Real Time Object Detection Using Deep Learning

Ashutosh Prajapati
Department of Computer Science and Engineering
Galgotias College of Engineering and Technology
Greator Noida, 201310, Uttar Pradesh, India

Mr. Ajeet Kumar Bhartee
Department of Computer Science and Engineering
Galgotias College of Engineering and Technology
Greator Noida, 201310, Uttar Pradesh, India
ajeet.bhartee@galgotiacollege.edu

Nitin Singh
Department of Computer Science and Engineering
Galgotias College of Engineering and Technology
Greator Noida, 201310, Uttar Pradesh, India
nitinsingh6218@gmail.com

Aryan yadav
Department of Computer Science and Engineering
Galgotias College of Engineering and Technology
Greator Noida, 201310, Uttar Pradesh, India
arsharyanyadav@gmail.com

***Abstract*— Real-time object detection is a large, dynamic, and challenging topic of computer vision. The detection of a single object in an image is referred to as image localization, whereas the detection of several objects in an image is referred to as object detection. This recognises semantic objects of a class in photographs and videos. Real-time object detection is widely available and uses by multiple applications, have items track feature, surveillancing videos, counting persons, self_drive automobiles, face matcher, and track sports ball. With OpenCV, a set of tools for write program primarily for real-time vision in computer, convolution neural networks is a Deep Learning technique for recognising objects. In this we do an extension of well know detection algorithm Yolov3 to detect 10 different classes of objects in a scene. We use Yolov3 as a feature extractor and added a dence layer block after it. which uses yolov3 learning to detect respective classes .**

Keywords—Deep Learning, Computer vision, Convolution Neural Networks.

I. INTRODUCTION

Deep Learning transformed compute power and improved the way apps are developed forever. Applications are becoming smarter, faster, and more capable of performing work that were earlier too difficult for computers to handle. Classifying and identifying items in a scene, analyzing enormous volumes of text, answering passage questions, making art, and competing against and defeating human players at complicated games like Chess are just a few of the more difficult jobs. Artificial Intelligence aspires to mimic the human brain's approach to processing enormous amounts of data including numerous patterns, as well as how the brain can recognise these patterns, reason about them, and then take action using Deep Learning. Deep Learning is capable of understanding data of varying patterns at an accurate rate. That is why it is used in most the innovations in understanding language and images. Research is moving forward at a fast pace with Deep Learning. Deep Learning has a major impact on object detection. Recognizing items is a word which refered a range in vision of device tasks that include activities such as recognising item in photographs. Picture classification entails prediction the class of a single object in a photograph. Items localize is the process to determined the positions of objects into picture, and creating a bound boxes around them. Items detector concat these two work by detecting and classifying single or multi item into pics. When someone mentions "item-recognition," they usual mean "item-detection." Starting person may have trouble distinguishing between different types of computer vision tasks. Object detection is performed by employing a bound to locate the available of items into imagex, and determining the classes types of the objects identified. For an inputs, a photograph made up of one or more objects is used, and as an out-value, box bound with a classes labeled for eached boxes. Humans have capablility, ability of finding and identifed objects in images. The human being eye system can do complexed work, differentiating many small object and recognising obstructions with their small conscioused thought. Because of larged quantity of data present, GPUs, and well formed algorithms, Now we quickly trained computers appliances to detect it and classifyed multiple item into pics with higher accuration.We'll looking terminology same to detect object, localize item, and elements detect and localize functional loss, as well an object finder algorithm "You Only Look Once" (YOLO).

Items locating is way to draw a box around multi-objects in an photo. whereas image differntiation is the best path of assigning a label to an items of class to which it belong. Finding items in input pics is always hard, as it concat these 2 jobs by constructing a box boundary surrounded each interested items in the photograph and assigning it a cls_ name. Problems are referred to as "item recognition."

The phrase "item-recognition" refered to a set of jobs that can use to distinguish item into images. RCNNs, or Region-based-networks, are a set of methods that are tuned for model performance when addressing object recognition and localization problems. YOLO is the second generation of real-time object recognition systems.

II. RELATED WORK

There are many different types of feature descriptors available for detection. The (SIFT) approached developed by Lowe allowing us to extracted distinct in variant characteristics from item we currently intent to detecting. A database have varied key-point featured can be created used a collection of trained photos. Object detector can be accomplished comparing featured key-points in a test image dictionary have varied key-points taken from trained photos for the items we seeked .[1]

This Shape Context Belongiee method involves sampled the edged of an item into interest points and captured the distributed of the sampled on the form with regard to specified points of body. relationship between a body point and other points was explained using distance and angle measurements. To make a histogram, you can bin distance and angles into multiple buckets. The similarity p and other points the form will be captured by the histogram. Matched two body is analogous to comparing spots each form having same boy context. These ways works well for finding text, fingerprint matcher, and other similar tasks..[2]

Another method presented is to utilize a Histogram of Orienting Gradents (H-O-G) as a descripting templates to recognise humans. Normal intensity gradients (edges) are utilized to efficiently describe the local object appearance and shape information.[3]

The region proposals in a CNN are defined by the identification of these regions of interest from the input frames (Convolutional neural network) To reduce the complexity and enhance the speed, these regions are wrapped and fed into CNN layers. SVM is used in RCNN to recognize different classes of cars for recognition purposes. 18 Convolutional layers, subsampling layers, and fully linked layers were combined to compute features for 128 x 128 input frames. Convoluted images are created by moving a kernel through the entire image and producing convolved images, which are then utilized to identify information from the input frames. The R-CNN's Region Proposals assisted in achieving a recognition accuracy of 91.3 percent for a variety of vehicle kinds, with a performance metric of. The suggested traffic surveillance system's competency with state-of-the-art approaches is demonstrated by the performance metric.[4]

By the help of hyper-parameter Fast RCNN get updated.It is made up of a CNN which is also called backbone, a final pooling layer called "ROI pooling," an FC layer called (a(K+1) softmax layer, and bounding box regression. For Fast R-CNN, weight values with back-propagation are critical in all training networks.[5]

Sliding window searching is one of the most widely used region-based approaches for object detection, but we created a novel method that employs fewer windows and has a higher recall rate. Instead of the serial feature extraction procedure used by the previous approach, Faster RCNN, the Fast RCNN method can execute feature extraction on the input image using a neural network. The last layer of the network's features is used for regression and classification in the traditional RCNN method. These features contain high-level semantic information. The Faster RCNN model processes fused features using a multi-scale feature fusion approach, as well as a residual module and pooling layer. To improve the usefulness and robustness of the suggested model.[6]

It is one of the best and faster item detecter for casual need in world, pushing the limits of real-time object detection. YOLO may be train along loss function which direct reflects its speed. maked it ideals for appliances requiring speed, accurate item detector. It would be multiple item finder that learned to detect many items at once. YOLO can find both bounding line and category probabilities for many items of various classes in a picture. If predicting confidence is replaced by single category predict then single item detect can also be perform by multi-box.[7]

The data set utilized in the experiment is an augmentation of the NWPU-VHR10 set of photographs. There are 650 original photos in the database and 398 extracted images. Instead of its predecessor, the YOLOv3-Mobilenet satellite is equipped with a weight less in a network called Mobilenet, which decreases the model scale and computation load having equal detection accuracy. In addition, an IoU K-medians technique is developed for predicting the target's position in real-time. The experiment sets the groundwork for future algorithm simulations on the embedded platform.[8]

III. NETWORK DESIGN

A. Yolov3

Object finding is problem of regression done in YOLOv3 and probable class are provide of such noticed photos. Neural interconnected layers which are convolution are use inside YOLOv3 algorithm for find real-time items. To finding items, through a neural layer interconnect this way just took a forward propagation which is single, as the name implies. This means that whole picture get analyse in single time. It get use to show multiple probable class as well as bounded line at similar time. ResNet and FPN were the inspiration for this algorithm.

B. Description of Each Steps

First, the image is divided into numerous matrix. The measure of single box in matrix are S-x-S. figure at end illustrates how a matrix box is generated from an photograph. In the photo, their are numerous matrix box cells of the same size. Each matrix box cell will find item that shown inside it. If item center showed inside a given matrix box , that cell would be the one to detect it.

The bounded line is a regression outline which act like highlighter of an photograph object. YoloV3 used bounded outline regression to find item width-height, equilibrium point, and category.

The given featured are available in every bounded lines: width-height_, category, and the center of the bound lines.

(I_O_U)Intersections of unions which is an concept in finding of object that explains overlapping of box_. YOLOv3 can use IoU in order find the best fitting boundry box and eliminating all other less feasible overlap boxes.
Bounding lines and their confidence score are predicted for each grid cell. After evaulation of IOU if it equate one, then the bounded lines is considered similar as real-block.That can eliminate all the box not equal to real box.

$$IOU = \frac{\text{Area of intersection regions of real and predicted box}}{\text{Area of Union regions of real and predicted box}}$$

Very Firstly, the photos is separated into matrix block. Each. matrix block B bounded block are forecasted, along with their confidenced_scores. The block estimate the category probability to determine object's category. When union intersects is utilized, the probable bounded block are same to the true bounded line block of the items. This concept eliminated any not necessary bounded line block that not correspond to the attributes of the items. The end finding have unique bounded line block that have items of best fit.

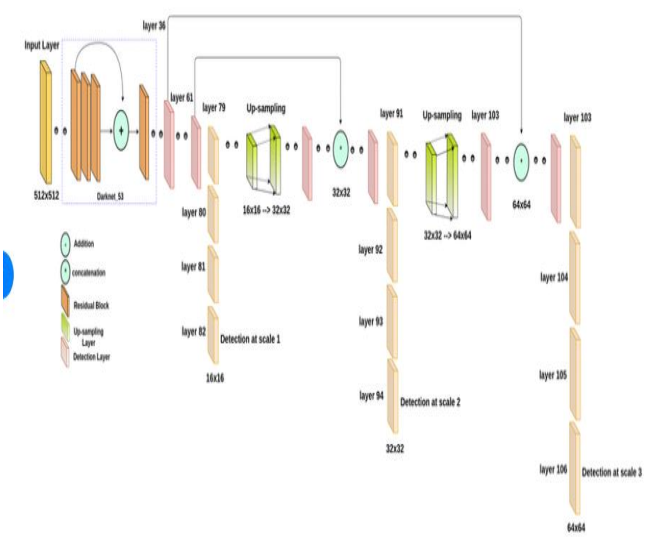

Fig.1 Yolov3 Architecture[9]

C. *Adaptive spacial fusion of feature pyramid*

Feature pyramids are employed in object identification networks to produce probablity of non identical attribute scale or the concatenation of several measure of information. YOLOv3, predicts on three separate scales with strides of 32, 16, and 8. In other words, given a 416 x 416 input image, it generates predictions on 13-X-13, 26-X-26, and 52-X-52 scales. At each level, all features are rescaled and their channel counts are changed. Assume we've got a 416 × 416 input image, and that we need to concat attribute at level_2 (attribute map_size is 26-X-26 and also the channel count is 512_) with attribute at level_3 of high_resolve (where map of atrribute sized is 26-X-26 and that's why the channel count is 512_). (have resolution_of 52-X-52,channel count 256_). The layer_ is get downscaled to 26-X-26 pixels, with the amount increase in channel_ to 512_. The lower_resolutions atrribute at level_1 with 13-X-13 resolve, channel count 1024_ , on the opposite hand, be sampledup to 26-X-26 and also the channel counting lowered till 512_.
Upscaling is accomplished by first compressing the amount of channels of features with a 1-X-1 layers of convolve, so upscaling with interpolation. A 3 by 3 layers of convolve have stride_of two employed to alter the channels amounts and also the resolution at identical time for down-sampling with a 1/2 ratio. Before the 2-stride convolution, a two_stride layers of max_pooling is used for dimensions ratio of 1/4.

D. *Extended architecture of yolo*

In this experiment we use yolov3 as a feature extractor by removing the detection layer from its architecture. Instead of detection at 3 different scale in yolo we add a block of layers at these 3 different scale which consist of a flatten layer then 3 dense layer and two normalisation layer between dense layers. Then all the output of the diffenent scale were concatenated and then pass through 1 dense, 1 normalisation and at the end softmax layer which used for detection purpose. Flatten layer convert the multi-dimentional feature in to a single dimentional feature map which is further given as a input to dense layers. Normalisation layer are for normalising the feature map value in the range of 0-1. At the end we are using a softmax layer which can detect 10 different class of vehicle.

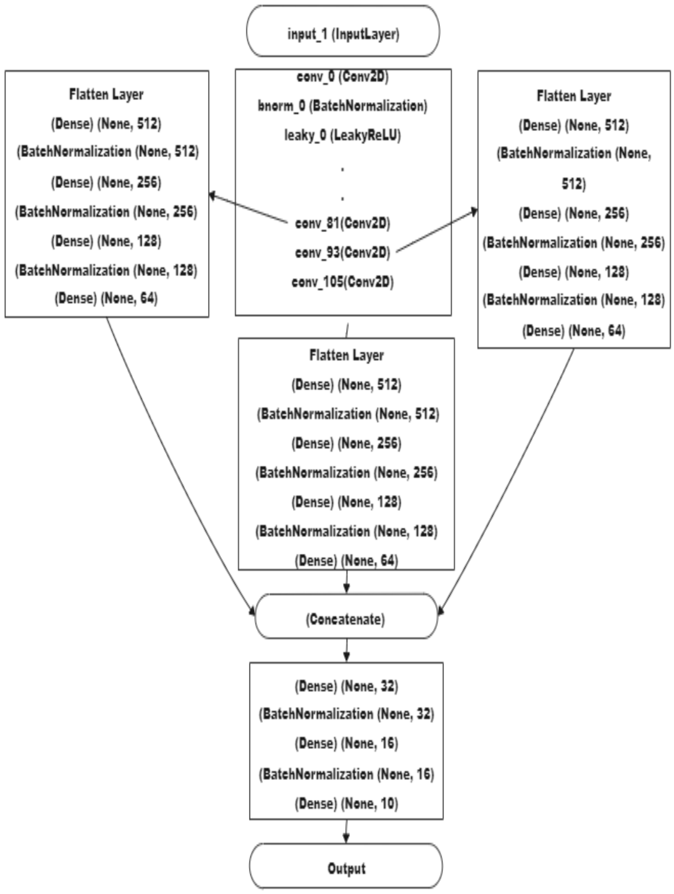

Fig 2 Extended Architecture of Yolov3

IV. EXPERIMENTS AND RESULTS

A. *DataSet*

In this experiment we use dataset is a custom dataset which consist of 10 different vehicle classes consist 800 training images. We use this dataset for training out yolov3 extended model which use yolov3 as a feature extracter and with some additional dense and normalisation layer and it also uses the learning of yolov3 to make a detection of new objects. The vehicle class it detects are ambulance,fire fighter vehicle, Army trucks, police car, JCB, police bike, Road Roller , Gov_official vehicle, armor trucks.

Fig.3 Custom dataset2 of 10 different class of vehicle.

B. *Experimental Setup*

Google Colab is the platform which is generally utilise for model test and train on the custom dataset. GPU(Graphic processing unit) used in this process is Nvidia. it saves time in training the model. It is Free GPU provide by the google Colab platform.

We use python as a programming language and also some of the powerfull libraries of pyhton like tensorflow, keras which consist of multiple inbuilt function and models for the use of research.

C. *Results and Discussions*

Fig3 represents the custom dataset 2 of 10 different class of vehicle. The vehicle class in this dataset are ambulance,fire fighter vehicle, Army trucks, police car, JCB, police bike, Road Roller , Gov_official vehicle, armor trucks.

Fig 4 represent the variation of accuracy with the number of epocs upto which the model is trained and we can se that we can achieve a accuracy of around 65.92% under 30 epocs by using our extended yolov3 model and transfer learning.

TableI shown the result given by the extended yolov3 model on the dataset2 consist 10 different classes of vehicle. In which we are using transfer learning approach and use yolov3 as a feature extractor. We freezed the yolov3 portion at the time of training and train the newly added layers only. So that we can use the learning of yolov3 model to detect our classes. Which can reduce time for training and we can make a detector by training it on small dataset also. By this we can achieve a desent accuracy and a fast trained model for detection.

TABLE I. Result of Ext. yolov3 on vehicle custom dataset.

Model	MAP(%)
Extended Yolov3	65.92

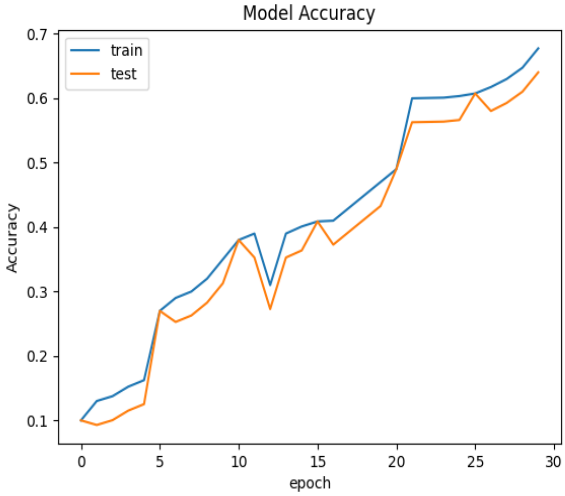

Fig. 4 Accuracy vs epocs graph.

```
[[8.6280698e-01 4.4364872e-04 1.2764307e-04 2.3444109e-03 1.0480608e-01
  2.4583612e-02 1.4328732e-03 8.6584267e-05 8.7538554e-04 2.4928388e-03]]
0
9
6
5
8
Ambulance
Road_roller
JCB
Government_Official_vehicle
Police_bike
```

Fig..5 Output after detection using trained model.

Fig.6 Predicted result by Extended model.

Fig 7 Model Predict the Ambulance in the traffic.

V. CONCLUSION

When compared to the other algorithm Yolov3 is a fast and accurate algorithm for detection and it is already trained over a large preprocessed dataset coco so instead of making a new model from scratch we can use the learning of these highly efficient models like yolo by use of transfer learning.This approach require less computaion power and can be used with small datasets. When compared to other object detection approaches, this technique delivers better detection results then SIFT, Shape context and Hog based approach.. Accuracy measuring facters such as precision, maP, F1-Score, Recall and Avg Iou is not dependent on the number of iteration or training time they are not propostional to the number of iteration.

ACKNOWLEDGMENT

We owe a huge debt of gratitude to Mr. Ajeet Kumar Bhartee for his continual advice and monitoring. We are gratefull to them for giving crucial details and their assistance in job. Dr. Vishnu Sharma, HOD, Computer Science Department, GCET, and Dr. Jaya Sinha, Project Coordinator,Computer Science Department, GCET, have been tremendously helpful and supportive throughout our project's duration. We'd like to extend our heartfelt gratitude to the whole teachers and staff of the GCET Computer Science Department.

REFERENCES

[1] Lowe, David G. "Object recognition from local scale-invariant features." Proceedings of the seventh IEEE international conference on computer vision. Vol. 2. Ieee, 1999.

[2] Belongie, Serge, Jitendra Malik, and Jan Puzicha. "Shape matching and object recognition using shape contexts." *IEEE* transactions on pattern analysis and machine intelligence 24.4 (2002): 509-522.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conferenceon, volume 1, pages 886–893. IEEE, 2005.

[4] Murugan, V., and A. Nidhila. "Vehicle logo recognition using RCNN for intelligent transportation systems." 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET). IEEE, 2019.

[5] Htet, Khaing Suu, and Myint Myint Sein. "Event Analysis for Vehicle Classification using Fast RCNN." 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE). IEEE, 2020.

[6] Yin, Xiaoqing, et al. "Enhanced Faster-RCNN Algorithm for Object Detection in Aerial Images." 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). Vol. 9. IEEE, 2020.

[7] Redmon, Joseph, et al. "You only look once: Unified, real-time object detection." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[8] Wang, Jintao, Wen Xiao, and Tianwei Ni. "Efficient object detection method based on improved YOLOv3 network for remote sensing images." 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD). IEEE, 2020.

[9] Mostofa, Moktari, et al. "Joint-SRVDNet: Joint super resolution and vehicle detection network." IEEE Access 8 (2020): 82306-82319.

[10] Roy, Shuvendu, and Md Sakif Rahman. "Emergency vehicle detection on heavy traffic road from cctv footage using deep convolutional neural network." *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2019.