CV / VLM

Unit 2: Introduction to Object Detection (OD)

TIL-AI
TODAY I LEARNED-AI

2.3.1

Single Shot Detectors (SSD) and YOLO

Introduction to SSD and YOLO architectures

OD Architectures: R-CNN Vs. SSD

Algorithms in the R-CNN family use a 2 step approach (Region proposal + CNN detection)

- + Better accuracy
- - Sub-seconds prediction

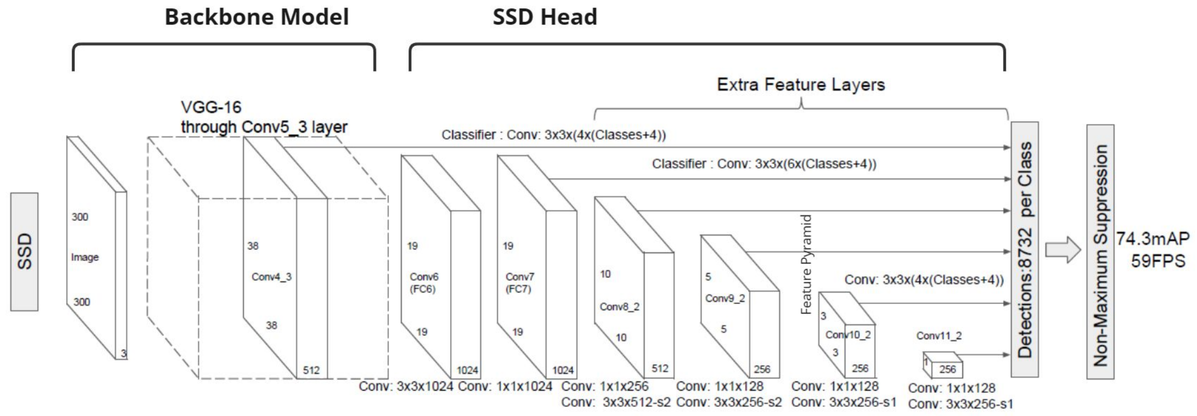
Single-shot detector (SSD) approaches utilize a deeper/custom neural network but only require a single pass (1 step)

- + More efficient
- + Real-time (predictions in milliseconds)
- - Moderate accuracy

Single-Shot Detector (SSD)

Two Components

1. **Backbone model:** Pre-trained image CNN (Resnet, VGG)
2. **SSD Head:** More convolutional layers added to the backbone, whereby the outputs are bounding boxes and classes of the objects in the spatial locations.

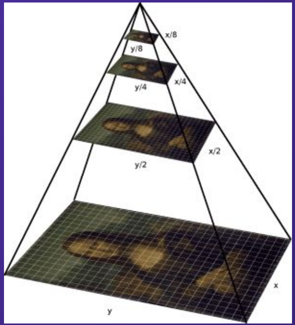


Wei Liu et al. in the paper SSD: Single Shot MultiBox Detector.

Single-Shot Detector (SSD)

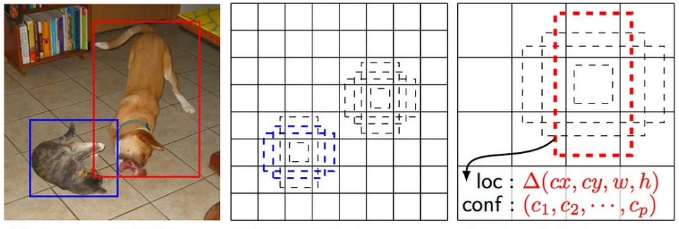
Innovations

- Applies various feature map grid cell sizes (e.g. 8x8, 6x6, ..., 1x1) to detect objects of different sizes [image pyramid]



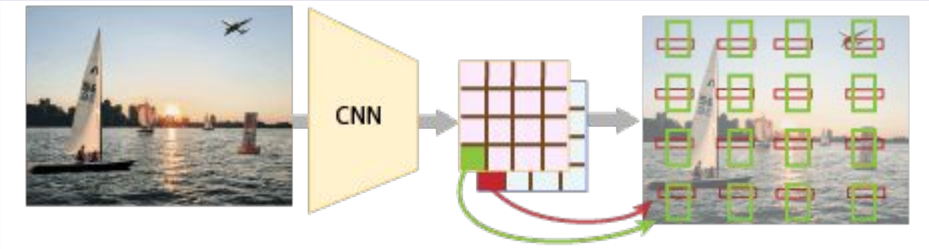
- Anchor Boxes

This is all done in the SSD Head network!


loc : $\Delta(cx, cy, w, h)$
conf : $(c_1, c_2, \cdots, c_p)$

(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

SSD default boxes at 8×8 and 4×4 feature maps



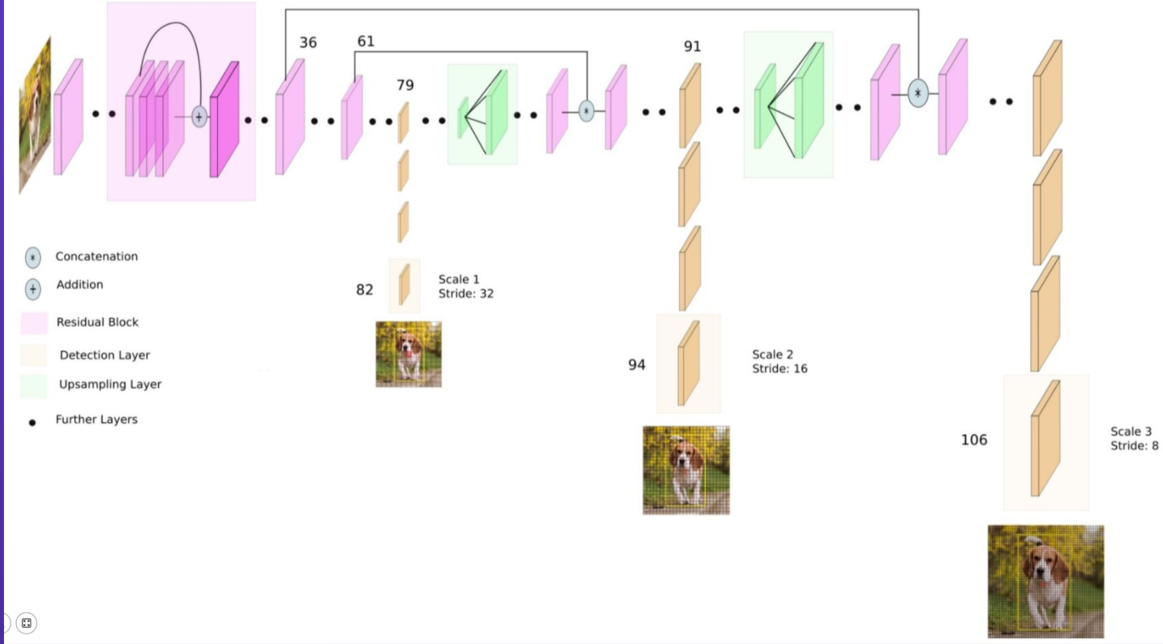
CNNoutput(i,j) maps to Image(i,j)

You Only Look Once (YOLO) v3 Architecture

Backbone model:
Pre-trained CNN (Darknet)

YOLO v3 unique customization:

- Adapted ResNet-style residual blocks.
- Upsampling and concatenation of feature layers with earlier feature layers which preserve fine-grained features
- Three scales for detection



Latest version: YOLO v8

YOLO v8 Architecture

High-level view of the architecture and its components.

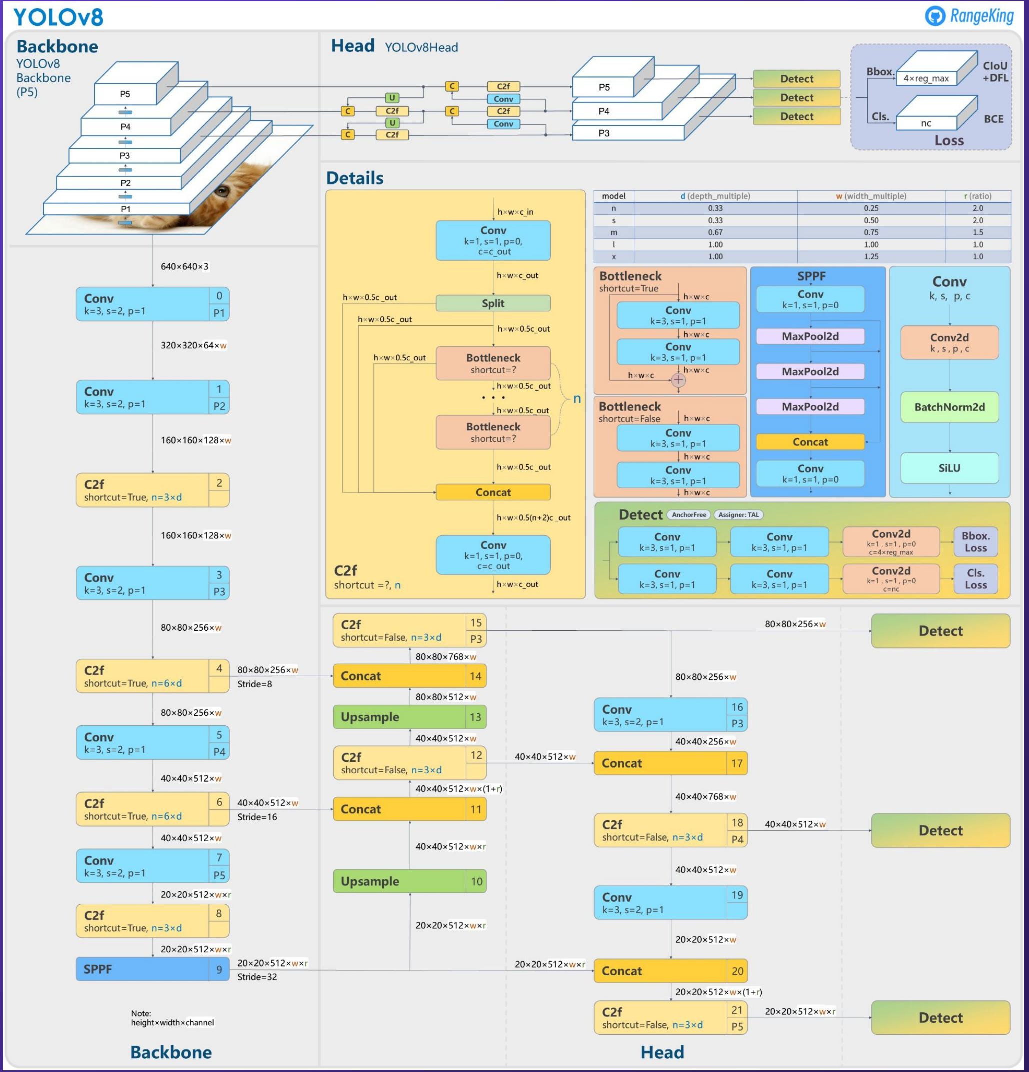
Backbone

- with Feature Pyramid (downscale through Conv from 640x640 to 20x20)

Head

- With C2F (bottleneck, SPPF) (Pooling/Reduce Dimension) and CoV[SiLU] (Activation)
- Bounding Box (without Anchor) and Classifications

- Latest version: YOLO v8
- Adapted from Brief summary of YOLOv8 model structure · GitHub

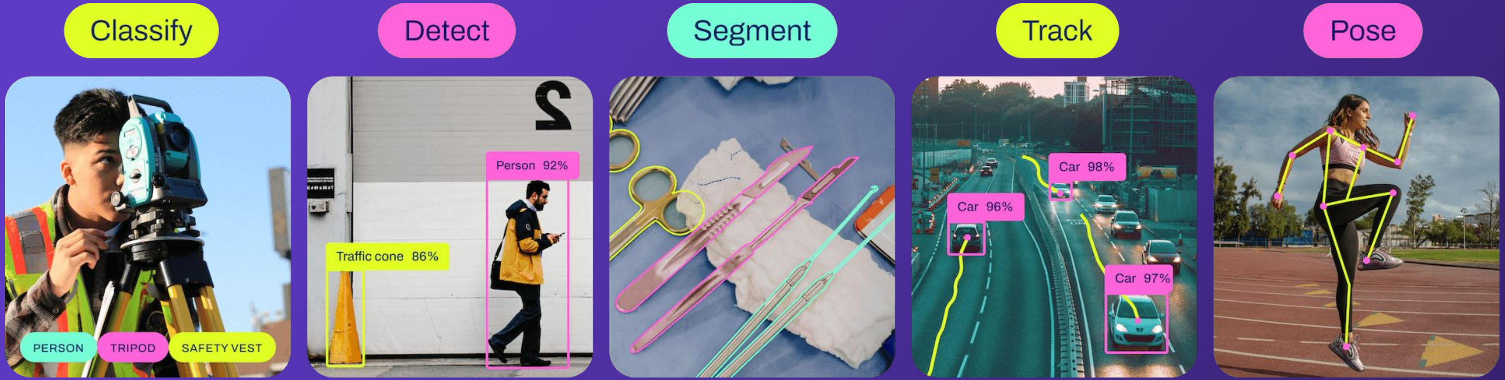


model	d (depth_multiple)	w (width_multiple)	r (ratio)
n	0.33	0.25	2.0
s	0.33	0.50	2.0
m	0.67	0.75	1.5
l	1.00	1.00	1.0
x	1.00	1.25	1.0

YOLO v8 Architecture

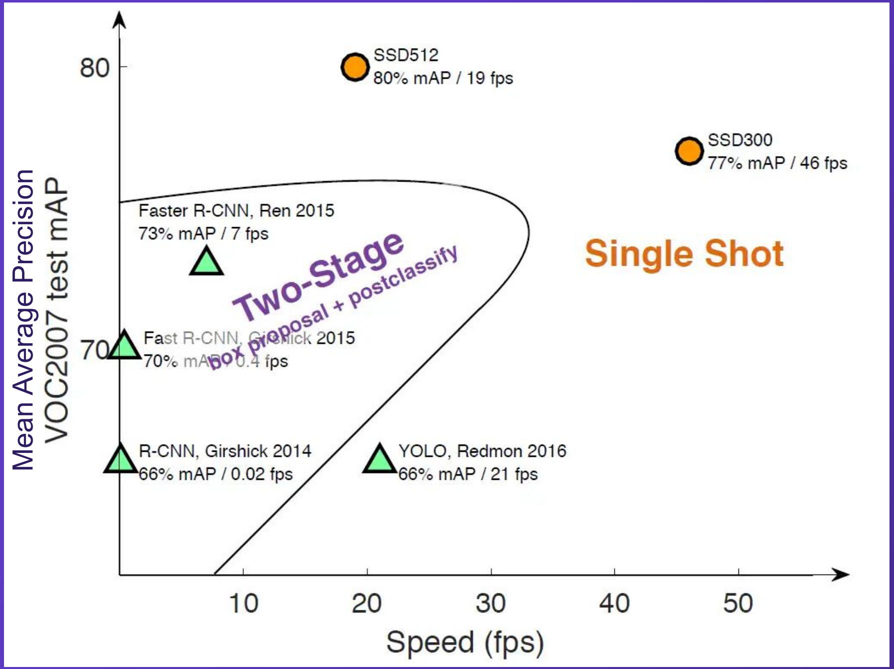
YOLOv8 has also integrated other submodules

- Classify models pretrained on the ImageNet dataset.
- Detect, Segment and Pose models pretrained on the COCO dataset (with track mode).



- Latest version: YOLO v8
- Adapted from Brief summary of YOLOv8 model structure · GitHub

Comparison Between One Stage and two-Stage Object detection



Single Stage Detectors (SSD)

- Focus on speed
- Harder on training

Two Stage Detectors

- Better precision
- More flexible

However, It is worthwhile to also note that SSD such as YOLO has undergone quite a bit of developments since.

Image adapted from computer vision笔记:SSD和DSSD | 高深远的博客 (gsy00517.github.io)