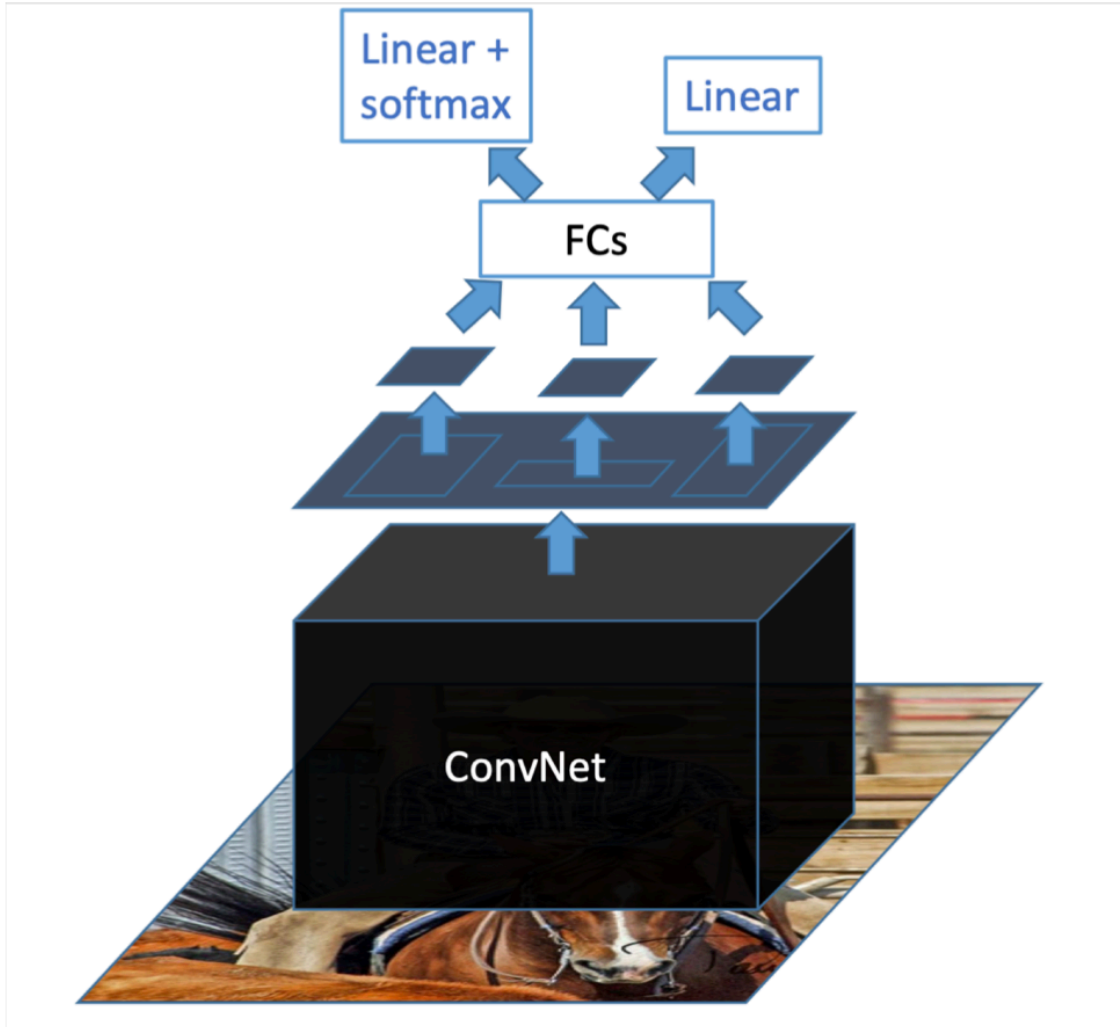
Slide credit: Ross Girshick

Fast RCNN (test time)



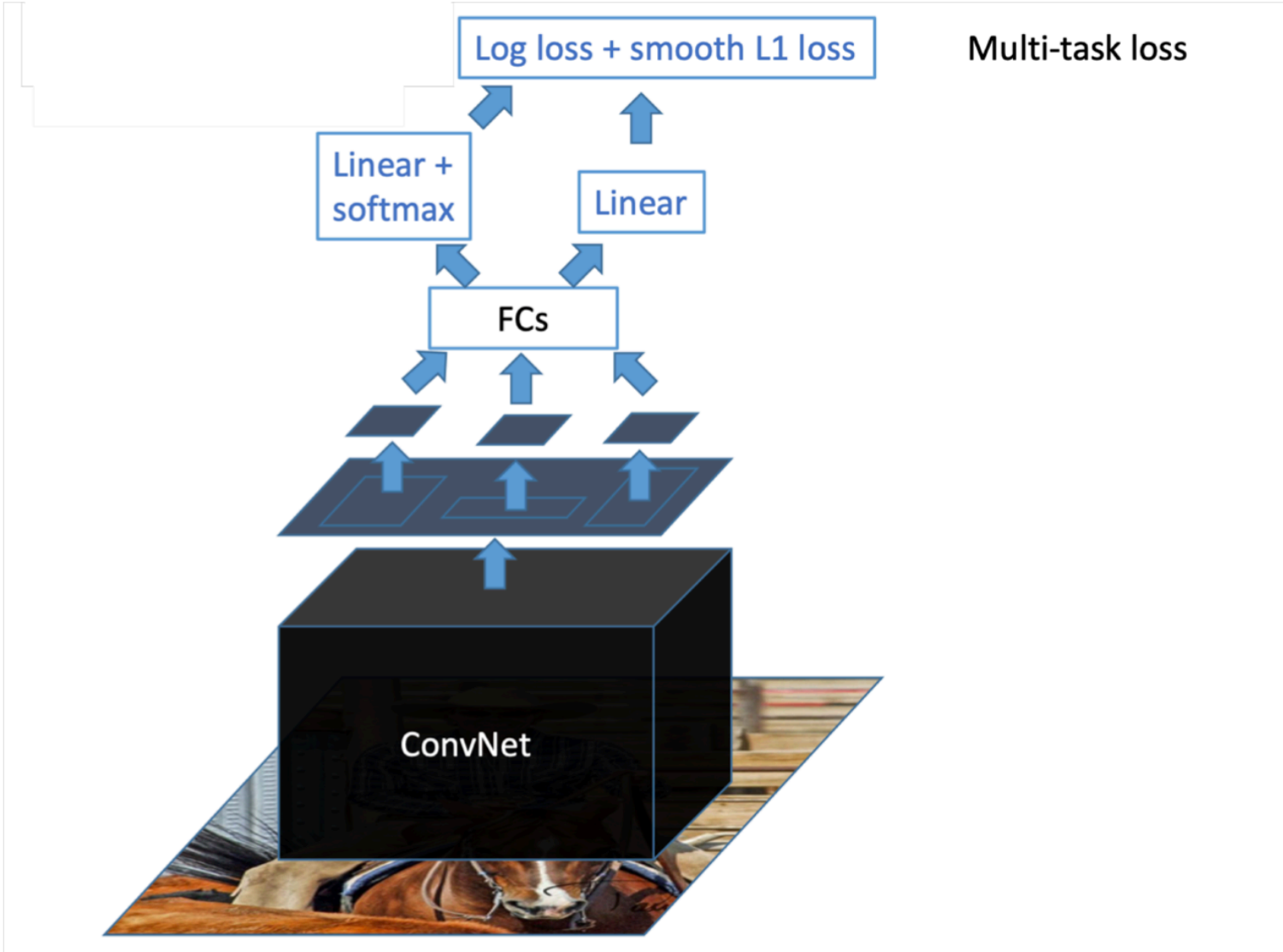
Slide credit: Ross Girshick

Fast RCNN (training)



Slide credit: Ross Girshick

Fast RCNN (training)



Slide credit: Ross Girshick

Results (RCNN vs Fast RCNN)

	R-CNN	**Fast R-CNN**
Training Time:	84 hours	**9.5 hours**
(Speedup)	1x	**8.8x**
Test time per image	47 seconds	**0.32 seconds**
(Speedup)	1x	**146x**
mAP (VOC 2007)	66.0	**66.9**

Fast RCNN

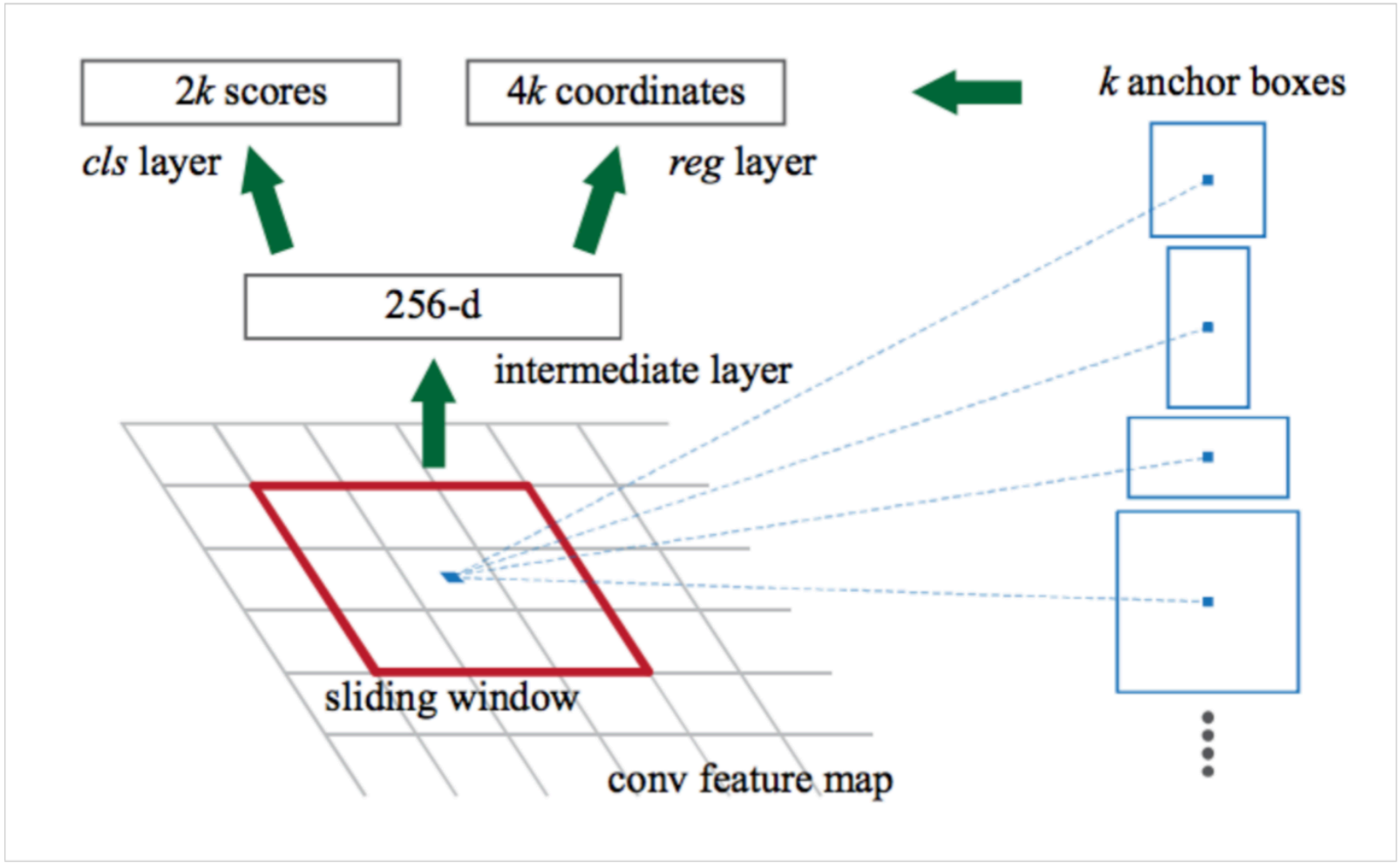
- Problem
  - Test-time speeds do not include region proposals

	R-CNN	Fast R-CNN
Test time per image	47 seconds	**0.32 seconds**
(Speedup)	1x	**146x**
Test time per image with Selective Search	50 seconds	**2 seconds**
(Speedup)	1x	**25x**

- Solution:
  - Just make the CNN do region proposals too!

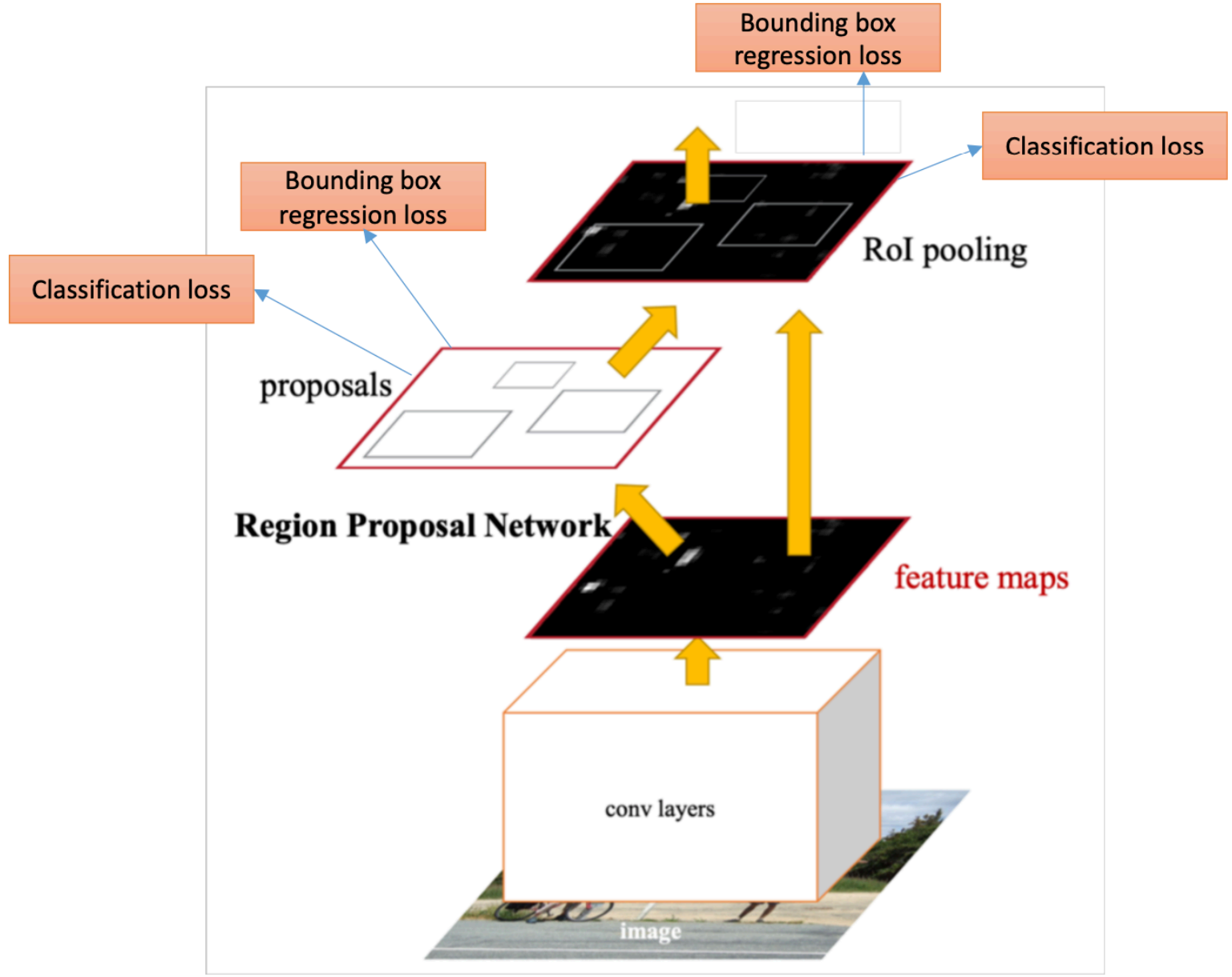
Faster RCNN

- Region Proposal Network (RPN)



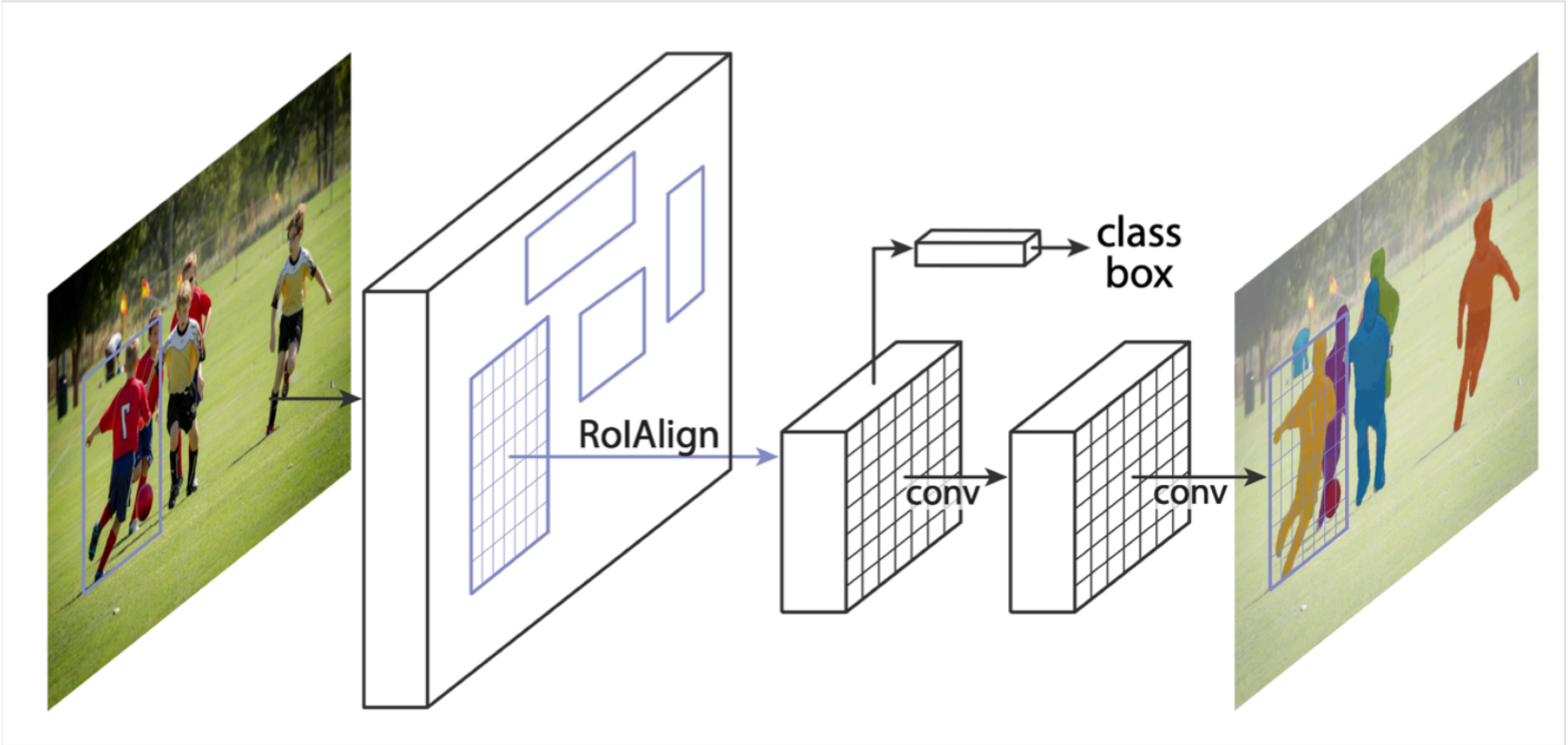
Faster RCNN

- **Training**
- One network, four losses
  - RPN classification
  - RPN regression
  - Fast R-CNN classification
  - Fast R-CNN regression



Object Instance Segmentation: Mask RCNN

- Extending Faster R-CNN for Pixel Level Segmentation

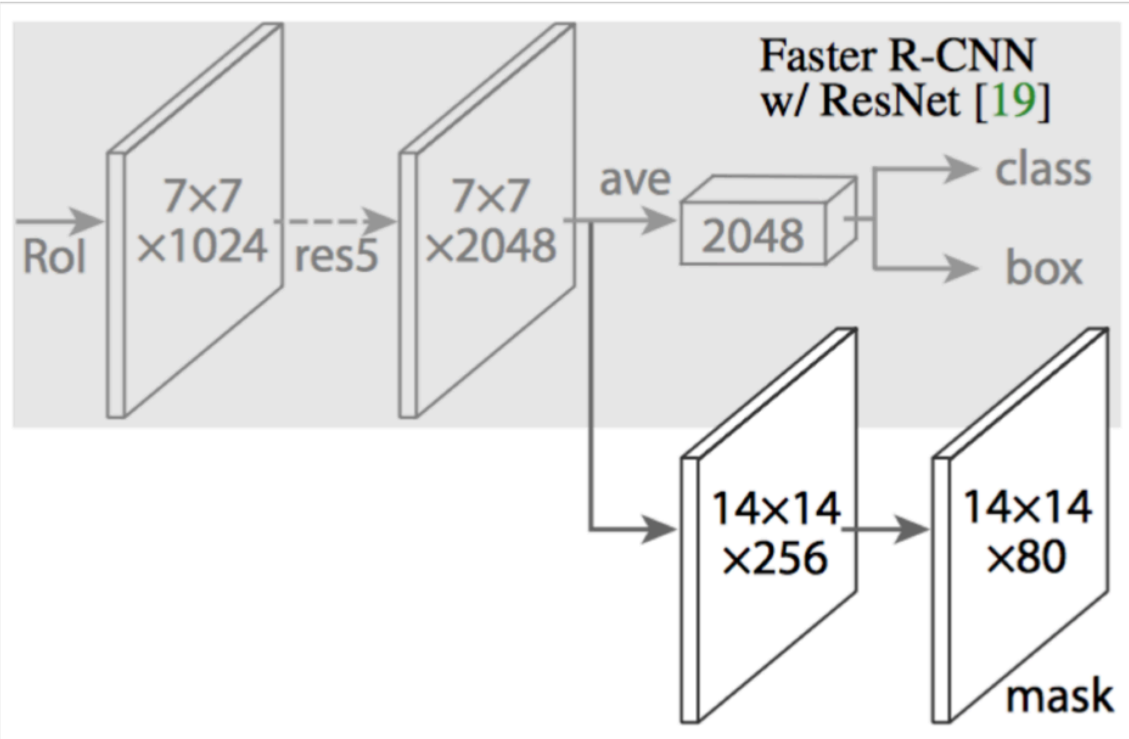


He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

Object Instance Segmentation: Mask RCNN

- Faster R-CNN has two outputs for each candidate object, a **class label** and a **bounding-box offset**
- In the Mask R-CNN they add a third branch that outputs the **object mask**.
- During training, they define a multi-task loss on each sampled RoI:

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



CNNs for Semantic Segmentation

CNNs for Semantic Segmentation

- Semantic Segmentation Problem
- Evaluation Metrics
- CNN Architectures for Semantic Segmentation
  - FCN
  - U-net

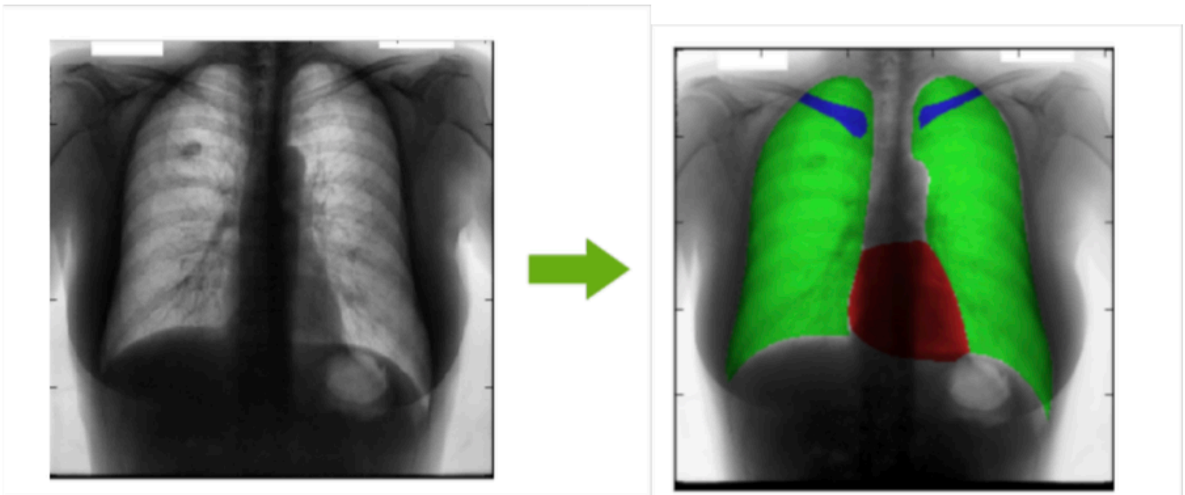
Semantic Segmentation Problem

- Classifying each pixel of an image into a category corresponding to an object or a part of the image (road, sky, …).

0: Background/Unknown
1: Person
2: Purse
3: Plants/Grass
4: Sidewalk
5: Building/Structures

Semantic Segmentation Problem

- Applications:
  - self-driving vehicles
  - human-computer interaction
  - virtual reality
  - bio Medical Image Diagnosis



Input image

Segmented image

Evaluation

- **Per Class Pixel Accuracy**
- The percent of pixels which were correctly classified for each class

$$pixel\ accuracy_i = \frac{the\ number\ of\ pixels\ of\ class\ i\ predicted\ to\ belong\ to\ class\ i}{the\ total\ number\ of\ pixels\ of\ class\ i}$$

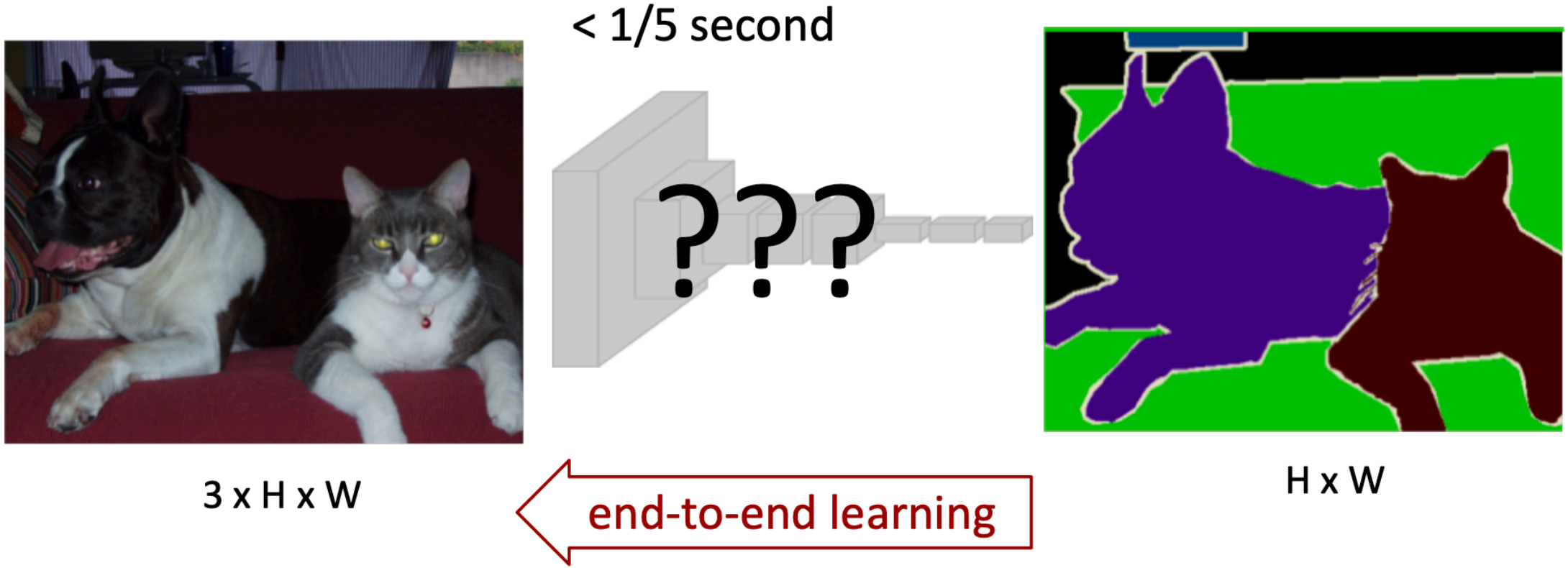
- Suitable for datasets with no background class but has a strong drawback for datasets with a large background class

Evaluation

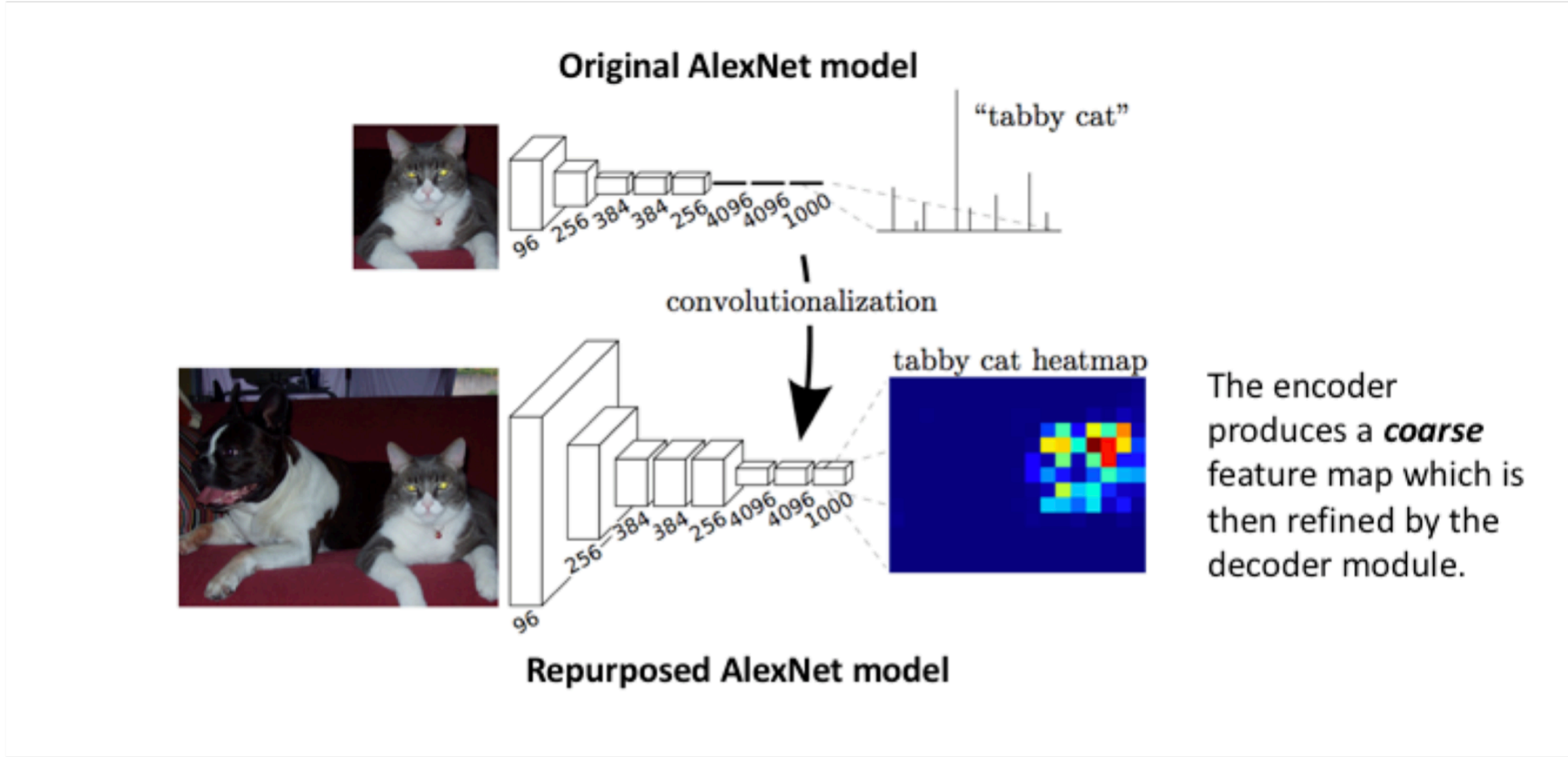
- **The Intersection over Union (IoU)**
  - also referred to as the Jaccard index

$$IoU = \frac{target \cap prediction}{target \cup prediction}$$

Semantic Segmentation Problem



Fully Convolutional Network (FCN)

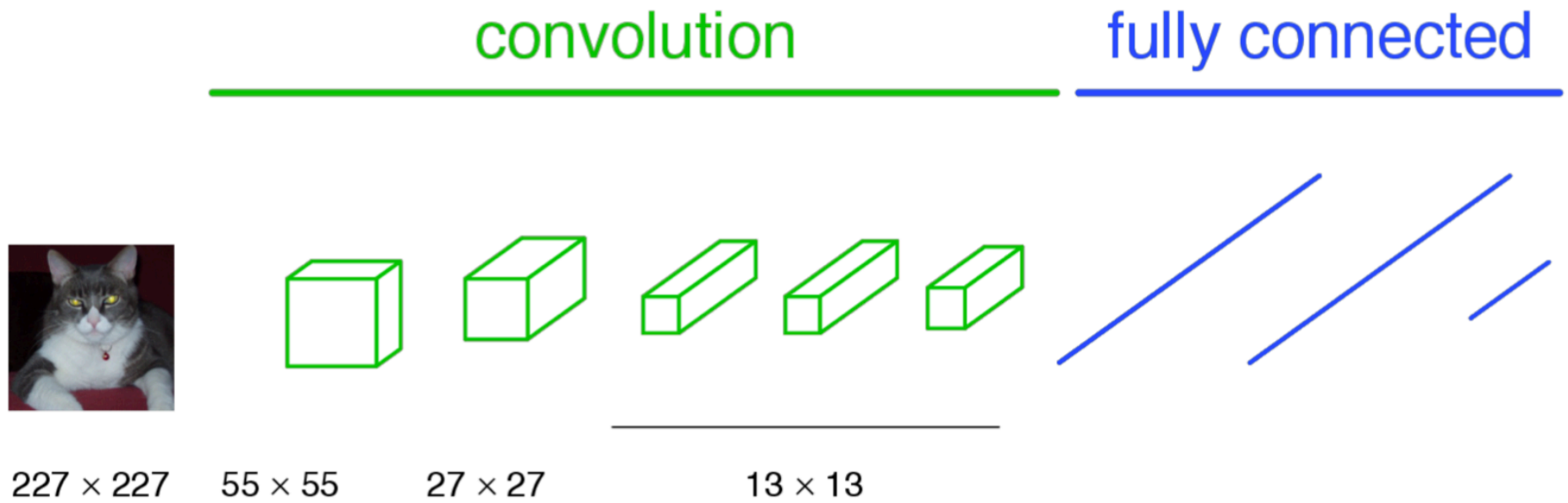


The encoder produces a ***coarse*** feature map which is then refined by the decoder module.

Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation.", ICCV. 2015.

Fully Convolutional Network (FCN)

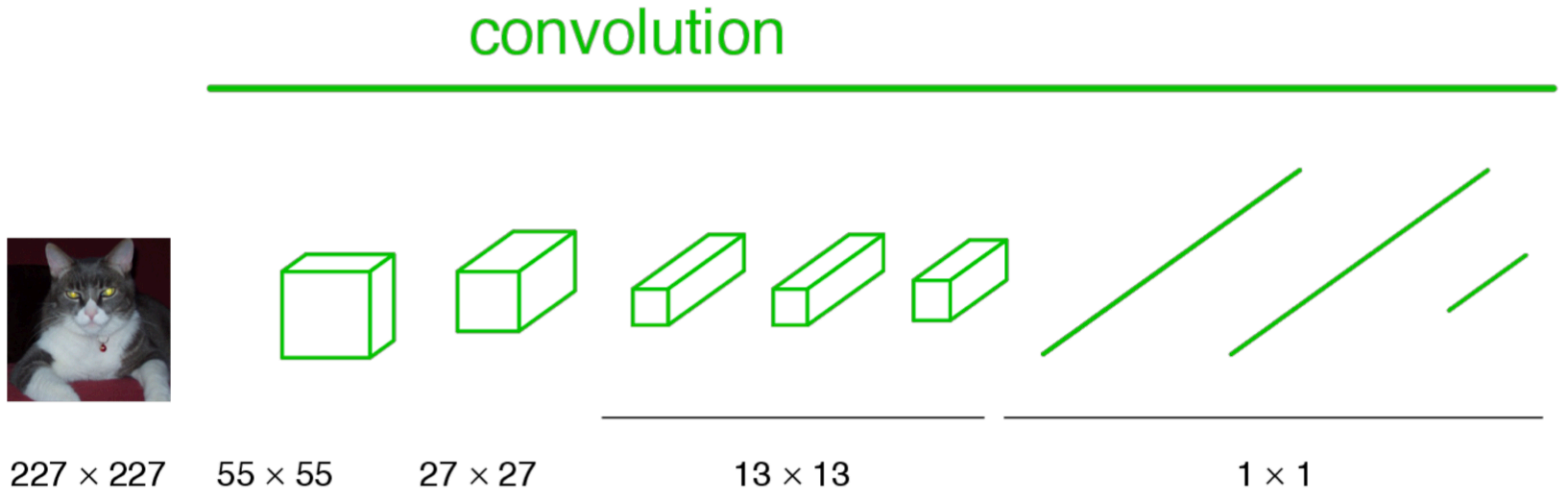
- a classification network



Slide credit: Jonathan Long

Fully Convolutional Network (FCN)

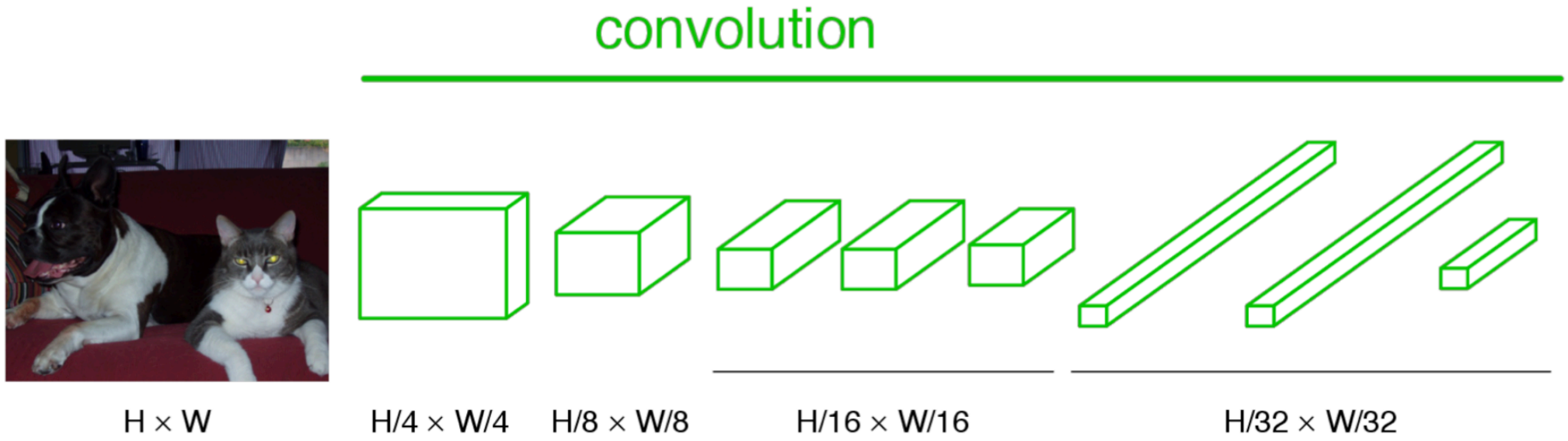
- becoming fully convolutional



Slide credit: Jonathan Long

Fully Convolutional Network (FCN)

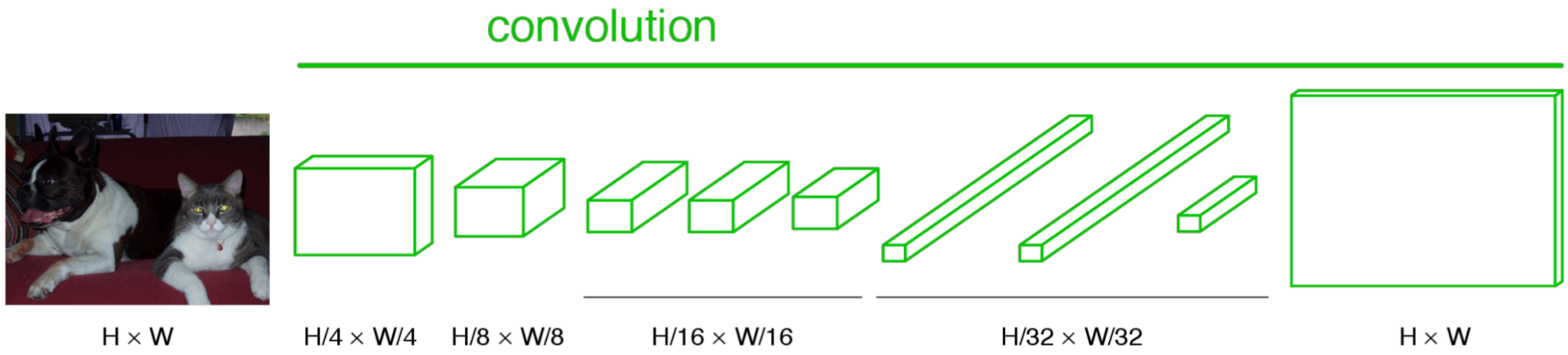
- becoming fully convolutional



Slide credit: Jonathan Long

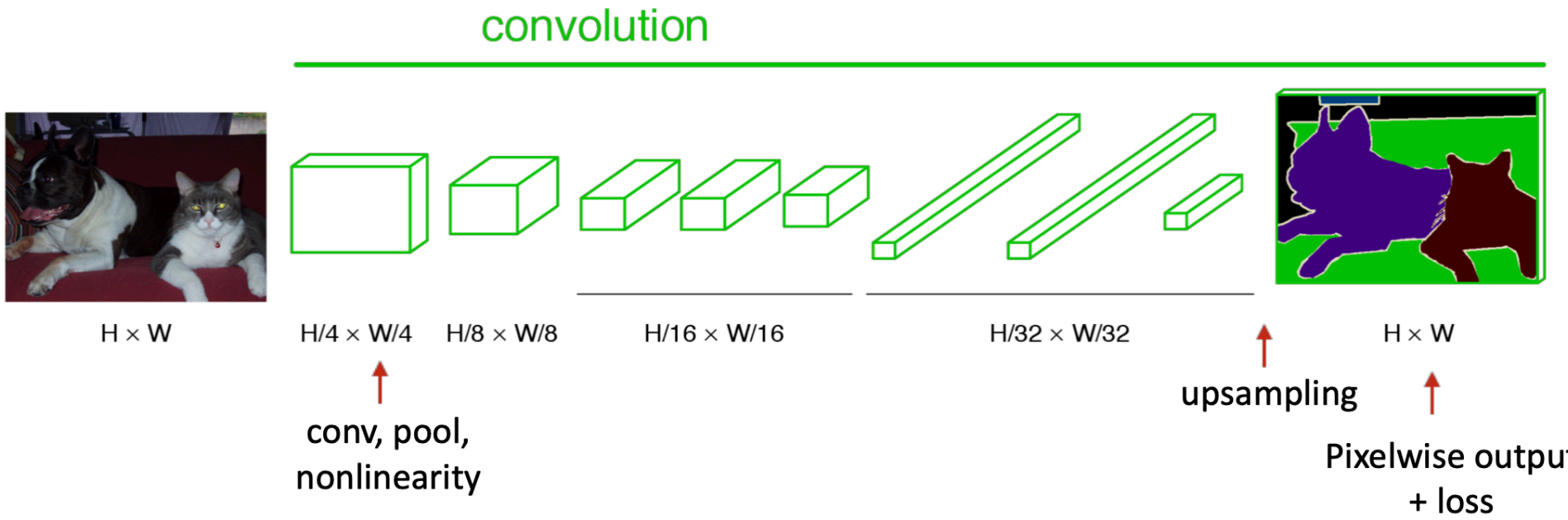
Fully Convolutional Network (FCN)

- **Upsampling output**



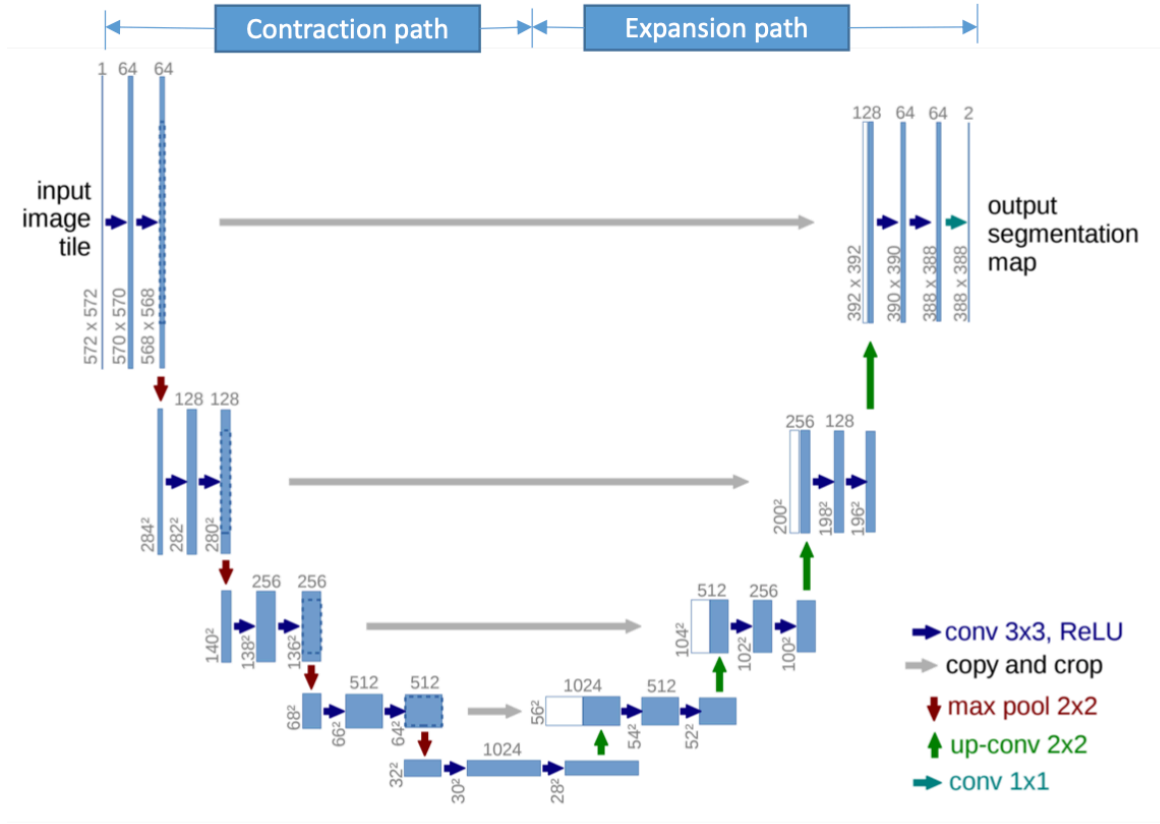
Fully Convolutional Network (FCN)

- End-to-end, pixel-to-pixel network



U-Net

- U-net composed in two parts:
  - contracting or downsampling
  - expanding or upsampling
- Feature maps from the downsampling part of the network are copied within the upsampling part
  - to avoid loosing pattern information.
- Finally, a 1x1 convolution processes the feature maps
  - to generate a segmentation map and thus categorise each pixel of the input image.



Ronneberger, Olaf, et al. "U-net: Convolutional networks for biomedical image segmentation.", MICCAI, 2015.