iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 8 **Issue:** VI **Month of publication:** June 2020

DOI: http://doi.org/10.22214/ijraset.2020.6098

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.429
Volume 8 Issue VI June 2020- Available at www.ijraset.com

An Overview of You Only Look Once: Unified, Real-Time Object Detection

Deepika H C[1], Vijayashree Raghuveer shetty[2], Dr G N Srinivasan[3]
[1, 2]*Student,* [3]*Professor, Department of Information Science and Engineering, R V College of Engineering, Bengaluru*

Abstract: ***This paper aims at reviewing existing YOLO architecture, its implementation and working. You only look once, is an architecture which is a regression problem. Yolo comes in different versions such as YoloV1, YoloV2 and YoloV3.The feature extractor for Yolo is Darknet.***
The network looks at the entire image only once. In one evaluation, a single neural network predicts bounding boxes and class probabilities directly from full images. The unified architecture is extremely fast. YOLO model processes images in real-time at 45 frames per second.
Fast YOLO, an extremely fast version of Yolo, processes 155 frames per second. Yolo is better at making less localization errors as it looks at the entire image to predict objects on individual cells. This paper also provides details on the evolution and evaluation of the architecture.
Keywords: Computer vision, Object detection, Convolutional neural networks, Multiclass problem, Multilabel problem, Anchor boxes, Loss function, Bounding boxes

I. INTRODUCTION

Computer vision is a process of using machines to understand imagery, both images and video. The tasks under computer vision include the methods of acquiring, processing, analyzing and understanding the details in the images. There are many computer vision techniques like, image classification, object localization, object detection, semantic segmentation, instance segmentation etc.
Object detection is a computer vision problem that deals with identifying and locating object of certain classes in the image. Bounding boxes are used to locate the objects. RCNN, Fast-RCNN, Faster-RCNN, YOLO, Single-shot detection, etc., falls under object detection architectures.
You Only Look Once (YOLO), an object detection approach which uses Convolutional neural networks for object detection. The main advantage of YOLO is that it supports real-time object detection. YOLO looks at the entire image only once and hence the name, “you only look once”. This paper is a review of the YOLO architecture and its working.

II. BACKGROUND

The object detection techniques can fall into two categories based on Classifications and Regressions. CNN, R-CNN comes under classifications and these are generally slow because, prediction should be run for every selected region. The interested regions are selected and are classified using convolutional neural networks. The algorithm like YOLO, which falls under regression category, predicts the classes of the objects in a single run and uses single neural network to detect multiple objects in the images. Since YOLO looks at the entire image to predict objects on the individual cells, it is better at making less localization errors. The algorithm has got the ability to process 45 frames per second.
For training, YOLO requires a neural network framework, Darknet. There are three existing versions of YOLO. YoloV1, YoloV2 and YoloV3 are the version of Yolo. The very first version of Yolo has 24layers in total, with 24 convolutional layers followed by 2 fully connected layers.
The algorithm was not good at detecting tiny objects and this turns out to be the major drawback of YoloV1. In YoloV2, there are 30 layers in total with no fully connected layers. Each Conv layer is followed by a batch normalization layer and Anchor boxes are introduced at this version.
This algorithm also failed to detect tiny objects and also multiclass problem was found. YoloV3 is the latest version which uses 106 neural network layers. Considers 9 anchor boxes, 3 per class and hence more bounding boxes are predicted. Multiclass problem in this is turned into a Multilabel problem. The algorithm works well with tiny objects and this turns out to be a positive outcome.

III. EXISTING IMPLEMENTATIONS OF YOLO

This section of the paper reviews various implementations of YOLO architecture.

A. In the paper, "You only look once: unified, real time detection" [1], a new approach to object detection is presented. The architecture comes under "Regression" problem. As per the details about implementation of the architecture given in this paper, the network uses entire image to predict each bounding box. The prediction of objects belonging to multiple classes is done simultaneously. Basically, the image is divided into an *S* x *S* grid. When the center of any object falls into a grid cell, that grid cell is responsible for the detection of the object. Any number of bounding boxes (*B*) with the confidence can be predicted by each grid cell. The bounding boxes have 5 prediction values like (*x,y,w,h,c*) which are coordinates of center of the box relative to the grid cell, width and height of the predicted relative to the whole image and confidence respectively. The model is implemented as a CNN and using PASCAL VOC dataset, it is being evaluated. There are two parts, the part which extracts the features from the image is the convolutional layers while the second part is the fully connected layers which provides the output probabilities and coordinates. The network includes, 24 convolutional layers followed by 2 fully connected layers. 1 x 1 reduction layers followed by 3 x 3 convolutional layers are used. The convolutional layers are pretrained with the ImageNet 1000-class dataset. The first 20 convolutional layers are used for trained, which is followed by an average-pooling layer and a fully connected layer. For training and inference, Darknet framework is used. The paper explains about the Loss functions. The main source of error was found during localizations. Strong spatial constraints are imposed on bounding box predictions by Yolo. These spatial constraints can make the model fail to predict nearby objects. This model developed cannot detect small objects. The model struggles to generalize objects in new or unusual aspect ratios. These are the main limitations drawn from the implementation. The paper provides the comparison to other detection systems and real-time systems. Fast YOLO is the fastest general-purpose object detector. Yolo is compatible which is ideal for any sort of applications and also, it is fast, robust in nature.

B. The paper, "Real-time object detection with Yolo" [2] presents the Object detection using Yolo approach. The paper initially briefs about Yolo algorithm and states Yolo is a regression problem. The method used has several advantages over other detection algorithms. Unlike the other detection algorithms, YOLO looks at the image completely at once and predicts the bounding boxes using CNN and class probabilities for the bounding boxes and the detection in the images happens way faster than the other algorithms. The paper describes the working of Yolo. For an input image, the algorithm is applied. The image is divided into number of grids. Each grid undergoes the process of classification and localization. For each grid, there is confidence score which is found. If there is no object found in the grid, then the objectness score and the value of the bounding box will be zero else the value will be 1 with corresponding values for bounding boxes. The bounding box prediction is being explained in the paper and also it shows the use of anchor boxes which is used to increase the accuracy of object detection. Each grid in the image undergoes classification and localization techniques. Each grid of the image is labeled. The label is a vector of different values in which the first value tells about the presence of object, next four are the bounding box details and the last values tells to which class, the object belongs to. The methods like, Intersection over union and non-maximum suppression helps when two or more grids contain the same object. It is called a good prediction when the IOU value is above the chosen threshold value. Greater the threshold value better is the accuracy. In Non-maximum suppression, boxes with high probability values are selected high probability boxes are taken and the ones with higher IOU values are suppressed. The types of loss functions are pointed out at the end of the paper. This algorithm compared to others is easy to build and can be trained on full images. And this is the best algorithm for real-time object detection.

C. The paper, "Evaluation and Evolution of Object Detection Techniques YOLO and R-CNN"[3], describes the algorithms and CNN family and evolution and improvements of YOLO. In the CNN family, CNN algorithm is explained at first which includes 3 layers like, convolutional layer, pooling layer and fully connected layer. Each layer has been defined and their functions are briefed out. According to the paper, CNN are for understanding the input features. Pooling layers are present between the 2 convolutional layers and are used to extract features and to reduce the dimensions of feature maps. Pooling layers can be Average or Max. Fully connected layers are present at the last part of the network mainly used for classification. RCNN is an architecture which is composed of 3 modules such as, module for region proposals production, module containing number of Convolutional layers and the last module comprises of SVM classifiers. The main difference between RCNN and Fast RCNN is that in Fast RCNN the region proposals are input at feature map level while in RCNN it is at pixel level. The advantage of Faster RCNN is that the computation time is very less due to the lesser feature map resolution than that of the original image. The comparison between the speed of the algorithms is listed out which states, it is the Faster RCNN which is

highly fast with test time per image as 0.2seconds. The YoloV1 is inspired by a network model GoogleNet which was used for image classification and this includes 24 convolutional layers and 2 fully connected layers. After fine-tuning of the network, the classification network of high resultion provides 4% increase in mAP. Anchor boxes are used in the prediction of bounding boxes. The pooling layers are removed. And the functioning and architecture of YoloV3 is similar to that given in paper [1] and [2]. The paper states that the heart of the Yolo algorithm is Loss function on which it is trained. Fast Yolo is the fastest version of Yolo. YOLOv3 using logistic regression predicts the boxes at 3 different levels. Moreover, applications that demand speed, robust object detection can depend on YOLO because YOLO generalizes representation of object other than models. These prominent points make YOLO a strongly recommended and widely spoken detection system.

D. This paper, "Understanding of Object Detection Based on CNN Family and YOLO", briefly discusses about the current algorithms in object detection, including the CNN family and YOLO. YOLO has more advanced application in practice compared to algorithms in CNN. It is unified object detection model. It's easy to construct and can be trained directly on full images. Loss function is the main part of YOLO, it is trained on a loss function that directly corresponds to detection performance and the entire model is trained jointly. YOLOv2 provides state-of-the-art the best tradeoff between real-time speed and excellent accuracy for object detection than other detection systems across a variety of detection datasets. These supporting advantages make it worthy of being strongly recommended and popularized. Except the structure of each algorithm, the scope of the dataset is the up-coming challenge for machine learning. The huge dataset is what is required and the important ones in the development of a model and acquiring good accuracy and to achieve an idea results.

E. In this paper," Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video",[5] Fast YOLO has been introduced which is a new architecture to detect objects in real time videos. Although YOLOv2 is very faster compared to others with real time inference on powerful GPUs, it is not possible to use it on embedded devices in real-time. Here, in this paper, with the advantage of the evolutionary deep intelligence framework, it is used to produce an optimized network architecture based on YOLOv2. This optimized network architecture is utilized within a motion-adaptive inference framework to speed up the detection process as well as reduce the energy consumption of the embedded device. Experimental results showed that the proposed Fast YOLO framework can achieve an average run-time that is ~3.3X faster compared to original YOLOv2, can reduce the number of deep inferences by an average of 38.13%, and possesses a network architecture that is ~2.8X more compact.

IV. CONCLUSION

From the reviews of different papers, it is understood that YOLO is the best and recommendable architecture for any sort of real-time object detections. YOLO is the better choice when one requires fast processing. Yolo is simple to implement, easy to understand and is robust. Since it generalizes well, it can be used in any applications. This review also gives the speed comparison with other CNNs. The first two paper briefs out the implementation methodologies. The papers later explain about the evolution of and Evaluation of Yolo. And after all the discussions, it is evident that using Yolo is the best solution to real-time object detections due to its speed and accuracy improvements with different techniques.

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788 (2016)

[2] Geethapriya. S, N. Duraimurugan, S.P. Chokkalingam, "Real-time object detection with Yolo", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-8, Issue-3S, February 2019

[3] K G Shreyas Dixit, Mahima Girish Chadaga, Sinchana S Savalgimath, G Ragavendra Rakshith, Naveen Kumar M R, "Evaluation and Evolution of Object Detection Techniques YOLO and R-CNN", International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S3, July 2019

[4] Juan du, "Understanding of Object Detection Based on CNN Family and YOLO", IOP Conf. Series: Journal of Physics: Conf. Series 1004 012029 doi:10.1088/1742-6596/1004/1/012029 (2018)

[5] Mohammad Javad Shafiee, Brendan Chywl ,Francis Li, Alexander Wong, "Fast YOLO: A Fast You Only Look Once System for Real-time Embedded Object Detection in Video", arXiv:1709.05943 (2019)

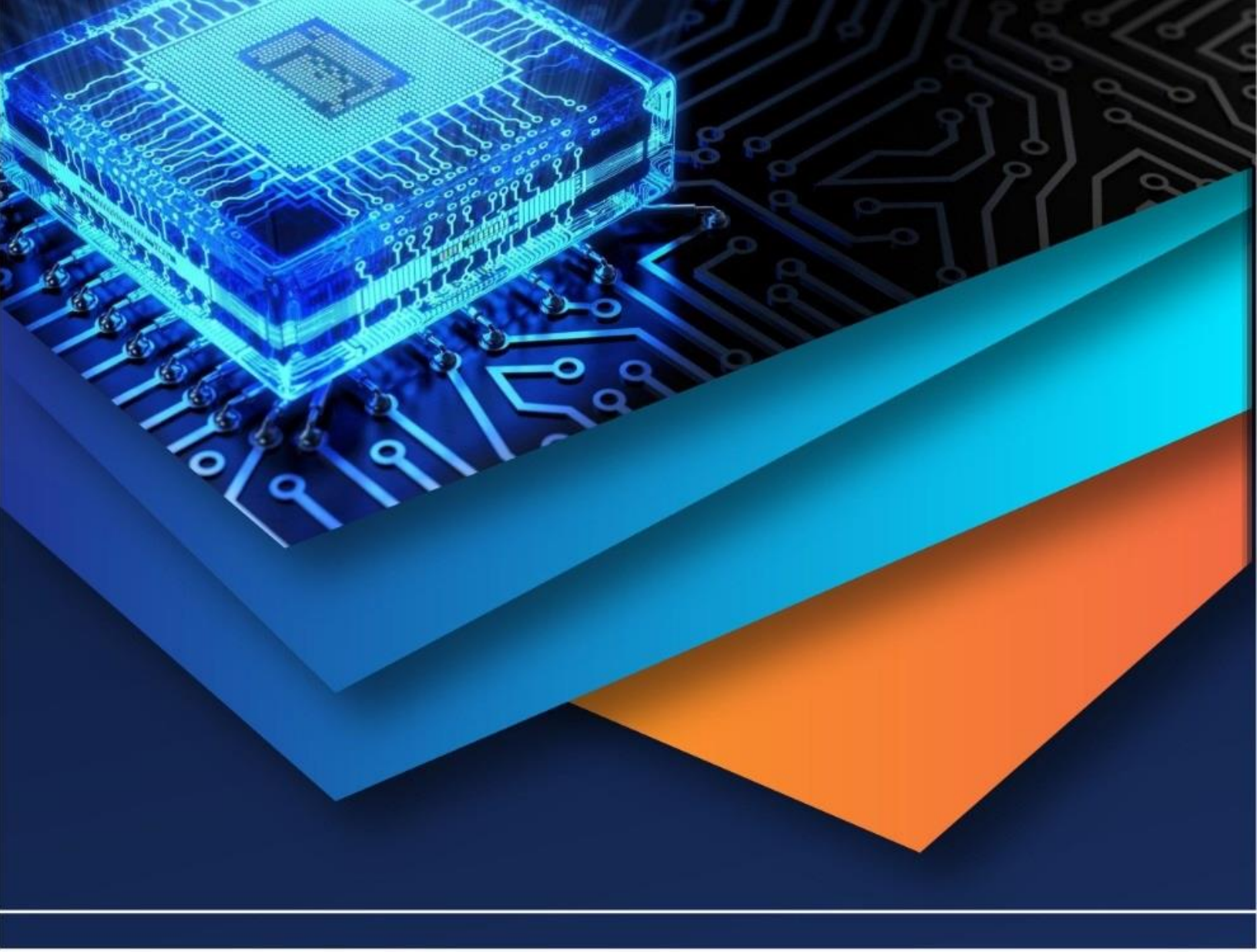
[6] Redmon, J., & Farhadi, A. (2016). YOLO9000: better, faster, stronger. arXiv preprint arXiv:1612.08242.

[7] W. Yang and Z. Jiachun, "Real-time face detection based on YOLO," in 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII), 2018, pp. 221– 224.

[8] P. Ren, W. Fang, and S. Djahel, "A novel YOLO-Based realtime people counting approach," in 2017 International Smart Cities Conference (ISC2), 2017, pp. 1–2

[9] Ankit Sachan, "Zero to Hero: Guide to Object Detection using Deep Learning: Faster R-CNN,YOLO,SSD," CV-Tricks.com, 2018. [Online]. Available: https://cv-tricks.com/objectdetection/faster-r-cnn-yolo-ssd/amp/. [Accessed: 27-January2019]

[10] 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)*, Belgrade, 2018, pp. 1-4, doi: 10.1109/TELFOR.2018.8611986.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY