**REPORT Date: 5 June 2020**

**Real-Time Object Detection with YOLO.**

**Introduction:**

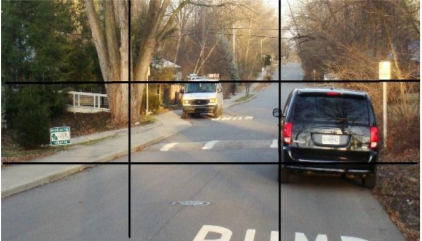
Object detection is a technology that detects the semantic objects of a class in digital images and videos. One of its real-time applications is self-driving cars. In this, our task is to detect multiple objects from an image. The most common object to detect in this application is the car, motorcycle, and pedestrian. For locating the objects in the image we use Object Localization and have to locate more than one object in real-time systems. There are various techniques for object detection, they can be split up into two categories, first is the algorithms based on Classifications. CNN and RNN come under this category. In this, we have to select the interested regions from the image and have to classify them using Convolutional Neural Network. This method is very slow because we have to run a prediction for every selected region. The second category is the algorithms based on Regressions. YOLO method comes under this category. In this, we won't select the interested regions from the image. Instead, we predict the classes and bounding boxes of the whole image at a single run of the algorithm and detect multiple objects using a single neural network. YOLO algorithm is fast as compared to other classification algorithms. In real time our algorithm process 45 frames per second. YOLO algorithm makes localization errors but predicts less false positives in the background.

**Working of YOLO Algorithm**:

First, an image is taken and YOLO algorithm is applied. In our example, the image is divided as grids of 3x3 matrixes. We can divide the image into any number grids, depending on the complexity of the image. Once the image is divided, each grid undergoes classification and localization of the object. The objectness or the confidence score of each grid is found. If there is no proper object found in the grid, then the objectness and bounding box value of the grid will be zero or if there found an object in the grid then the objectness will be 1 and the bounding box value will be its corresponding bounding values of the found object. The bounding box prediction is explained as follows. Also, Anchor boxes are used to increase the accuracy of object detection which also explained below in detail.

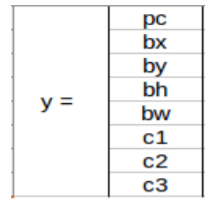
**Bounding box predictions:**

YOLO algorithm is used for predicting the accurate bounding boxes from the image. The image divides into S x S grids by predicting the bounding boxes for each grid and class probabilities. Both image classification and object localization techniques are applied for each grid of the image and each grid is assigned with a label. Then the algorithm checks each grid separately and marks the label which has an object in it and also marks its bounding boxes. The labels of the gird without object are marked as zero.



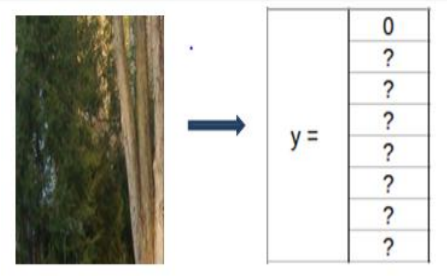
Example image with 3x3 grids

Consider the above example, an image is taken and it is divided in the form of 3 x 3 matrixes. Each grid is labelled and each grid undergoes both image classification and objects localization techniques. The label is considered as Y. Y consists of 8 values.



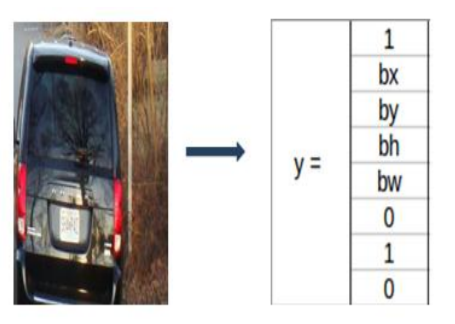
Elements of label Y

Pc – Represents whether an object is present in the grid or not. If present pc=1 else 0. bx, by, bh, bw – are the bounding boxes of the objects (if present). c1, c2, c3 – are the classes. If the object is a car then c1 and c3 will be 0 and c2 will be 1. In our example image, the first grid contains no proper object. So it is represented as,



Bounding box and Class values of grid 1

In this grid, there exists no proper object so the pc value is 0. And rest of the values are doesn’t matter because there exist no object. So, it is represented as ?. Consider a grid with the presence of an object. Both 5th and 6th grid of the image contains an object. Let’ consider the 6th grid, it is represented as.



Bounding box and Class values of grid 6

In this table, 1 represents the presence of an object. And bx, by, bh, bw are the bounding boxes of the object in the 6th grid. And the object in that grid is a car so the classes are (0,1,0). The matrix form of Y in this is Y=3x3x8. For the 5th grid also the matrix will be little similar with different bounding boxes by depending on the objects position in the corresponding grid. If two or more grids contain the same object then the center point of the object is found and the grid which has that point is taken. For this, to get the accurate detection of the object we can use to methods. They are Intersection over Union and Non-Max Suppression. In IoU, it will takes the actual and predicted bounding box value and calculates the IoU of two boxes by using the formulae, IoU = Area of Intersection / Area of Union. If the value of IoU is more than or equal to our threshold value (0.5) then it's a good prediction. The threshold value is just an assuming value. We can also take greater threshold value to increase the accuracy or for better prediction of the object. The other method is Non-max suppression, in this, the high probability boxes are taken and the boxes with high IoU are suppressed. Repeat this until a box is selected and consider that as the bounding box for that object.

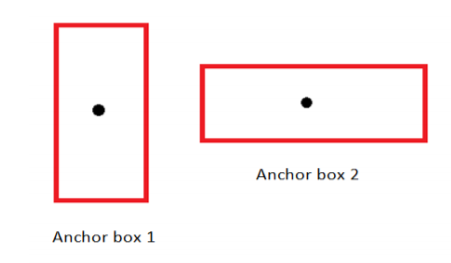
**Anchor Box**:

By using Bounding boxes for object detection, only one object can be identified by a grid. So, for detecting more than one object we go for Anchor box.



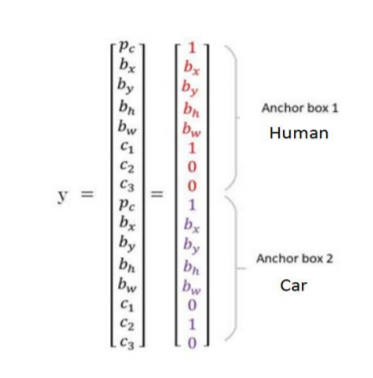
An example image for anchor box

Consider the above picture, in that both the human and the car’s midpoint come under the same grid cell. For this case, we use the anchor box method. The red color grid cells are the two anchor boxes for those objects. Any number of anchor boxes can be used for a single image to detect multiple objects. In our case, we have taken two anchor boxes.



Anchor boxes

The above figure represents the anchor box of the image we considered. The vertical anchor box is for the human and the horizontal one is the anchor box of the car. In this type of overlapping object detection, the label Y contains 16 values i.e, the values of both anchor boxes

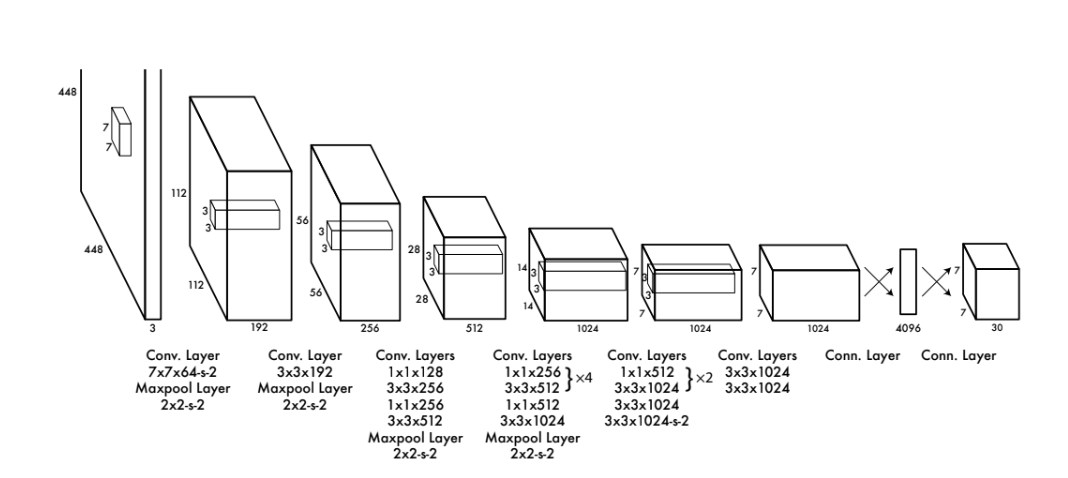


Anchor box prediction values.

Pc in both the anchor box represents the presence of the object. bx, by, bh, bw in both the anchor box represents their corresponding bounding box values. The value of the class in anchor box 1 is (1, 0, 0) because the detected object is a human. In the case of anchor box 2, the detected object is a car so the class value is (0, 1, 0). In this case, the matrix form of Y will be Y= 3x3x16 or Y= 3x3x2x8. Because of two anchor box, it is 2x8.

**RESULTS**:

The idea of YOLO is to make a Convolutional neural network to predict a (7, 7, 30) tensor. It uses a Convolutional neural network to scale back the spatial dimension to 7x7 with 1024 output channels at every location. By using two fully connected layers it performs a linear regression to create a 7x7x2 bounding box prediction. Finally, a prediction is made by considering the high confidence score of a box.

CNN Network Design.