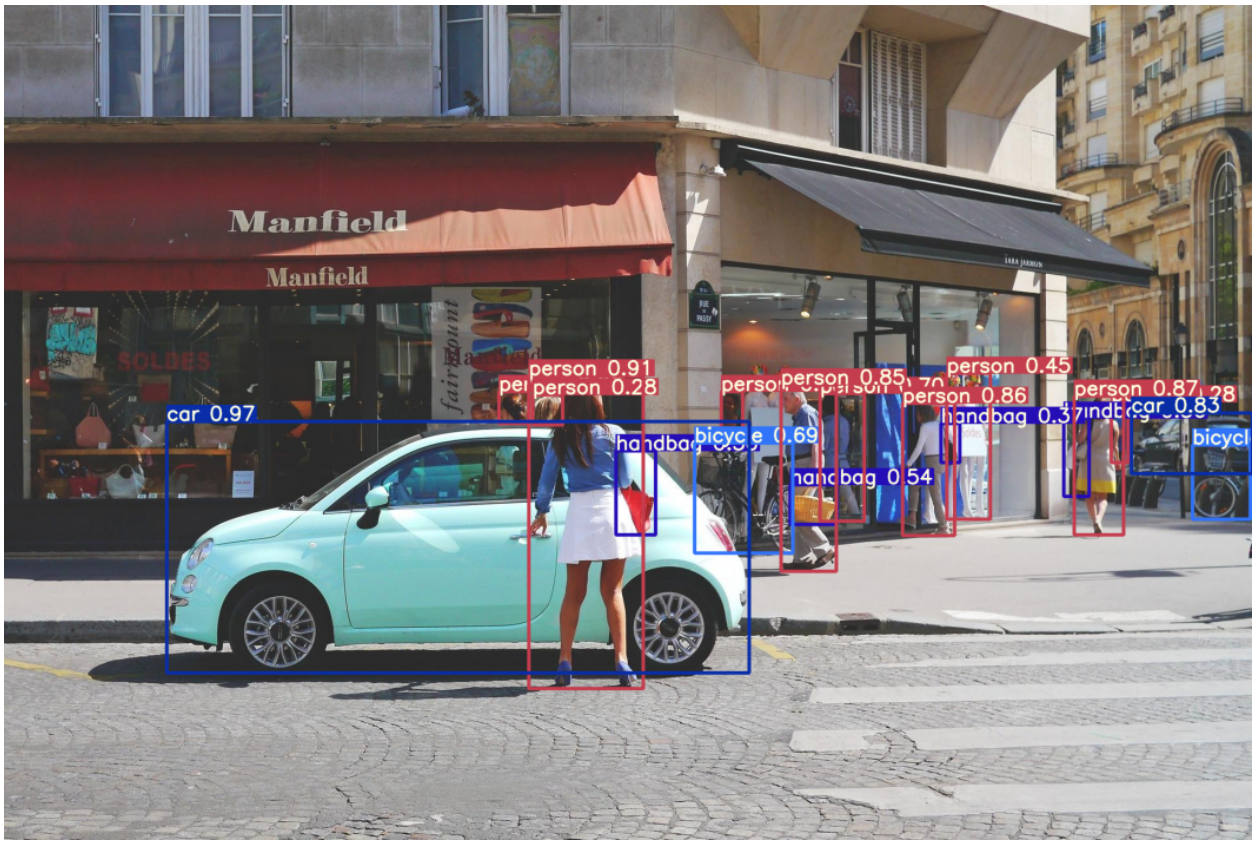
Neural Network Architecture

According to the research paper from Cornell University (https://arxiv.org/abs/2207.02696) YOLOv7 surpasses all the known object detectors in terms of speed and accuracy in the range from 5 FPS to 160 FPS and has the highest accuracy 56.8% AP among all known real-time object detectors with 30 FPS or higher on GPU V100. YOLOv7-E6 object detector (56 FPS V100, 55.9% AP) outperforms both transformer-based detector SWIN-L Cascade-Mask R-CNN (9.2 FPS A100, 53.9% AP) by 509% in speed and 2% in accuracy, and convolutional-based detector ConvNeXt-XL Cascade-Mask R-CNN (8.6 FPS A100, 55.2% AP) by 551% in speed and 0.7% AP in accuracy, as well as YOLOv7 outperforms: YOLOR, YOLOX, Scaled-YOLOv4, YOLOv5, DETR, Deformable DETR, DINO-5scale-R50, ViT-Adapter-B and many other object detectors in speed and accuracy. Moreover, we train YOLOv7 only on MS COCO dataset from scratch without using any other datasets or pre-trained weights.



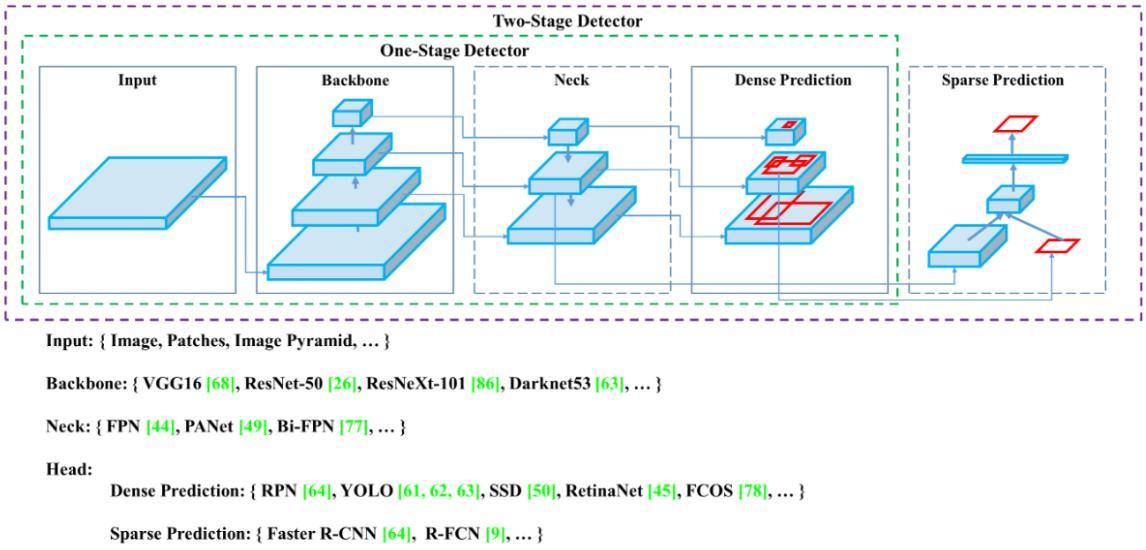


Figure 2: Object detector.

Figure 2 shows the architecture of YOLOv7 in detail. YOLOv7 2 stage detectors. First stage consists of Input, Backbone Neck and Dense Prediction parts. Let me discuss each of these parts briefly here:

1. **Input:** The input layer is nothing but the images we pass to the model as input.
2. **Backbone:** It's a Deep Neural Network composed mainly of Convolutional Neural Networks (CNN's) It's objective is to extract the main features from the input image. Backbone consists of many pre-trained Neural Networks. They are VoVNET, CSPVoVNET, ELAN, and E-LAN
3. **Neck:** Neck is used for enhanced feature detection from input images. Neck consists of FPN, RFB, PAN, etc.
4. **Dense Prediction**: Used for the prediction of objects from the output of above layers.
5. **Sparse Prediction:** This is also responsible for predictions and finding probabilities for the match with respective objects. It consists of FRCNN and RFCN.

Optimization Techniques in YOLOv7

Bag of Freebies:

Bag of freebies refers to making improvements in the model without actually increasing the training cost. The previous models of YOLO also used similar, but less advanced concepts for model optimization. It consists of 3 major steps called: Batch Normalization, Implicit Knowledge, and EMA model. To know more about them, you may read the research papers published on YOLO.

For YOLOv7, IoU threshold was increased which resulted in better detection of objects IoU refers to intersection over union.

Apart from this, our model uses unsupervised machine learning as well to detect criminal activities with greater accuracy.