**Aero2Astro**

**YOLO V1**

**Report**

**By**

**Anshul Singh**

**Research intern-Inspect**

# How is YOLO different?

YOLO is different from all these methods as it treats the problem of image detection as a regression problem rather than a classification problem and supports a single convolutional neural network to perform all the tasks.

Benefits:

1. Speed
2. Less background mistake
3. Highly generalizable

# Network Design

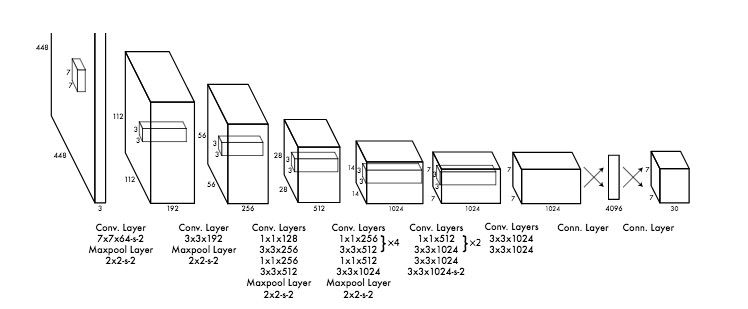
YOLO is implemented as a convolution neural network and has been evaluated on the PASCAL VOC detection dataset.

It consists of a total of 24 convolutional layers followed by 2 fully connected layers.

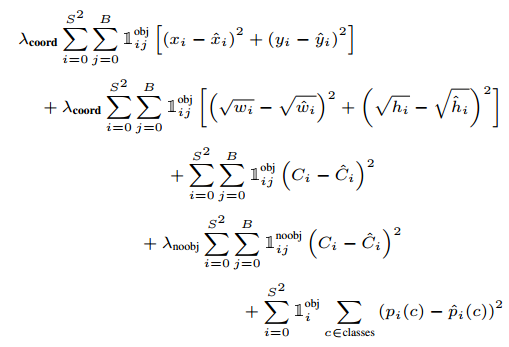
* First 20 convolutional layers followed by an average pooling layer and a fully connected layer is pre-trained on the ImageNet 1000-class classification dataset
* The pretraining for classification is performed on dataset with resolution 224 x 224
* The layers comprise of 1x1 reduction layers and 3x3 convolutional layers
* Last 4 convolutional layers followed by 2 fully connected layers are added to train the network for object detection
* Object detection requires more granular detail hence the resolution of the dataset is bumped to 448 x 448
* The final layer predicts the class probabilities and bounding boxes.

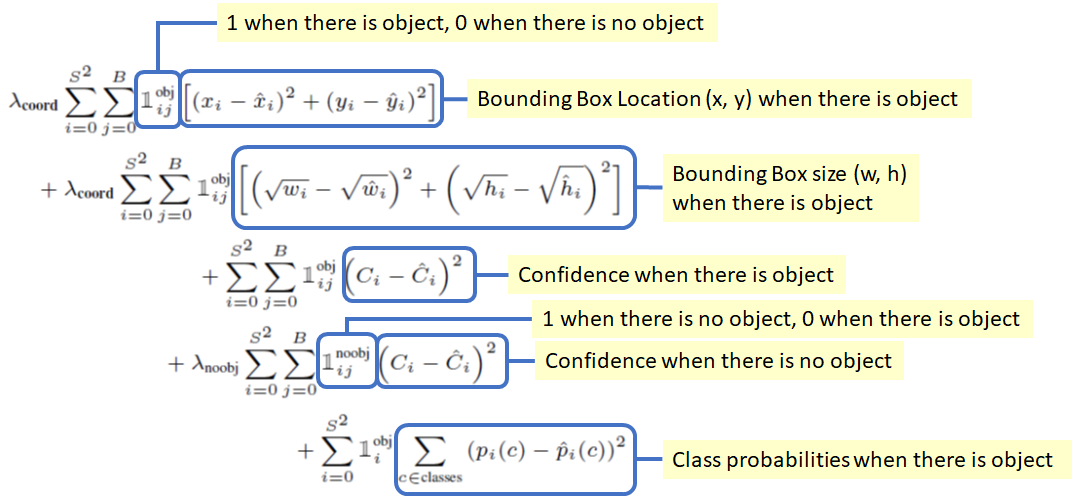
**Each bounding box consists of 5 predictions: x, y, w, h, and confidence.**

* The (x, y) coordinates represent the center of the box relative to the bounds of the grid cell.
* The width w and height h are predicted relative to the whole image.
* The confidence represents the Intersection Over Union (IOU) between the predicted box and any ground truth box.



# Loss Function





There are **5 terms in the loss function** as shown above.

1. **1st term (x, y)**: The bounding box x and y coordinates is parameterized to be offsets of a particular grid cell location so they are also bounded between 0 and 1. And the sum of square error (SSE) is estimated only when there is object.
2. **2nd term (w, h)**: The bounding box width and height are normalized by the image width and height so that they fall between 0 and 1. SSE is estimated only when there is object. Since small deviations in large boxes matter less than in small boxes. square root of the bounding box width w and height h instead of the width and height directly to partially address this problem.
3. **3rd term and 4th term (The confidence)**(i.e. the IOU between the predicted box and any ground truth box): In every image many grid cells do not contain any object. This pushes the “confidence” scores of those cells towards zero, often overpowering the gradient from cells that do contain objects, and makes the model unstable. Thus, the loss from confidence predictions for boxes that don’t contain objects, is decreased, i.e. λnoobj=0.5.
4. **5th term (Class Probabilities):** SSE of class probabilities when there is objects.
5. **λcoord:** Due to the same reason mentioned in 3rd and 4th terms, λcoord = 5 to increase the loss from bounding box coordinate predictions.

# Limitations

1. A cell can only detect one object
2. It finds it difficult to localize small objects or groups of small objects.
3. The model samples down the input image to an SxS grid where every grid cell is responsible for making bounding box predictions. Thus, due to the downsampling the model uses rather coarse features to predict the bounding boxes.