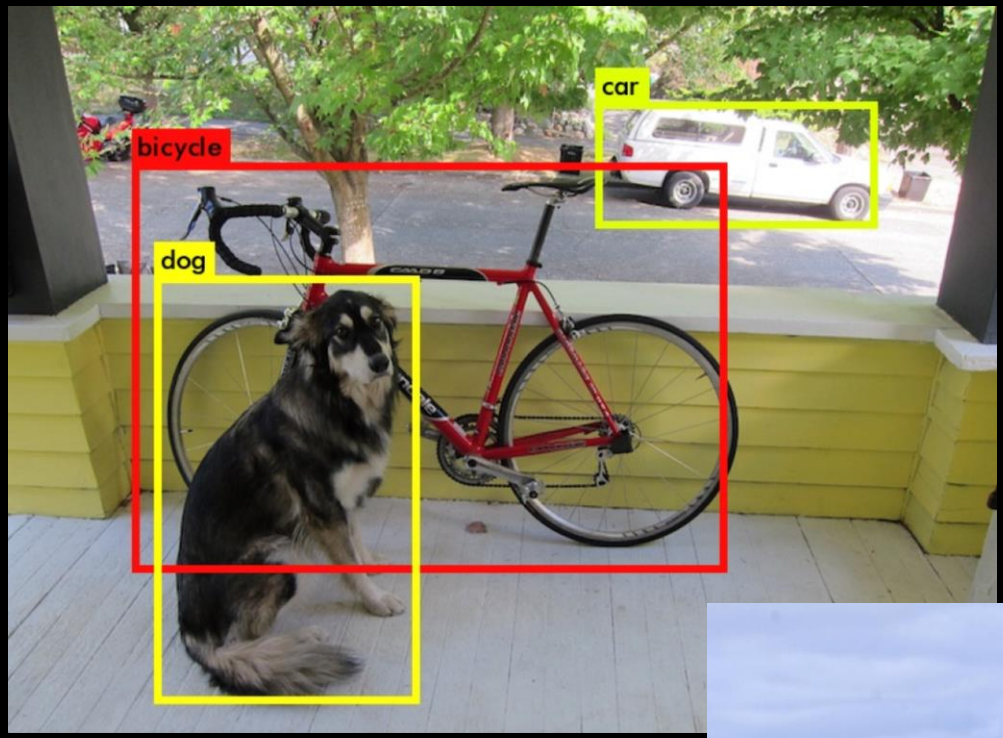
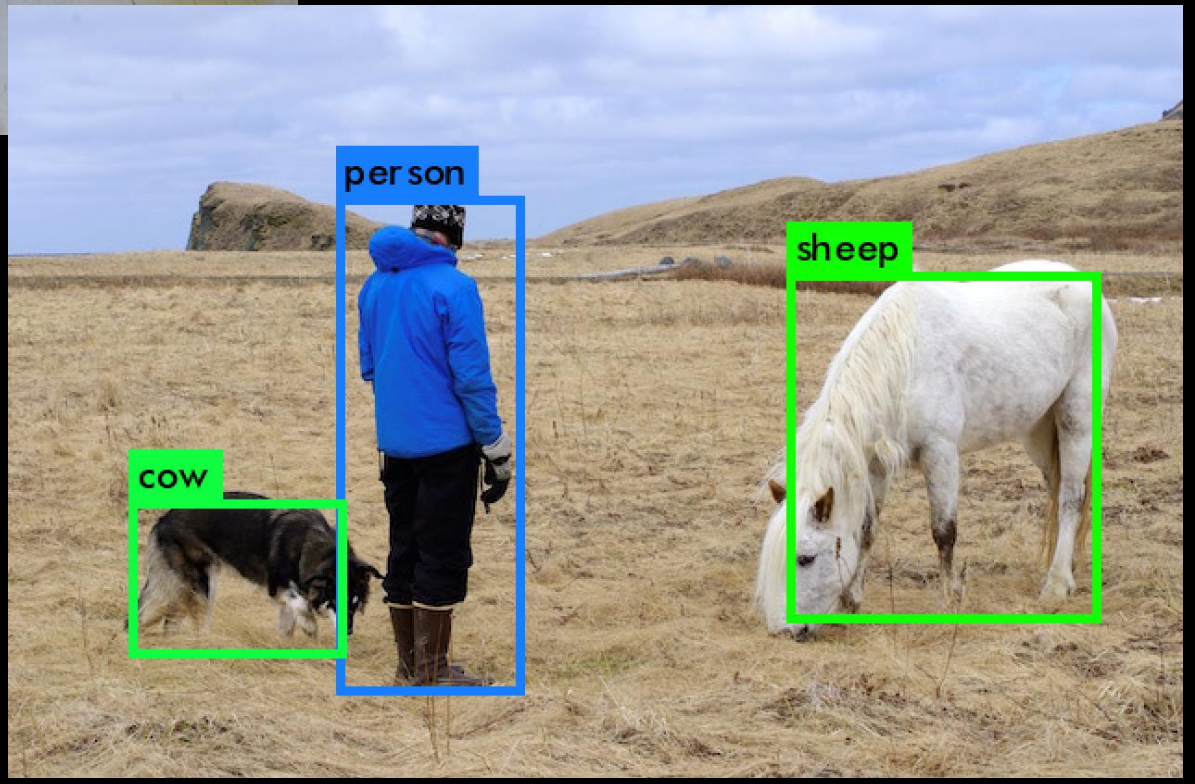
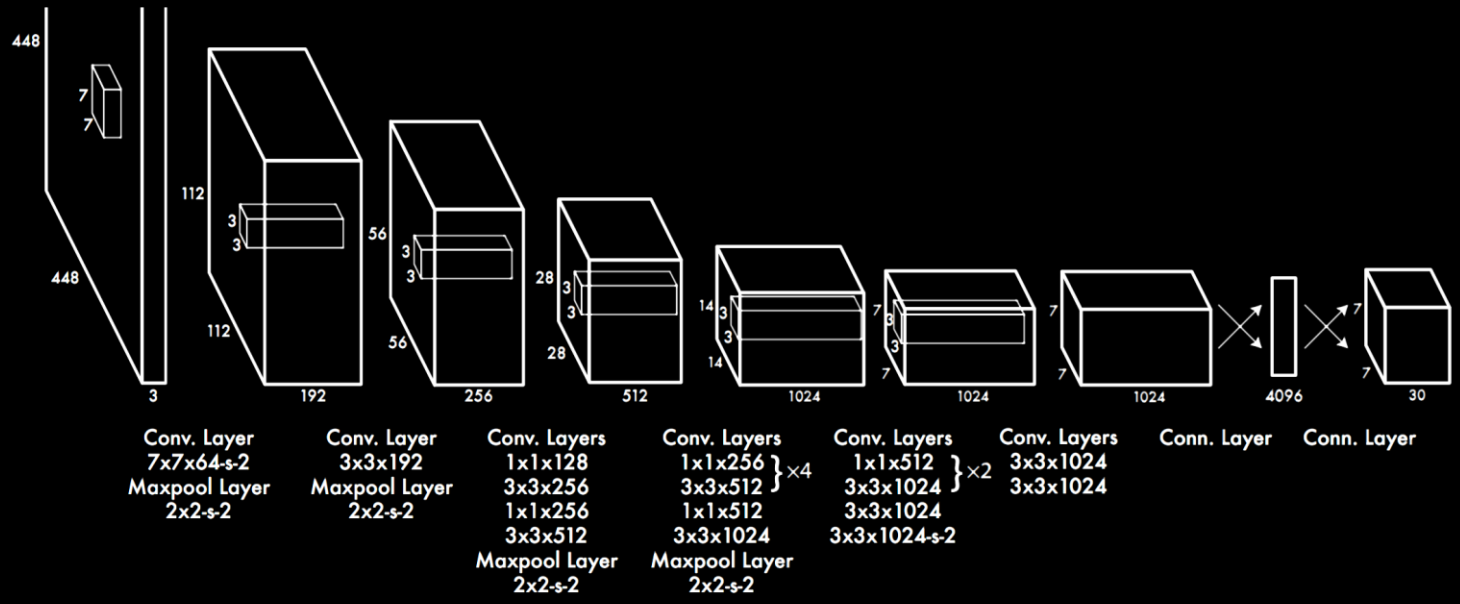
You Only Look Once: Unified, Real-Time Object Detection

Joseph Redmon University of Washington
Santosh Divvala Allen Institute for Artificial Intelligence
Ross Girshick Facebook AI Research
Ali Farhadi University of Washington

IEEE Conference on Computer Vision and Pattern Recognition, 2016

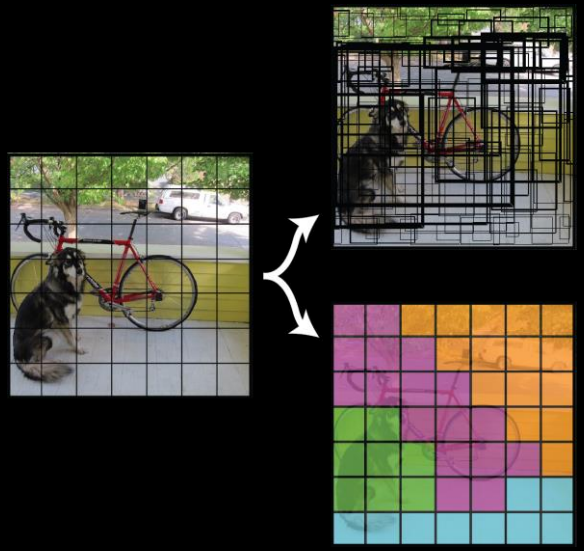

car
bicycle
dog


person
sheep
cow



The output block is $7 \times 7 \times (5B + K)$.

- 7×7 is a grid of output cells
- Each cell is asked to produce $B = 2$ bounding boxes consisting of (x, y, w, h, c)
- $(x, y) \in [0,1]^2$ is the centre of the box w.r.t. the cell
  $(w, h) \in [0,1]^2$ is the size of the box w.r.t. the image
  $c \in [0,1]$ is a confidence score
- Each cell is asked to produce $K = 20$ logit scores, for each of K classes



It seems to me peculiar to encode “This scene has two objects, a k_1 and a k_2 in boxes b_1 and b_2” as data in a 7×7 array. Why not encode it as a sequence, or a maybe a tree if there is more semantic knowledge?

This approach to object detection has been extended, for detecting moving objects in a video stream. (How fast is the car in this bounding box moving?) The approach has been “Can I predict the boxes and their data, given two successive frames of the video?” Sequence/recurrent approaches have not yet been successful.

MultiNet: Real-time Joint Semantic Reasoning for Autonomous Driving (2016)

Marvin Teichmann, Michael Weber, Marius Zoellner, Roberto Cipolla, Raquel Urtasun

MODNet: Motion and Appearance based Moving Object Detection for Autonomous Driving (2017)

Mennatullah Siam, Heba Mahgoub, Mohamed Zahran, Senthil Yogamani, Martin Jagersand