Artificial Intern 2021
Deep Mind Creations Pvt Ltd
Presented by
Hem Bahadur Gurung (Yolo)
and Sachin Awal (SSD)

Object Detection: YOLO and SDD

YOLO (You only look Once)

1.Yolo was created by Joseph Redmon and Ali Farhadi in 2016.Yolo is based on Google Net.

2. Uses single convolutional network predicts the bounding boxes and the class probabilities.

3.Take an image and fragments into a S* S grid within each of the grid we take n bonding box. Usually we used 17*17 grid size.

4.The network ouputs a class probability and off values for the bounding box for each of the bounding box.

5.The bounding box which have class probability above a threshold value is selected and use to localize the object within the image. Finally, Non-Max Suppression and IOU are used to eliminate overlapping boxes.

6.Superfast 45 frames per second and accuracy 80.3%.

7. Yolo is difficulty for detection with small objects that appear in groups, such as flocks of birds. The main source of error is incorrect localizations.

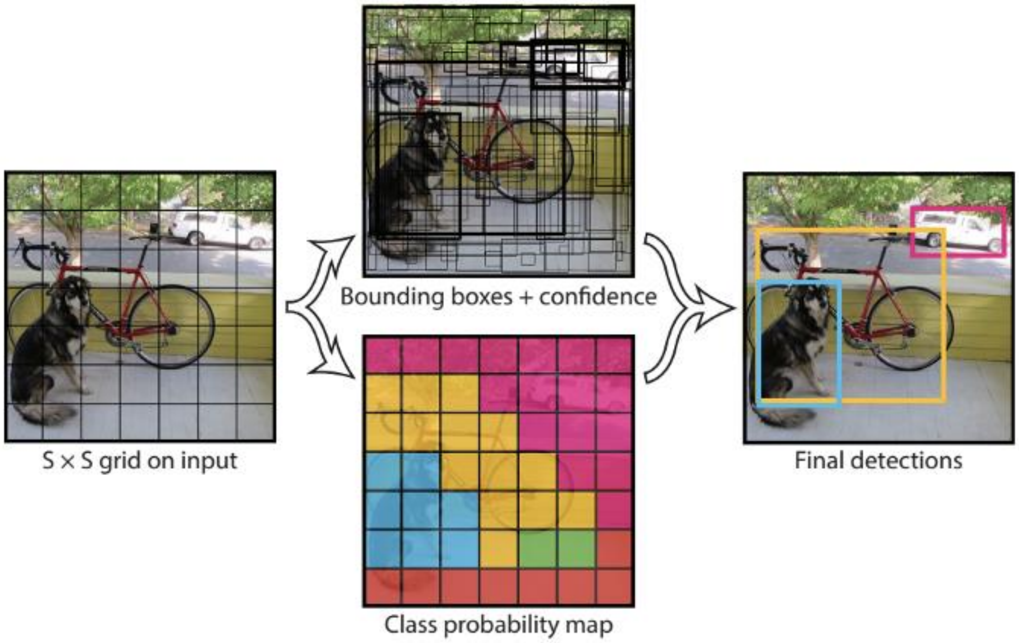

Fig: Workflow of Yolo model

The final prediction of image S×S×(5B+C) is produced by two fully connected layers over the whole conv feature map where B bounding boxes, confidence for those boxes, and C class probabilities

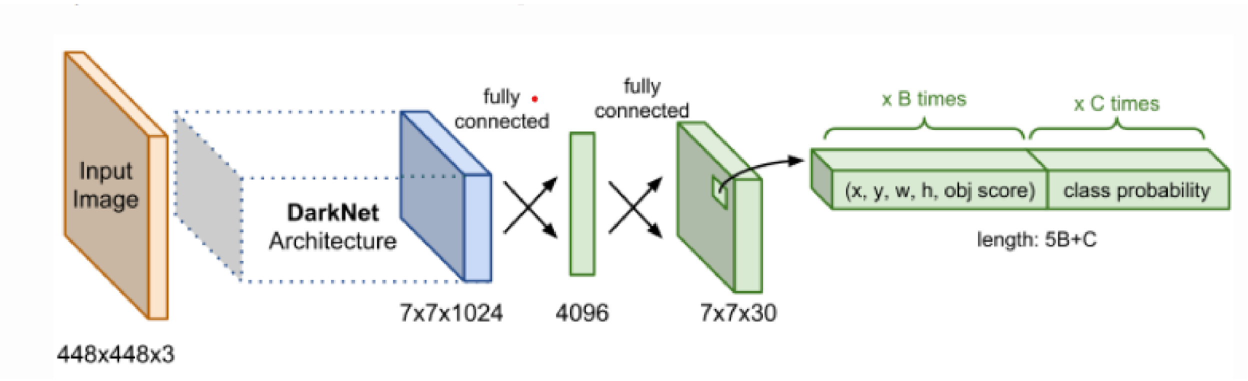

Fig: The network architecture of YOLO

Some Facts about YOLO Version

1.Yolo tiny 3 is 2/3rd times accurate and 8 times frame rate per second (FPS) than YOLOV4.

2. Fast YOLO is a fast version of YOLO, which uses 9 convolutional layers instead of 24. It is faster than YOLO but has lower mAP

3.YoloV3 uses a variant of Darknet architecture and has 53 layers training with ImageNet dataset

4.YoloV4 architecture is made of CSPDarkness53, spatial pyramid pooling, additional module PANet path aggregation neck and yolov3head.

5.Yolov3 is better and faster than SSD and worse than RetinaNet but 3.8Xfaster.

6. Yolov5 was released by a company Ultralytics in 2020.

SSD(Single Shot Detector)

• SSD work or modification of the architecture of the VGG-16.

• SSD decarded the use of the fully connected layers.

• SSD has two components: Backbone Model and SSD Model

− Backbone Model is predefined image classification network as extract feature maps.

− SSD Model is the convolution filter added to backbone model where the output are interpreted as the bounding boxes and classes of the objects in the spatial location of the final filter’s activations.

• SSD is more efficient and has a good accuracy.

• Instead of sliding window SSD divides the images using the grid where each grid cell is responsible for the detecting objects in the of region of the image.

Here, detecting object is predicting the class and location of an object with in that region. If no object is present, we consider it as background where that location is ignored. • SSD is uses non-maximum suppression to remove duplicate predictions pointing to same objects.

• Anchor box is used to detect the multiple objects in an image. It is simple boxes assigned with multiple prior boxes which are predefined and have fixed size and shape within the grid cell.

• MultiBox's loss function is combined of two critical components Confidence Loss and Location Loss.

− Confidence Loss measures how confident the network is of the object of the computed bounding box where categorical cross-entropy is used to compute this loss.

− Location Loss measures how far away the network's predicted bounding boxes are from the ground truth ones from the training set.

− FORMULA:

Multibox_loss = Confidence loss + alpha * Location_loss

Where, the term alpha helps in balancing the contribution of the location loss.