Object Recognition Using YOLOv3

SHAYAN KUMAR
Signal Processing and Machine Learning
Department of Electronics and Electrical Engineering
Indian Institute of Technology Guwahati
Guwahati, India
k.shayan@iitg.ac.in

***Abstract*—Due to device memory and computing power limitations, real-time object recognition with a desired response time is a difficult task. Deep learning models needed to be modified to deal with these challenges. Based on the Darknet-53, "YOLO" is a well-lightweight network that does not significantly reduce detection accuracy. The 'You Only Look Once' v3 (YOLOv3) approach is one of the most often used deep learning-based object identification algorithms. It employs the k-means cluster method to calculate the initial width and height of the expected bounding boxes. The architecture of YOLO is presented here in order to assess its operational functionality.**

***Index Terms*—deep-learning,object-detection, resnet, darknet, convolutional-block, residual-block, intersection-over-union, anchor-box, non-max suppression.**

I. Introduction

Object recognition is a technique of computer vision which is used for identifying and localizing objects in images or videos. This is a task that combines object detection and image classification. Deep learning and machine learning algorithms provide several algorithms of object recognition. We can quickly identify people, objects, scenes, and visual elements when we look at an image or a video. The idea is to train a computer to do what people do naturally: obtain an understanding of what an image includes.

Objects can be recognized from various perspectives, including front, back, and side views. The object can also be recognized in different sizes and when it is partially obscured from view. Various object recognition tasks, such as handwriting, speech, face, license plate, lane line, ship and military objects, and underwater creatures, and so on, have received a lot of attention in recent years.

Several deep learning models, including R-CNN, Fast R-CNN, and Faster, have also been proposed to minimise training time and enhance mean average accuracy. The structure of region-based approaches is complicated, and object detection takes time. Because the detection accuracy and speed are nicely balanced, YOLOv3 may be the most common object detector in real applications. .

II. Methodology

Object identification is treated as a regression problem by the YOLOv3 approach. With a single feed forward convolution neural network, it predicts class probabilities and bounding box offsets from complete pictures. This approach improves detection time and takes images of various sizes as input. Darknet-53 is used by YOLOv3 to do feature extraction. YOLOv3 employs multi-scale prediction, which implies it is identified on feature maps of several scales. As a result, target detection accuracy has increased. Figure 1 depicts its detailed structure.

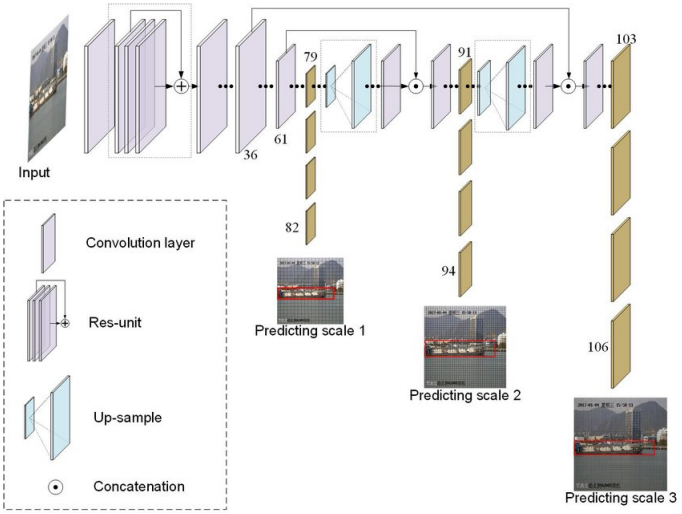

Fig. 1. Framework of YOLOv3 Neural Network

The following is the YOLOv3 testing procedure:

Step 1: Insert the image and scale it to the standard size.

Step 2: Split the input picture into 13*13, 26*26, 52* 52 grids of three scales. If an item's centre point falls within the grid unit, the grid unit forecasts the object.

Step 3: Using k-means clustering, we can compute the bounding box priors for each grid unit. Each grid unit has three clusters. There are nine clusters per grid unit due to the three scales.

Step 4: Upload the image to the network to extract the features. The model initially generates a small-scale feature map of 13*13.

Step5: The 13*13 small-scale feature map is initially subjected to a convolutional set and two times upsampling before being coupled to the 26*26 feature map and predicting the result.

Step6. The 26*26 feature map generated in step 5 is subjected to convolutional set and two times upsampling before

being coupled to the 52*52 feature map and producing the prediction result.

Step 7: Combine characteristics from three scale predictive outputs. Following that, use a likelihood score as a threshold to exclude most anchors with low scores. Then, for post-processing, use Non Maximum Suppression (NMS) leaving more accurate boxes.

III. Conclusion

The architecture and functioning functions of YOLOv3 are examined here. Bounding boxes, union-over-intersection, non-max suppression, and anchor boxes are used. Its backbone is DrakNet-53. YOLO provides a good blend of accuracy and quickness. Figure 2 shows how photos are evaluated to ensure accuracy.

Fig. 2. Object Recognition Using YOLOv3