IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Object Detection Using YOLO V3

Kasham Malini[1*], Manshika Nalla, Gattu Shivani[3], Sarikonda Sree Hari Raju[4]

[1,2,3]*B.Tech. Student, Department of Computer Science and Engineering, Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India*

[4]*Associate Professor, Department of Computer Science and Engineering, Nalla Narasimha Reddy Education Society's Group of Institutions, Hyderabad, India*

***Abstract*: In the field of object detection, recently, tremendous success is achieved, but still it is a very challenging task to detect and identify objects accurately with fast speed. Human beings can detect and recognize multiple objects in images or videos with ease regardless of the object's appearance, but for computers it is challenging to identify and distinguish between things. In this paper, a modified YOLOv1 based neural network is proposed for object detection. The new neural network model has been improved in the following ways. Firstly, modification is made to the loss function of the YOLOv1 network. The improved model replaces the margin style with proportion style. Compared to the old loss function, the new is more flexible and more reasonable in optimizing the network error. Secondly, a spatial pyramid pooling layer is added; thirdly, an inception model with a convolution kernel of 1 1 is added, which reduced the number of weight parameters of the layers. Extensive experiments on Pascal VOC datasets 2007/2012 showed that the proposed method achieved better performance**

***Keywords*: YOLO, CNN, image classification, image identification ,image tracking, bounding boxes, object identification.**

1. Introduction

In computer vision, object detection refers to finding and identifying an object in an image or video. The main steps involved in object detection include feature extraction [1], feature processing [2–4], and object classification [5]. Object detection achieved excellent performance with many traditional methods that can be described from the following four aspects: bottom feature extraction, feature coding, feature aggregation, and classification. The feature extraction plays an essential role in the object detection and recognition process [6]. There will be more redundant information which can be modelled to achieve better performance than previous point-of-interest detection. Previously used scale-invariant feature transformations (SIFT) [7] and histogram of oriented gradients (HOG) [8] belong to this category. The object detection is critical in different applications, such as surveillance, cancer detection, vehicle detection, and underwater object detection. Various techniques have been used to detect the object accurately and efficiently for different applications. However, these proposed methods still have problems with a lack of accuracy and efficiency. To tackle these problems of the object detection, machine learning and deep neural network methods are more effective in correcting object detection. Traditional visual surveillance system uses human to monitor cameras for detecting any unpleasant events. If more number of cameras are to be monitored more number of man power is required which puts a limitation on the manpower use in visual surveillance. Hence the surveillance system becomes weaker. To have a better solution for this problem researchers work on Automated Visual Surveillance System which detects events requiring attention as it happens, and take action immediately. Most of the automated visual surveillance system used to detect people and/or vehicle movement in the area of interest. Thus, in this study, a modified new network is proposed based on the YOLOv1 [9] network model. The performance of the modified YOLOv1 is improved through the following points:

(i). The inception model structure is added.

(ii). A spatial pyramid pooling layer is used.

(iii) The loss function of the YOLOv1 network is optimized.

(iv)The proposed model effectively extracts features from images, performing much better in object detection.

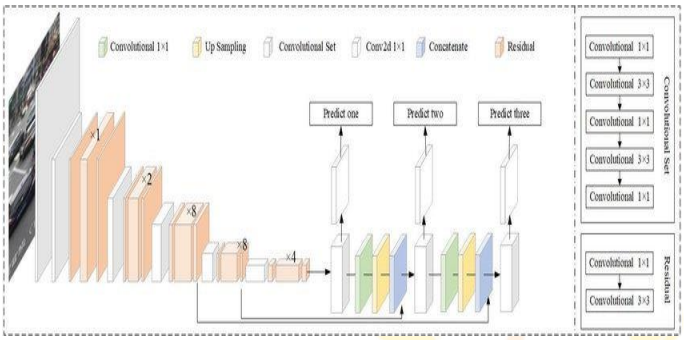
2. Algorithm and Technique

A. YOLO V3

YOLOv3 (You Only Look Once, Version 3) is a real-time object detection algorithm that identifies specific objects in videos, live feeds, or images. The YOLO machine learning algorithm uses features learned by a deep convolutional neural network to detect an object. Versions 1-3 of YOLO were created by Joseph Redmon and Ali Farhadi, and the third version of the YOLO machine learning algorithm is a more accurate version of the original ML algorithm. The first version of YOLO was created in 2016, and version 3, which is discussed extensively in this article, was made two years later in 2018. YOLOv3 is an improved version of YOLO and YOLOv2. YOLO is implemented using the Keras or How does YOLOv3 work?

(Overview) YOLO is a Convolutional Neural Network (CNN) for performing object detection in real-time. CNNs are classifier-based systems that can process input images as structured arrays of data and recognize patterns between them (view image below). YOLO has the advantage of being much faster than other networks and still maintains accuracy. It allows the model to look at the whole image at test time, so its predictions are informed by the global context in the image.

YOLO and other convolutional neural network algorithms "score" regions based on their similarities to predefined classes. High-scoring regions are noted as positive detections of whatever class they most closely identify with. For example, in a live feed of traffic, YOLO can be used to detect different kinds of vehicles depending on which regions of the video score highly in comparison to predefined classes of vehicles

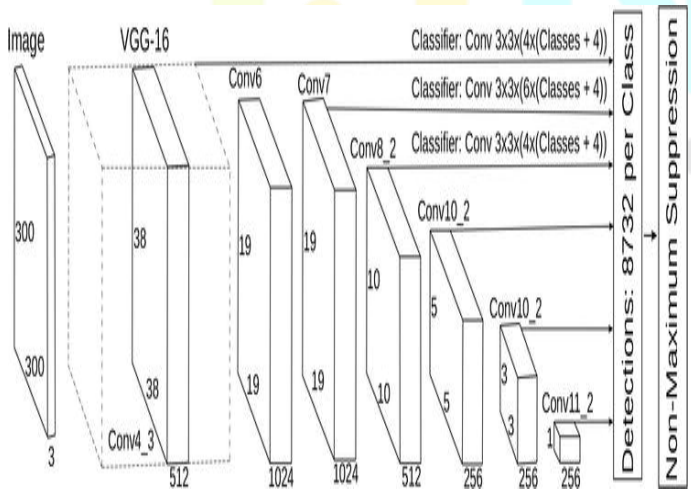


3. Implementation

The single shot multibox detector is one of the best detectors in terms of speed and accuracy comprising two main steps, feature map extraction and convolutional filter applications, to detect objects.

The SSD architecture builds on the VGG-1616 network and this choice was made based on the strong performance in high-quality image classification tasks and the popularity of the network in problems where transfer learning is involved. Instead of the original VGG fully connected layers, a set of auxiliary convolutional layers change the model, thus enabling to extract features at multiple scales and progressively decrease the size of the input to each subsequent layer.

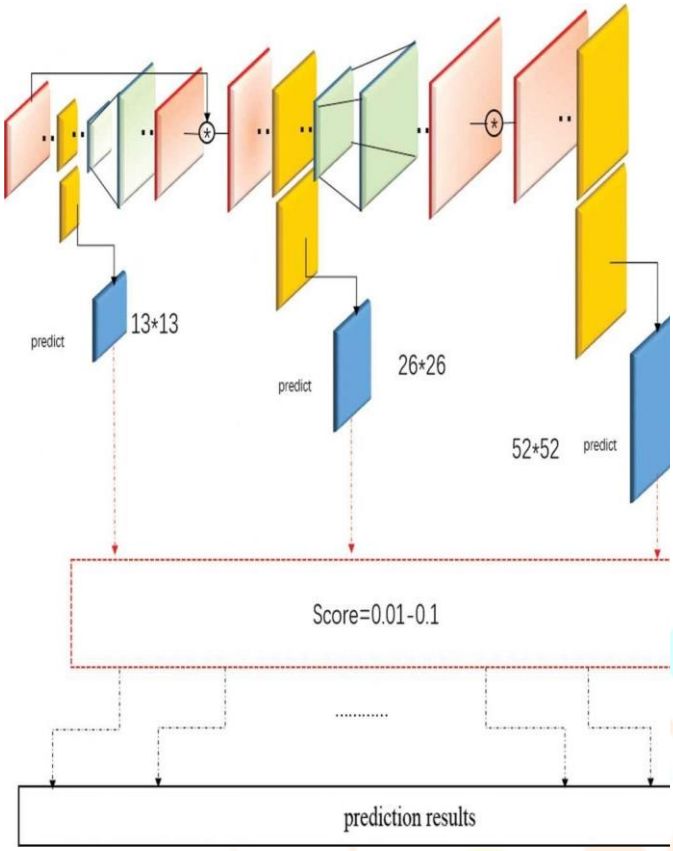
The bounding box generation considers the application of matching pre-computed, fixed-size bounding boxes called *priors* with the original distribution of ground truth boxes. These *priors* are selected to keep the intersection over union (IoU) ratio equal to or greater than 0.50.5.

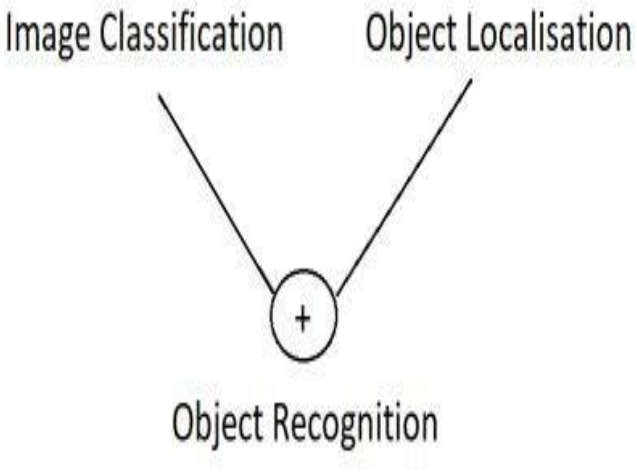


YOLO V3

The YOLOv3 algorithm first separates an image into a grid. Each grid cell predicts some number of boundary boxes (sometimes referred to as anchor boxes) around objects that score highly with the aforementioned predefined classes. Each boundary box has a respective confidence score of how accurate it assumes that prediction should be and detects only one object per bounding box. The boundary boxes are generated by clustering the dimensions of the ground truth boxes from the original dataset to find the most common shapes and sizes. Other comparable algorithms that can carry out the same objective are R-CNN (Region-based Convolutional Neural Networks made in 2015) and Fast R-CNN (R-CNN improvement developed in 2017), and Mask R-CNN. However, unlike systems like R-CNN and Fast R-CNN, YOLO is trained to do classification and bounding box regression at the same time. Update: Check out our article about the new YOLOv7 model, which is widely expected to become the new industry standard for object detection. There are major differences between YOLOv3 and older versions occur in terms of speed, precision, and specificity of classes. YOLOv2 and YOLOv3 are worlds apart in terms of accuracy, speed, and architecture. YOLOv2 came out in 2016, two years before YOLO v3. The following sections will give you an overview of what's new in YOLOv3. Speed YOLOv2 was using Darknet-19 as its backbone feature extractor, while YOLOv3 now uses Darknet-53. Darknet-53 is a backbone also made by the YOLO creators Joseph Redmon and Ali Farhadi. Darknet-53 has 53 convolutional layers instead of the previous 19, making it more powerful than Darknet-19 and more efficient than competing backbones (ResNet-101 or ResNet-152.)

Backbone	Top-1	Top-5	Ops	BFLOP/s	FPS
Darknet-19	74.1	91.8	7.29	1246	**171**
ResNet-101	77.1	93.7	19.7	1039	53
ResNet-152	**77.6**	**93.8**	29.4	1090	37
Darknet-53	77.2	**93.8**	18.7	**1457**	78





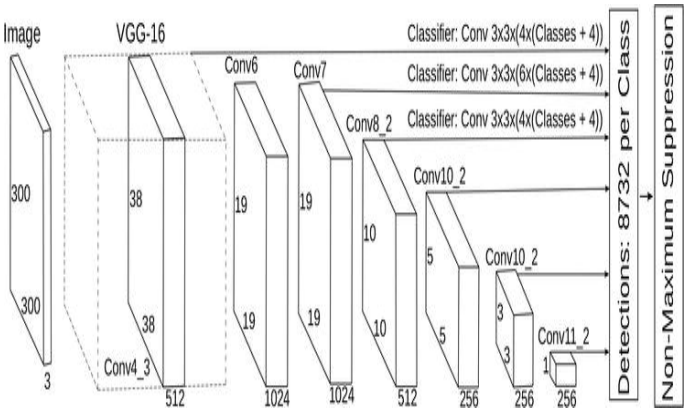
While the image classification problem focuses on classifying the images, in 1 image there may be more than 1 class we are searching for and in object recognition, our task is to find all of them placed in the most appropriate **boxes**. we try to recognize the objects, we use bounding boxes in which the object will be possibly detected. We will learn how to obtain as possible as close boxes to the object detected later.

As you may notice, object recognition is a bit more complex task than image classification where we try to localize and recognise the object in images

YOLO gives us the maximum flexibility, it gives us the bounding box and the confidence score which directly refers to the accuracy of the recognized object from the image

1.Selective search for image segmentation: image segmentation without using AI-based methods but using Classical Computer Vision-Based methods
·first the similarities between all neighbouring regions are calculated. The two most similar regions are grouped together, and new similarities are calculated between the resulting region and its neighbours. This process is then repeated until the whole object is covered in a single region

2.Classification with SVM and Bounding Box Prediction
Finally, using SVM (support vector machine) for classification and a bounding box regressor, the model gives us the final bounding boxes along with detected classes where the bounding box regressor's task is just to improve the proposed box to encircle the object better



You only look once is a state-of-the-art object detection algorithm which targets real-time applications, and unlike some of the competitors, it is not a traditional classifier purposed as an object detector. YOLO works by dividing the input image into a grid of S×SS×S cells, where each of these cells is responsible for five bounding boxes predictions that describe the rectangle around the object. It also outputs a confidence score, which is a measure of the certainty that an object was enclosed. Therefore the score does not have any relation with the kind of object present in the box, only with the box's shape.

For each predicted bounding box, a class it's also predicted working just like a regular classifier giving resulting in a probability distribution over all the possible classes. The confidence score for the bounding box and the class prediction combines into one final score that specifies the probability for each box includes a specific type of object. Given these design choices, most of the boxes will have low confidence scores, so only the boxes whose final score is beyond a threshold are kept.

2.CNN (Convolutional Neural Networks)

This is the main methodology of CNN

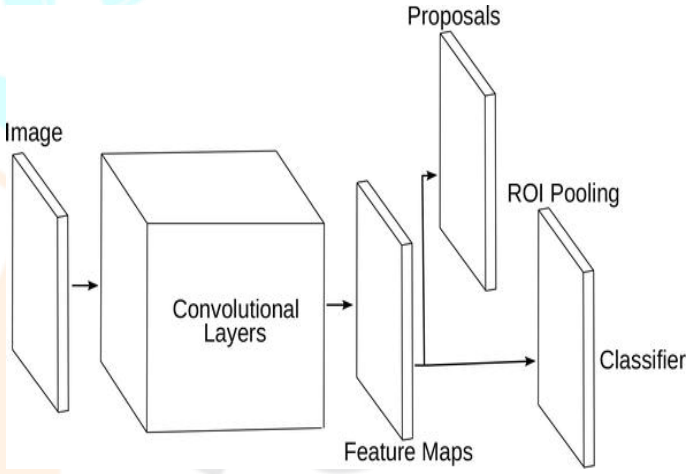
- Selective Search is already a complex algorithm and using this only for the first step increase the model computational cost too much. -> SLOW
- Obtaining 2000 regions to apply feature extraction 1 by 1 is too much computational again! -> SLOW

which gives us 47 seconds for 1 image detection, therefore it's not possible to use this model for real-time object detection tasksIt's not an end-to-end trainable model since Selective Search Algorithm is not a trainable method which makes it impossible to develop region proposals by training RCNN. -> NOT TRAINABLE for some partUsing SVM is another reason for not having an end-to-end architecture where we need to train SVM and CNN separately which poses a more difficult task.As a result, although it was a state-of-art architecture back in time having better accuracies than the previous models, it's clear that the model needs to get improved for especially speed performances. For this reason, we find ourselves examining the Fast RCNN model which is an improved version of RCNN.

The faster region convolutional neural network is another state-of-the-art CNN-based deep learning object detection approach. In this architecture, the network takes the provided input image into a convolutional network which provides a convolutional feature map. Instead of using the selective search algorithm to identify the region proposals made in previous iterations, a separate network is used to learn and predict these regions. The predicted region proposals are then reshaped using a region of interest (ROI) pooling layer, which is then used to classify the image within the proposed region and predict the offset values for the bounding boxes.

The strategy behind the region proposal network (RPN) training is to use a binary label for each anchor, so the number one will represent the presence of an object and number zero the absence; with this strategy any IoU over 0.70.7 determines the object's presence and below 0.30.3 indicates the object's absence.

The picture below depicts the unified network for object detection implemented in the Faster RCNN architecture. Using the recently popular terminology of neural networks with "attention" mechanisms the region proposal network module tells the Fast RCNN module where to look



4.Applications

- Law enforcing set of government
- Identification
- Traffic analysis
- Anamoly detection
- Face detection

5.Conclusion:

Due to its powerful learning ability and advantages in dealing with occlusion, scale transformation and background switches, deep learning based object detection has been a research hotspot in recent years. This paper provides a detailed review on deep learning based object detection frameworks which handle different sub-problems, such as occlusion, clutter and low resolution, with different degrees of modifications on R-CNN. The review starts on generic object detection pipelines which provide base architectures for other related tasks. Then, three other common tasks, namely salient object detection, face detection and pedestrian detection, are also briefly reviewed. Finally, we propose several promising future directions to gain a thorough understanding of the object detection landscape. This review is also meaningful for the developments in neural networks and related learning systems, which provides valuable insights and guidelines for future progress.

6. Future Enhancements

- Pixel level detection
- Multi modal detection
- Object part detection
- Active vision

7. References

[1] P. Adarsh, P . Rathi and M . Kumar , "YOLO V3-tiny:Object Dection and Recognition using one stage improved model," 2020 6th International Conference on Advanced Computing and Communication System (ICACCS),pp.687-694,2020.

[2] D. Pestana et al., " A Full Featured Configurable Accelerator for Object Detection With YOLO" in IEEE Access, vol.9,pp.75864-75877 , 2021.

[3] A. Ćorović, V. Ilić, S. Đurić, M. Marijan and B. Pavković, "The Real-Time Detection of Traffic Participants Using YOLO Algorithm," 2018 26th Telecommunications Forum (TELFOR),2018,pp.1-4,2018.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real- time object detection," IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 2016-Decem, pp. 779–788, 2016, doi: 10.1109/CVPR.2016.91.

[5] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR, vol. 2017-Janua, pp. 6517–6525,2017, doi: 10.1109/CVPR.2017.69