Spring 2024

CMPE 258-01 Deep Learning

Dr. Kaikai Liu, Ph.D. Associate Professor

Department of Computer Engineering

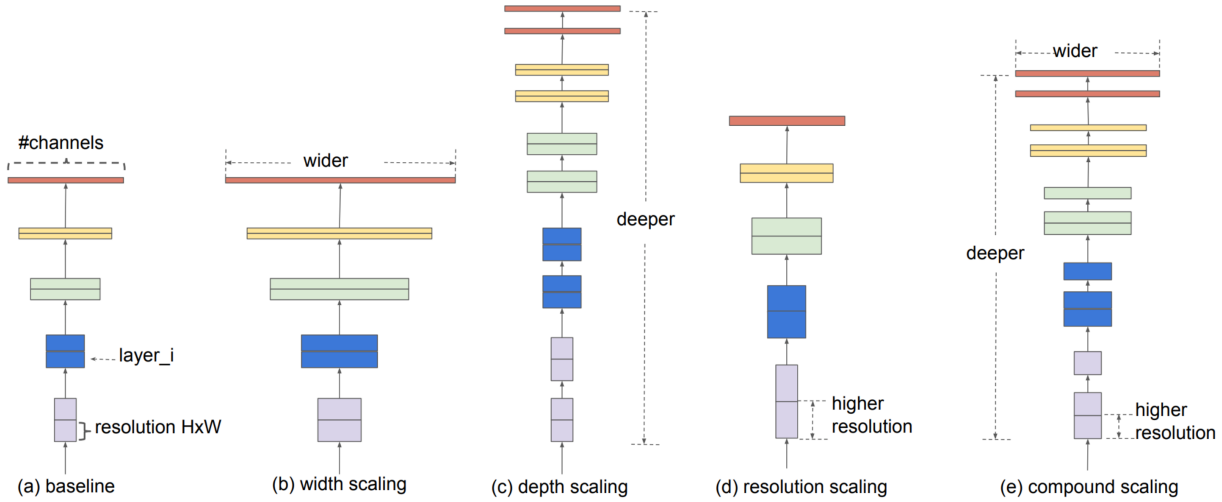
San Jose State University

Email: kaikai.liu@sjsu.edu

Website: https://www.sjsu.edu/cmpe/faculty/tenure-line/kaikai-liu.php

EfficientNet

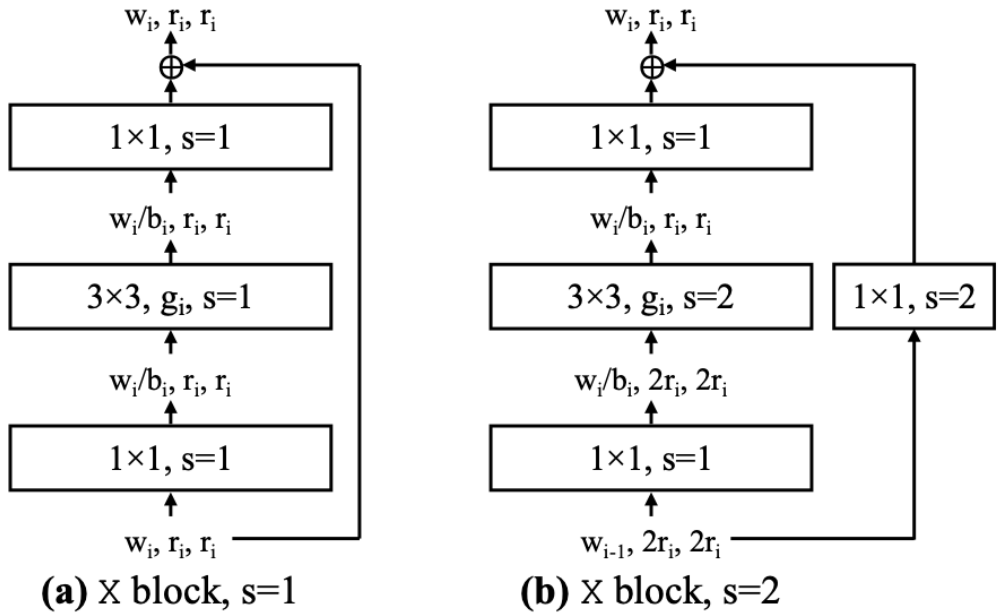
- EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
  - Paper: https://arxiv.org/abs/1905.11946
  - EfficientNet is a convolutional neural network architecture and scaling method that uniformly scales all dimensions of depth/width/resolution using a compound coefficient.
    - uniformly scales network width, depth, and resolution with a set of fixed scaling coefficients.



The base EfficientNet-B0 network is based on the inverted bottleneck residual blocks of MobileNetV2, in addition to squeeze-and-excitation blocks.

RegNet

- Designing Network Design Spaces (2020)
  - Paper: https://arxiv.org/pdf/2003.13678.pdf
  - https://github.com/facebookresearch/pycls/tree/main
  - Design space is defined by model building parameters which have its own range, and therefore defines the range of possible model structures.
    - by chasing design spaces instead of individual networks, we can discover general design principles that work across general settings.



(a) X block, s=1 **(b)** X block, s=2

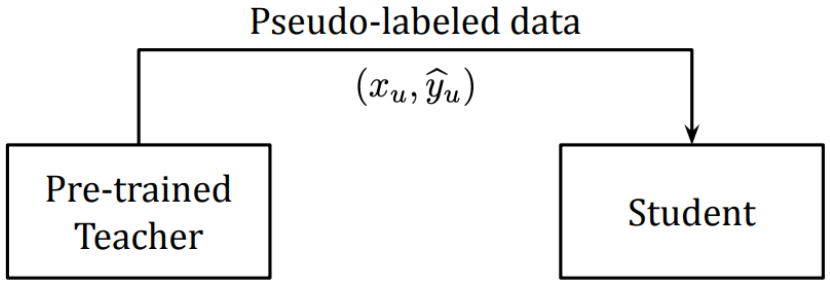
Figure 4. The **X block** is based on the standard residual bottleneck block with group convolution [26]. (a) Each X block consists of a 1×1 conv, a 3×3 group conv, and a final 1×1 conv, where the 1×1 convs alter the channel width. BatchNorm [9] and ReLU follow each conv. The block has 3 parameters: the width w_i, bottleneck ratio b_i, and group width g_i. (b) The stride-two ($s = 2$) version.

BiT

- Big Transfer (BiT): General Visual Representation Learning (2020)
  - Paper: https://arxiv.org/pdf/1912.11370.pdf
  - We revisit the paradigm of pre-training on large supervised datasets and fine-tuning the model on a target task. We scale up pre-training, and propose a simple recipe that we call Big Transfer (BiT).
  - We transfer BiT to many diverse tasks; with training set sizes ranging from 1 example per class to 1M total examples. These tasks include ImageNet's ILSVRC-2012, CIFAR-10/100, Oxford-IIIT Pet, Oxford Flowers-102 (including few-shot variants), and the 1000-sample VTAB-1k benchmark, which consists of 19 diverse datasets. BiT-L attains state-of- the-art performance on many of these tasks, and is surprisingly effective when very little downstream data is available. We also train BiT-M on the public ImageNet-21k dataset, and attain marked improvements over the popular ILSVRC-2012 pre-training.

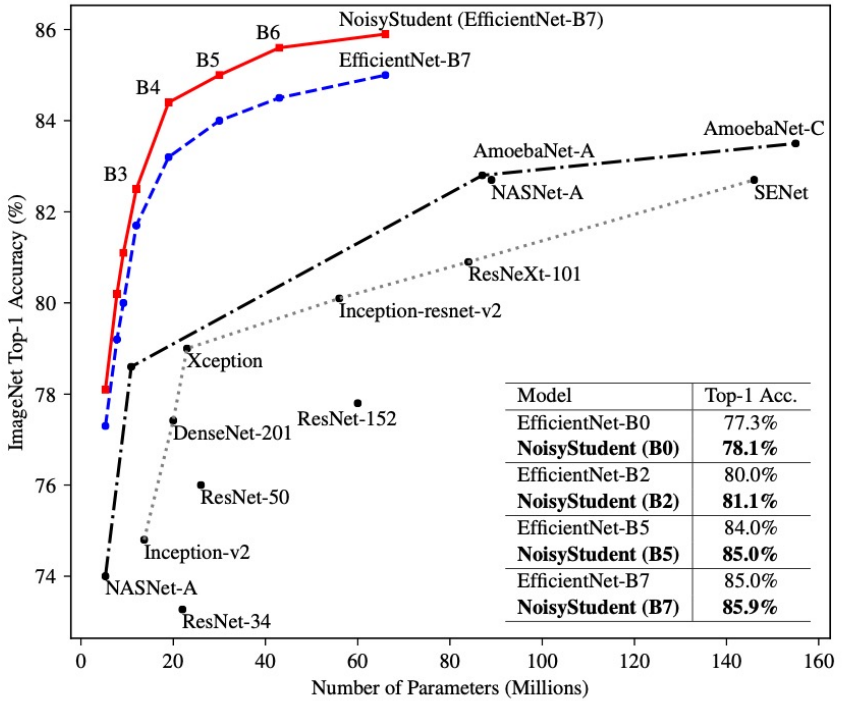
Pseudo Labels or self-training

- Pseudo Labels methods work by having a pair of networks, one as a teacher and one as a student. The teacher generates pseudo labels on unlabeled images. These pseudo labeled images are then combined with labeled images to train the student. Thanks to the abundance of pseudo labeled data and the use of regularization methods such as data augmentation, the student learns to become better than the teacher
- One main drawback: if the pseudo labels are inaccurate, the student will learn from inaccurate data. As a result, the student may not get significantly better than the teacher. This drawback is also known as the problem of confirmation bias in pseudo-labeling

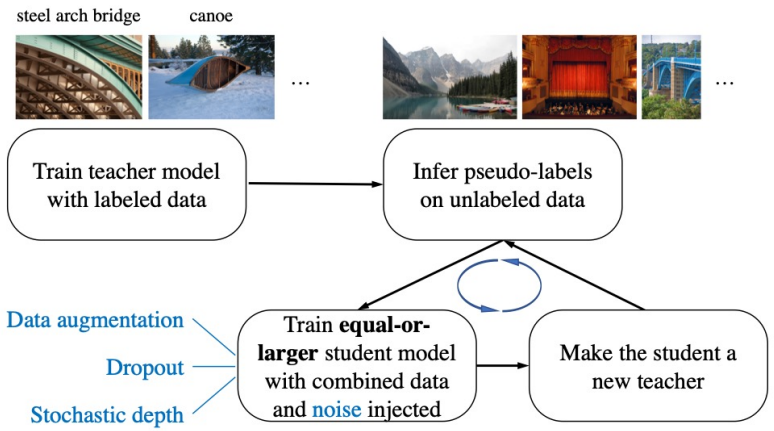


Noisy Student

- Self-training with Noisy Student improves ImageNet classification (2020)
  - Paper: [link]
  - First train an EfficientNet model on labeled ImageNet images and use it as a teacher to generate pseudo labels on 300M unlabeled images. Then train a larger EfficientNet as a student model on the combination of labeled and pseudo labeled images.
  - Iterate this process by putting back the student as the teacher. During the generation of the pseudo labels, the teacher is not noised so that the pseudo labels are as accurate as possible. However, during the learning of the student, we inject noise such as dropout, stochastic depth and data augmentation via RandAugment to the student so that the student generalizes better than the teacher.

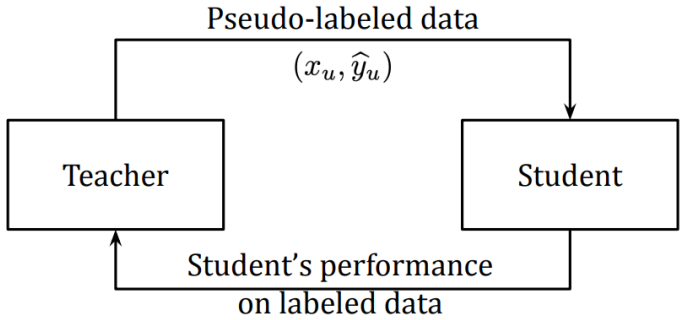


Model	Top-1 Acc.
EfficientNet-B0	77.3%
NoisyStudent (B0)	**78.1%**
EfficientNet-B2	80.0%
NoisyStudent (B2)	**81.1%**
EfficientNet-B5	84.0%
NoisyStudent (B5)	**85.0%**
EfficientNet-B7	85.0%
NoisyStudent (B7)	**85.9%**



Pseudo-Labels

- Meta Pseudo-Labels (2021) -- Google AI, Brain Team
  - Paper: https://arxiv.org/abs/2003.10580
  - We present Meta Pseudo Labels, a **semi-supervised learning** method that achieves a new state-of-the-art top-1 accuracy of 90.2% on ImageNet, which is 1.6% better than the existing state-of-the-art.
  - Like Pseudo Labels, Meta Pseudo Labels has a teacher network to generate pseudo labels on unlabeled data to teach a student network. However, unlike Pseudo Labels where the teacher is fixed, the teacher in Meta Pseudo Labels is constantly adapted by the feedback of the student's performance on the labeled dataset. As a result, the teacher generates better pseudo labels



Knowledge Distillation

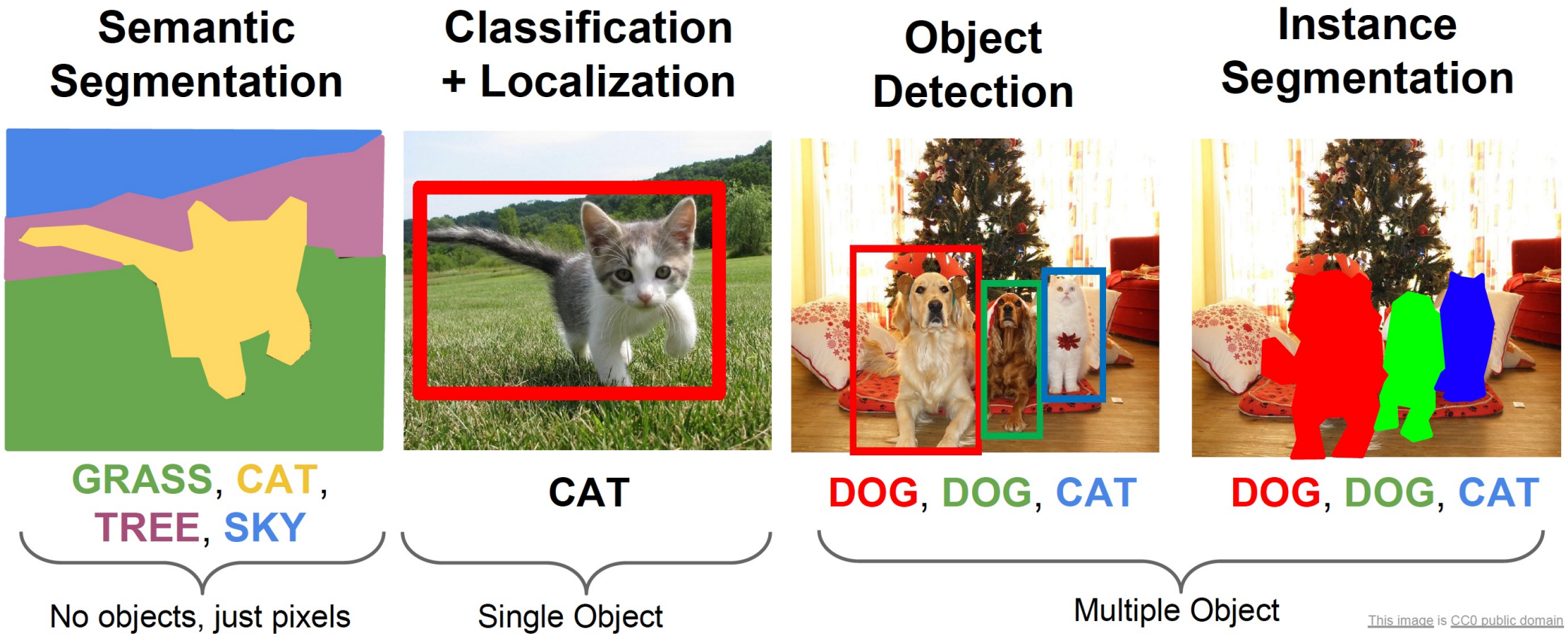
- Knowledge Distillation and Label Smoothing.
  - The teacher in Meta Pseudo Labels uses its softmax predictions on unlabeled data to teach the student. These softmax predictions are generally called the soft labels, which have been widely utilized in the literature on knowledge distillation

Model name	Number of parameters [Millions]	ImageNet Top 1 Accuracy	Year
AlexNet	60 M	63.3 %	2012
Inception V1	5 M	69.8 %	2014
VGG 16	138 M	74.4 %	2014
VGG 19	144 M	74.5 %	2014
Inception V2	11.2 M	74.8 %	2015
ResNet-50	26 M	77.15 %	2015
ResNet-152	60 M	78.57 %	2015
Inception V3	27 M	78.8 %	2015
DenseNet-121	8 M	74.98 %	2016
DenseNet-264	22M	77.85 %	2016
BiT-L (ResNet)	928 M	87.54 %	2019
NoisyStudent EfficientNet-L2	480 M	88.4 %	2020
Meta Pseudo Labels	480 M	90.2 %	2021

Deep Learning based Object Detection

Deep learning based object detection

- Perception tasks based on Computer Vision



From cs231n 2018 Stanford, Fei-Fei Li

Deep learning based object detection

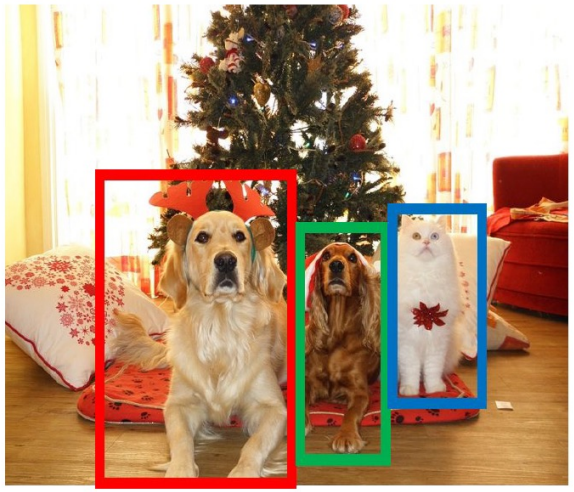
- Perception tasks based on Computer Vision

Semantic Segmentation

GRASS, CAT, TREE, SKY

No objects, just pixels

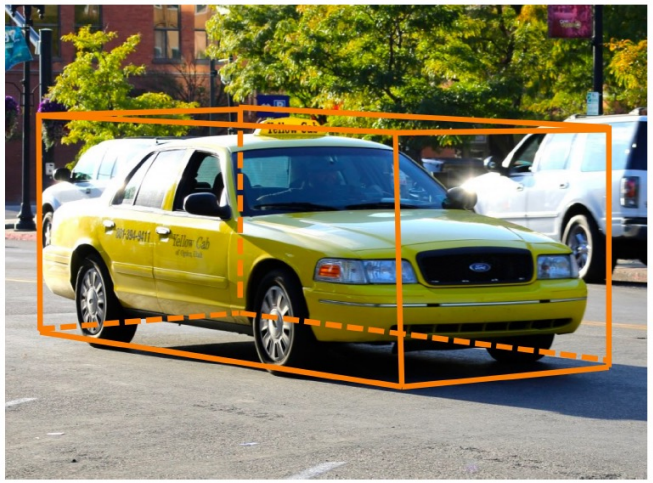
2D Object Detection



DOG, DOG, CAT

Object categories + 2D bounding boxes

3D Object Detection



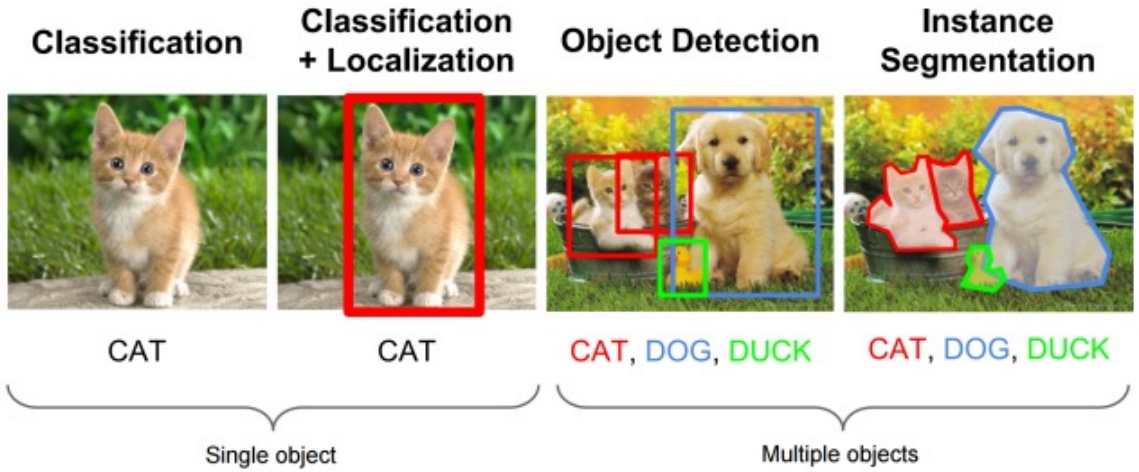
Car

Object categories + 3D bounding boxes

From cs231n 2018 Stanford, Fei-Fei Li

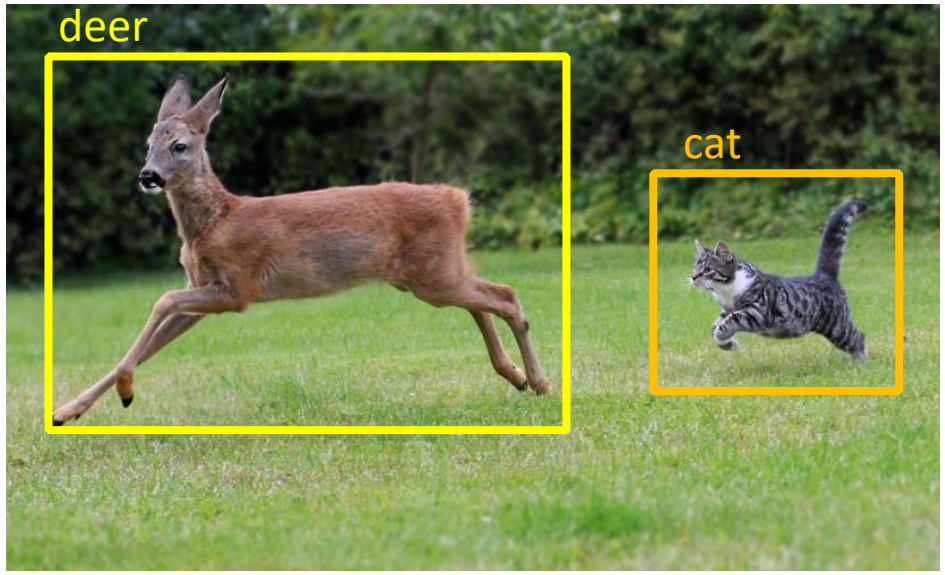
Deep learning based object detection

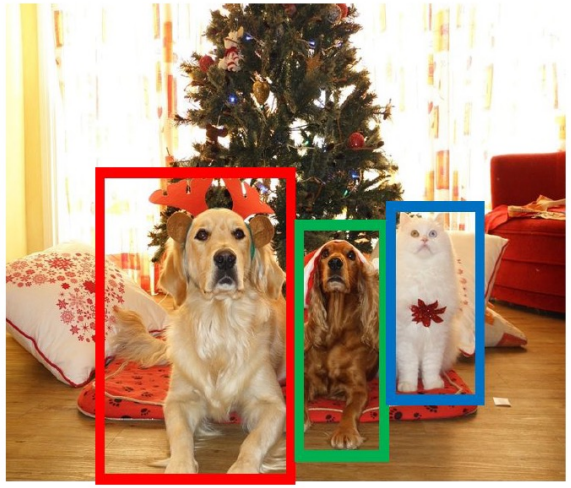
- (Classic) Deep learning based object detection
  - The RCNN Object Detector (2014)
  - The Fast RCNN Object Detector (2015)
  - The Faster RCNN Object Detector (2016)
  - The YOLO Object Detector (2016)
  - The SSD Object Detector (2016)
  - Mask-RCNN (2017)



Object Detection

- Object detection
  - Object detection as classification
  - 2D object detection with object categories+2D bounding boxes



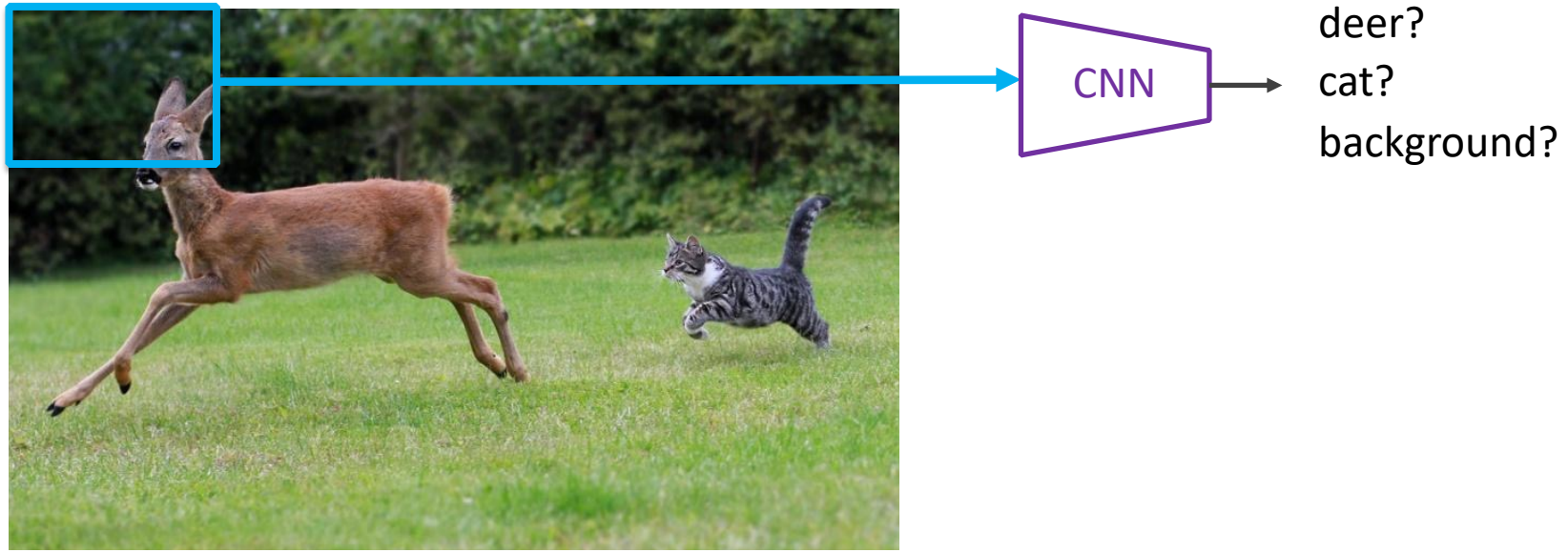


DOG, DOG, CAT

Object categories +
2D bounding boxes

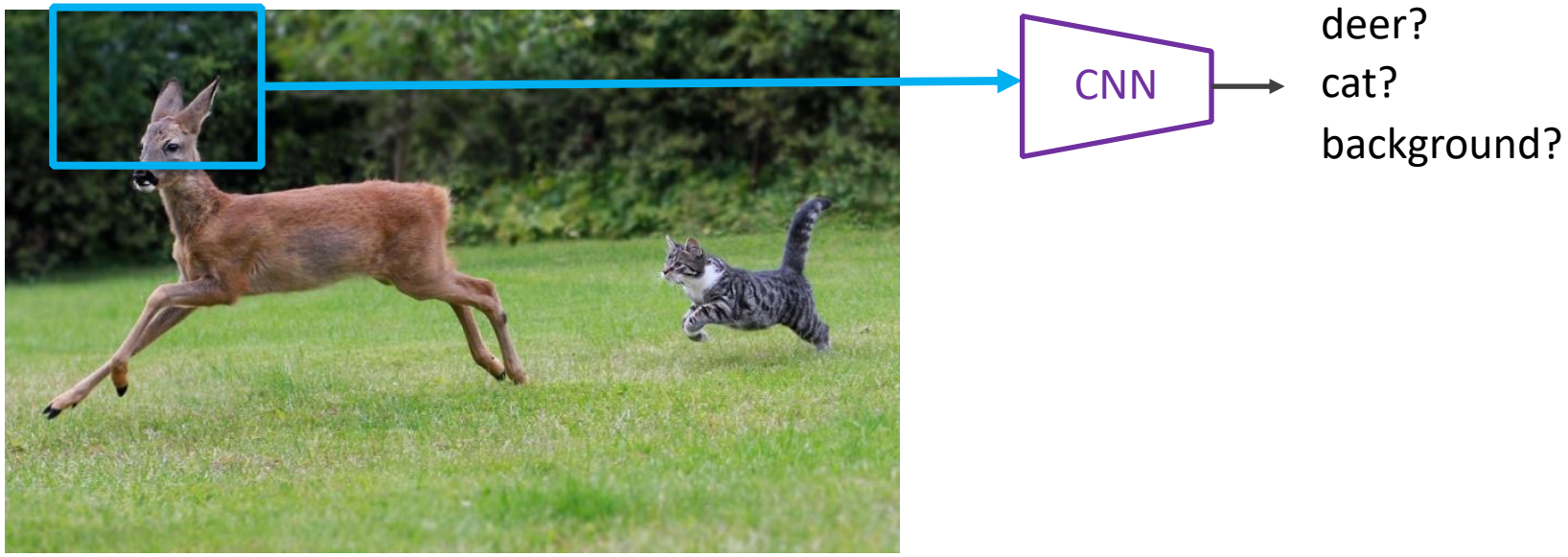
Object Detection

- Object detection
  - Object detection as classification



Object Detection

- Object detection
  - Object detection as classification with **Sliding Window**

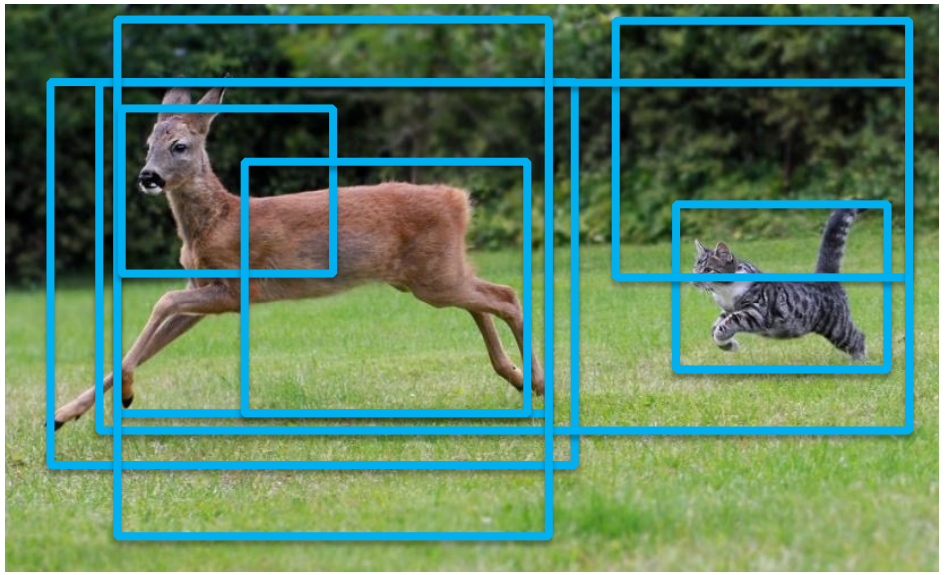


Object Detection

- Object detection
  - Object detection as classification with **Sliding Window**
- Sliding Window Process
  - Take a set of image patches
  - For each patch, classify what object the patch contains, using an object classifier trained with hand-crafted features
- Do not know the location of the object of interest
  - sliding window at small strides
- Do not know the size/ aspect ratio of the object of interest
  - windows of different sizes
- Need to apply CNN to huge number of locations, scales, and aspect ratios, very computationally expensive!

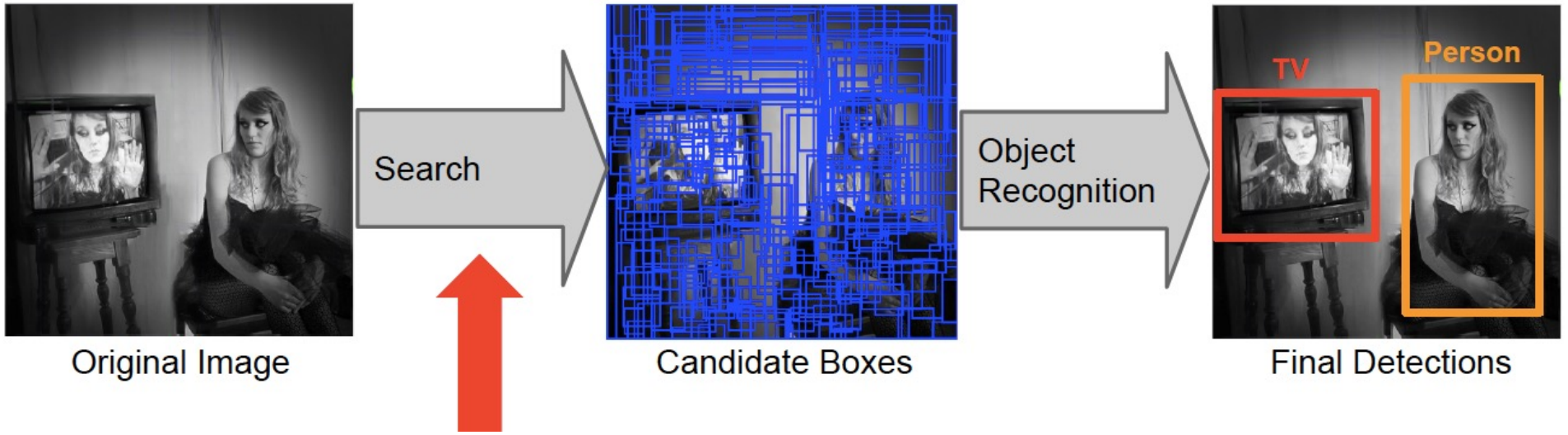
Object Detection

- Object detection
  - Object detection as classification with **Box Proposals**



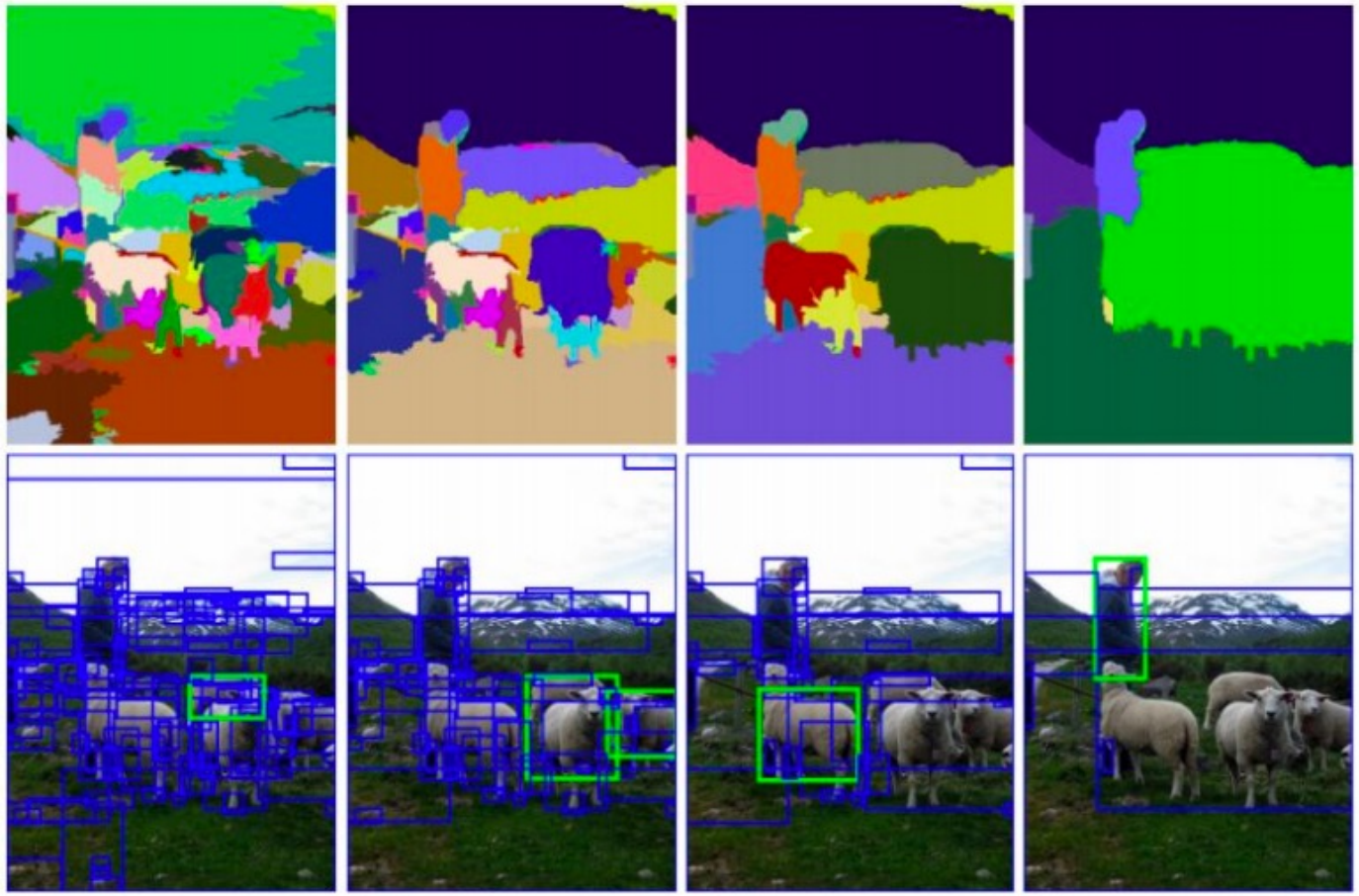
Object Detection

- Object detection
  - Object detection as classification with **Box Proposals**
- As an alternative to sliding window search, evaluate a few hundred region proposals
  - Can use slower but more powerful features and classifiers
  - Take advantage of low-level perceptual organization cues
  - Proposal mechanism can be category-independent
  - Proposal mechanism can be trained



Box Proposal

- Box Proposal Method – SS: Selective Search



Segmentation As Selective Search for Object Recognition. van de Sande et al. ICCV 2011

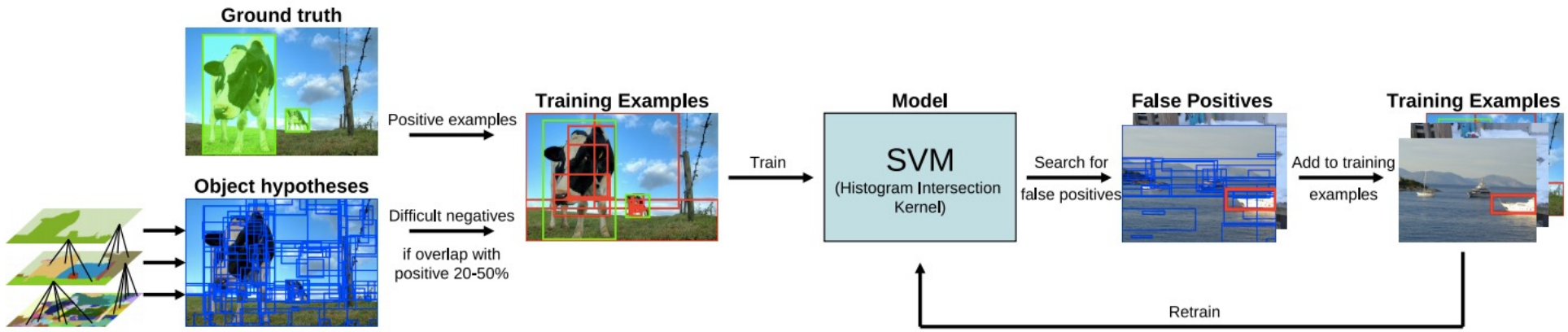
J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013

Box Proposal

- Box Proposal Method – SS: Selective Search
- Use hierarchical segmentation: start with small *superpixels* and merge based on diverse cues

Selective search detection pipeline

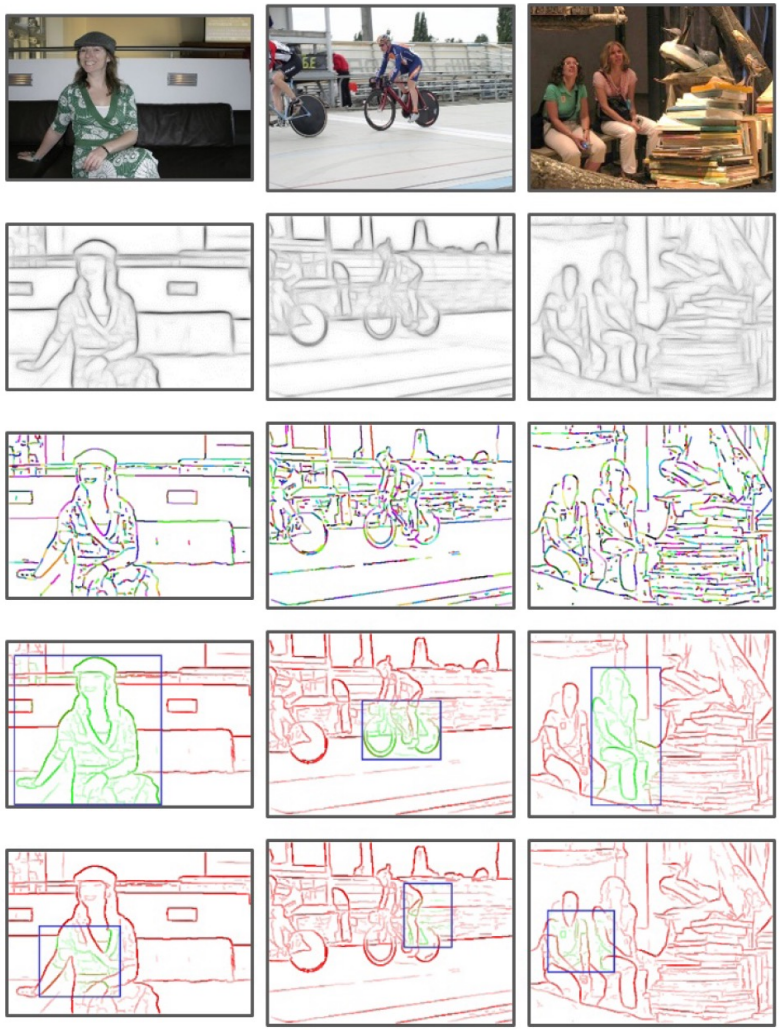
- Selective search detection pipeline
  - Feature extraction: color SIFT, codebook of size 4K, spatial pyramid with four levels = 360K dimensions



J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013

Box Proposal - EdgeBoxes

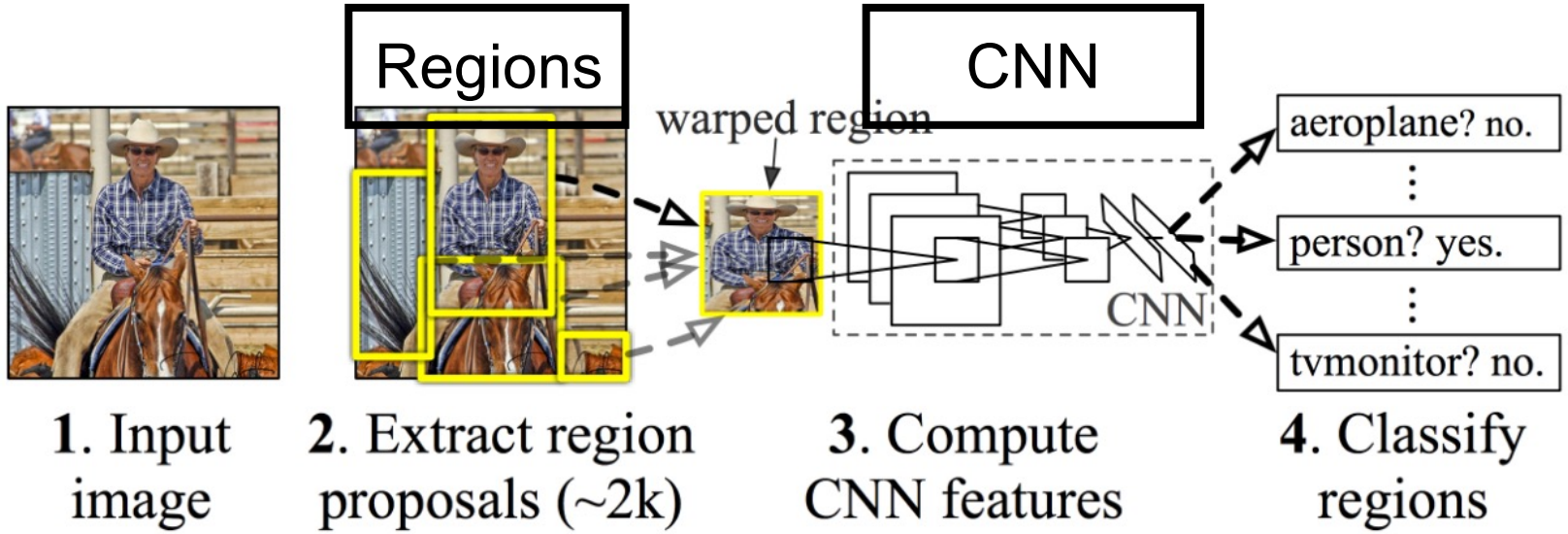
- Another proposal method: EdgeBoxes
  - Box score: number of edges in the box minus number of edges that overlap the box boundary
  - Uses a trained edge detector
  - Uses efficient data structures (incl. integral images) for fast evaluation
  - Gets 75% recall with 800 boxes (vs. 1400 for Selective Search), is 40 times faster



C. Zitnick and P. Dollar, Edge Boxes: Locating Object Proposals from Edges, ECCV 2014.

R-CNN

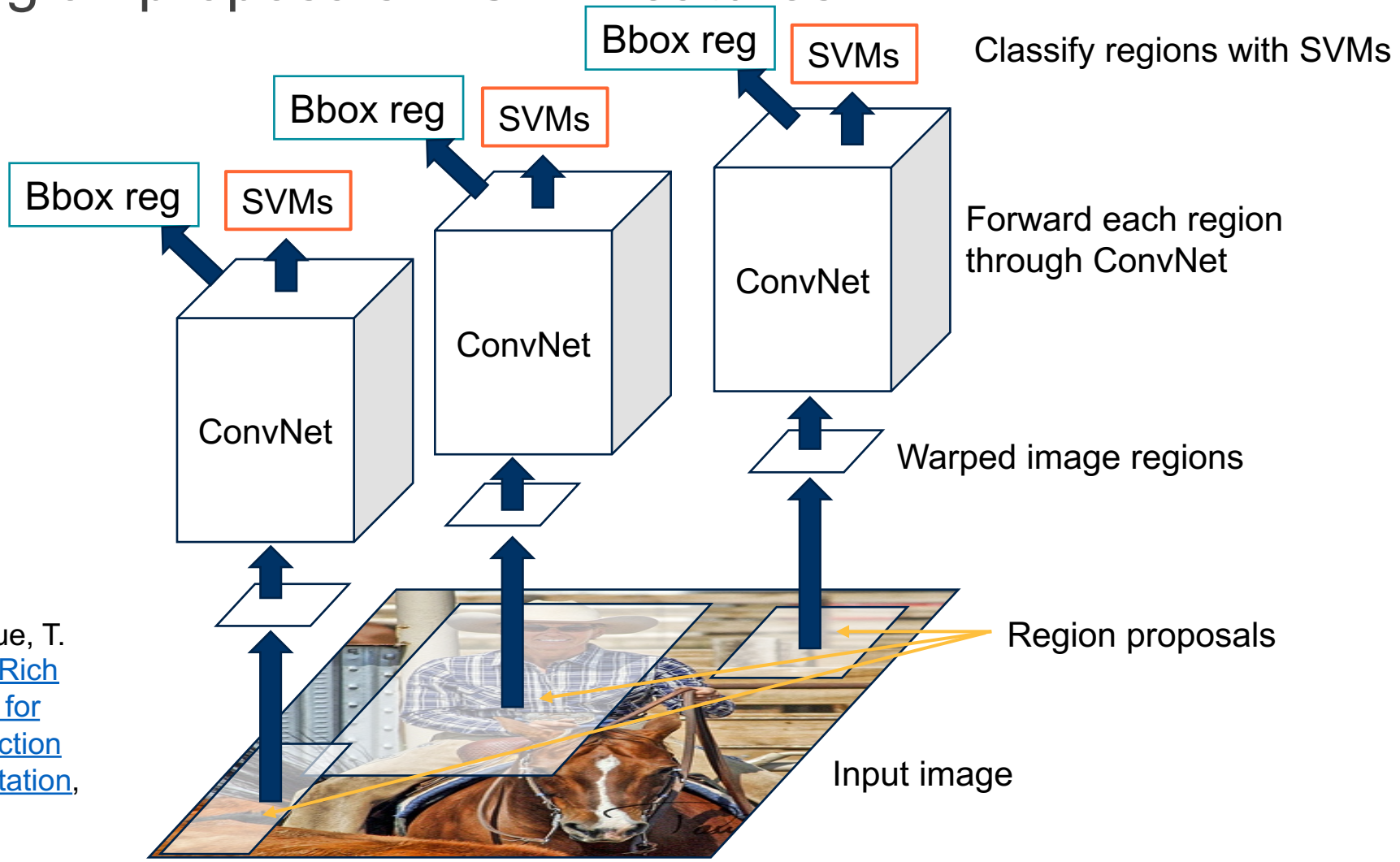
- R-CNN (Girshick et al. CVPR 2014)
  - Region proposals reduce brute force search as in sliding window
  - Each proposed region is warped to match input size expected by CNN
  - CNN gives more discriminatory features, compared to hand-crafted features
  - Multi-class SVM trained with CNN features



R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014.

R-CNN

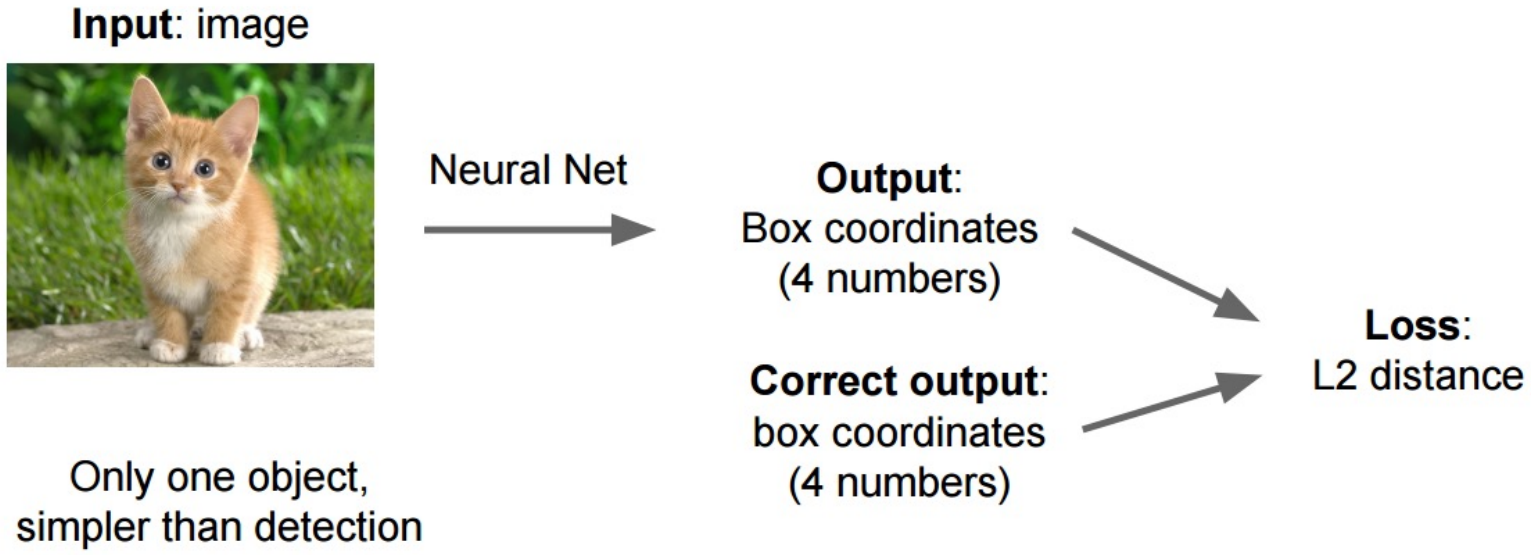
- R-CNN: Region proposals + CNN features



R. Girshick, J. Donahue, T. Darrell, and J. Malik, Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014.

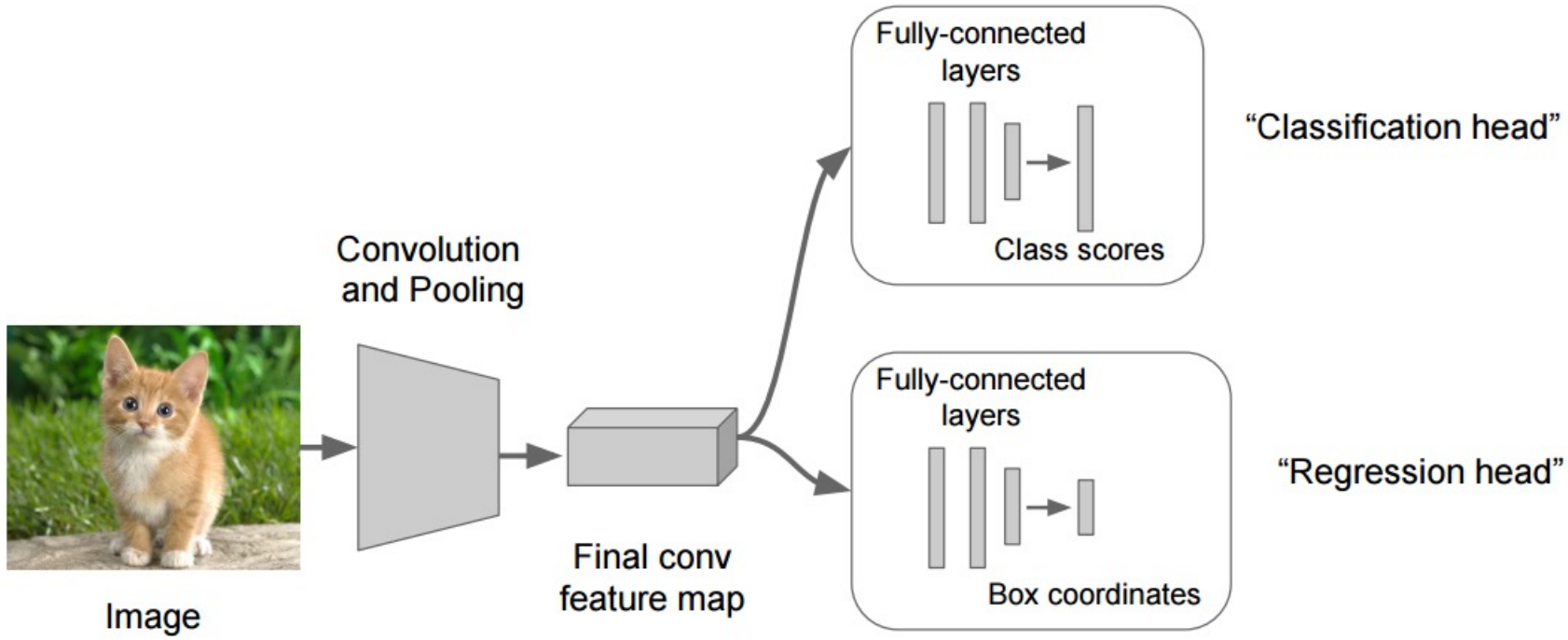
Localize objects with regression

- Regression is about returning a number instead of a class, in our case we're going to return 4 numbers (x0,y0,width,height) that are related to a bounding box.
  - You train this system with an image an a ground truth bounding box, and use L2 distance to calculate the loss between the predicted bounding box and the ground truth.

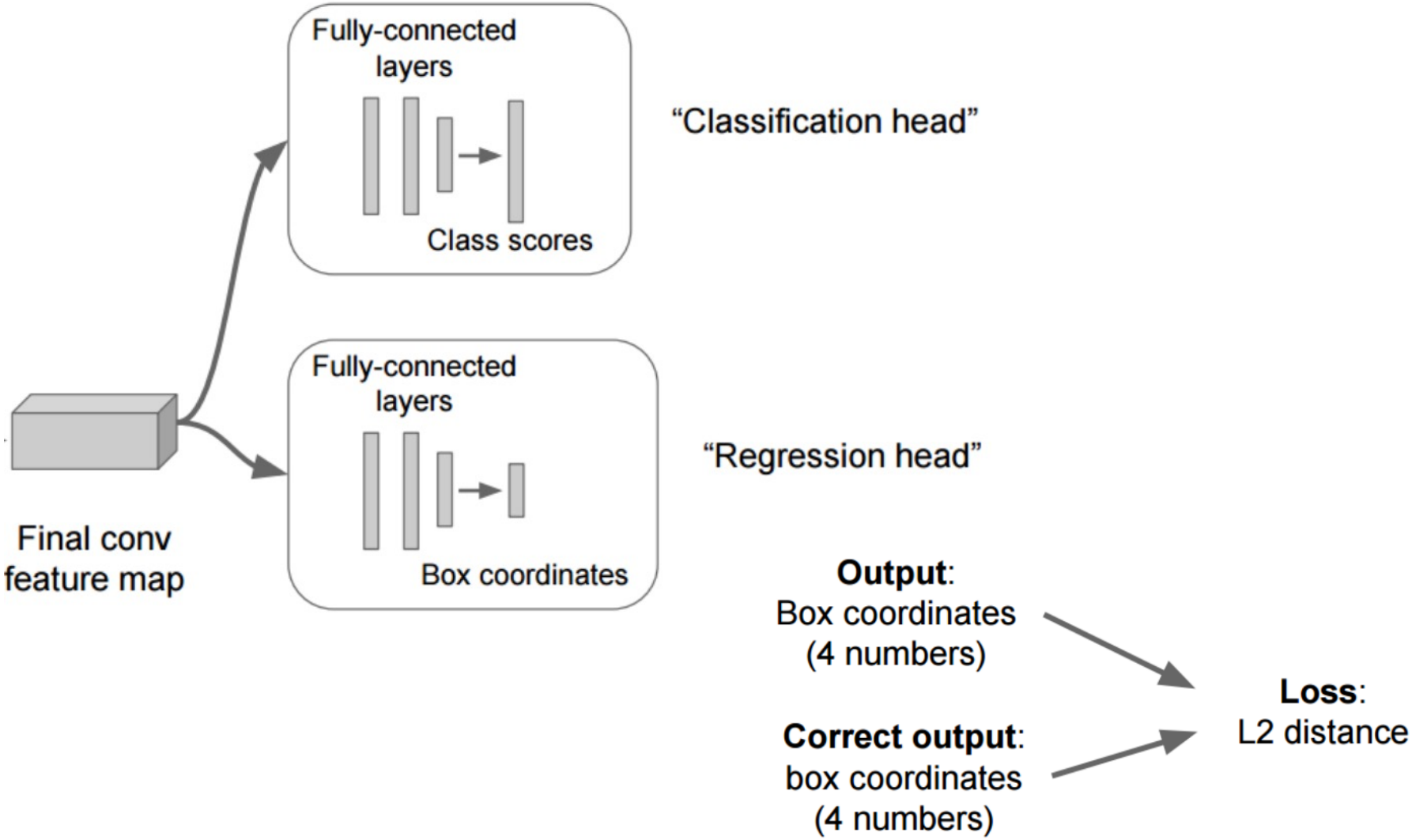


Localize objects with regression

- Normally what you do is attach another fully connected layer on the last convolution layer

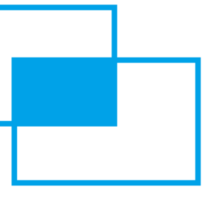


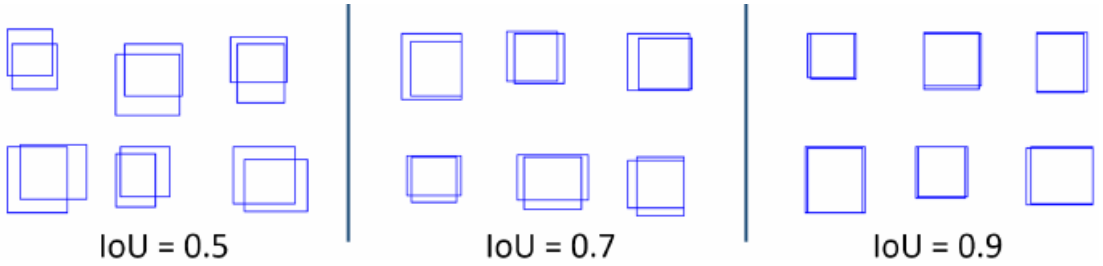
Bounding box regressor

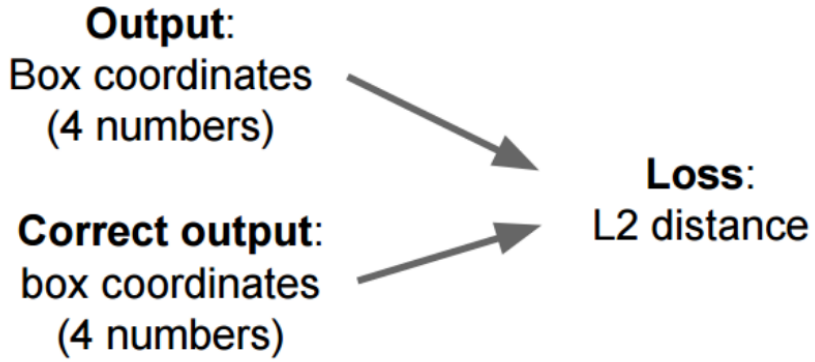


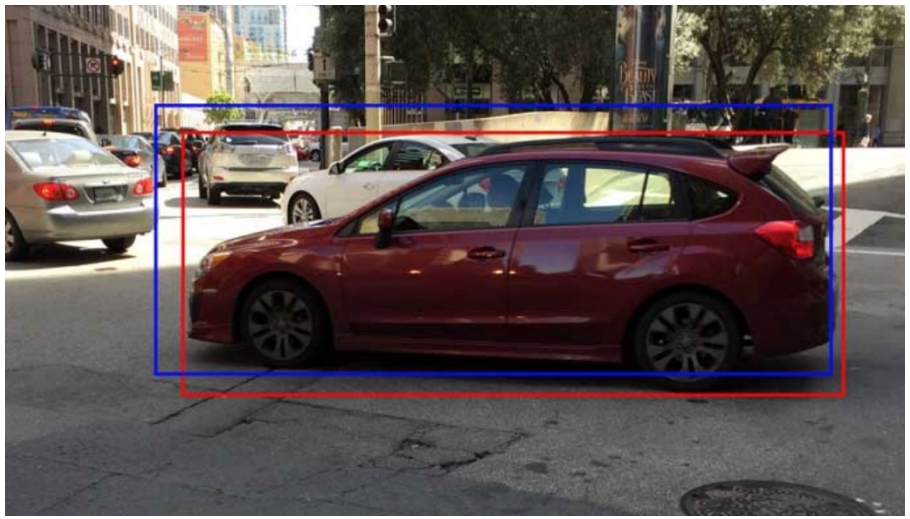
R-CNN

- Bounding box regressor
  - During training, a bounding box regressor is learnt minimizing regression loss compared to ground truth bounding boxes.
- Comparing bounding box prediction accuracy
  - Compare if the Intersect Over Union (ioU) between the prediction and the ground truth is bigger than some threshold (ex > 0.5)









Thank You

Address:
ENG257, SJSU

Email Address:
Kaikai.liu@sjsu.edu