Object detection using MobileNet SSD in OpenCV python and comparison with YOLO

S Anusha[1], M Bindu[2], G Navya[3] , Venkata Sai[4] ,A.Ajil [5]
[1,2,3,4]Student, Dept. of CSE Engineering, REVA University, Bangalore, Karnataka,India.[5]
Professor, Dept. of CSE Engineering, REVA University, Bangalore, Karnataka, India

Abstract

These days we can see that object detection has a huge number of applications in different sectors. Some of the applications are healthcare monitoring, video surveillance, autonomous driving, robot vision, anomaly detection. One of the application of object detection which is face mask detection was of prominent use during the covid situation. As days are passing the application of object detection is also increasing. So improving the accuracy of object detection by improving the algorithms used, methods used would be beneficial. There are different algorithms available for object detection like SSD, YOLO, R-CNN, Fast R-CNN, Faster R-CNN. We have chosen to implement object detection using SSD and YOLO. First we have implemented object detection using SSD algorithm. Then we implemented using YOLO algorithm. Later we have compared the accuracy of both the algorithms.

Index Terms-- ssd, yolo, tensor flow, openCV, matplotlib, mobie net, deep neural networks, coco

1.INTRODUCTION

Object detection has two parts object localization and object classification. Object localization locates object in an image. Object classification classifies located object in appropriate category. There are many object detection algorithms available like SSD(Single Shot Detector), YOLO(You Only Live Once), R-CNN, Fast R-CNN, Faster R-CNN. We are carring out the object detection using SSD and then YOLO and doing the comparision. [6][7]SSD which stands for single shot detector is the object detection algorithm which carries out both the steps i.e localization and classification in single shot.

[2]Hence detects fast. There are many classification algorithms available ResNet, DarkNet-53, VGG-16, MobileNet. We are using MobileNet as our classification algorithm. MobileNet is based on CNN(Convolutional Neural Network). The job of MobileNet layers is to convert the pixels from the input image into features, these features describe the contents of the image, and pass these to the other layers. OpenCV is the (Open Source Computer Vision library) is an open source computer vision and machine learning software library. Python is the programming language we are using in the jupyter notebook to implement the project.

Face unlock in our smartphone and self driving cars are applications of object detection. Technically object detection is a technology that includes computer vision and image processing used to detect objects in images and videos. To make it more clear we can take the example of self driving cars. Self driving cars make use of the moving object detection technology where computer vision and image processing are used to determine the distance between the car and the moving objects to create alert and accordingly guide the self driving cars. Artificial intelligence is the basic principle that actually drives object detection. Definitely we can see a lot of applications of artificial intelligence and object detection in the upcoming days and hence we can see many more job opportunities related to this field. Many pre trained models are already available so we don't have to train each and every category of the object and that model can be used with the help of few lines of codes. It's all possible because of deep learning algorithms or availability of the computational resources.

1.1 Literature Survey

[1]In the paper object detection for autonomous driving using YOLO(You Only Look Once) written by the authors Abhishek Sarda, Dr.Shubhra Dixit, Dr.Anupama Bhan and published on 31st march 2021 the algorithm used is You Only Look Once. The advantages are it has fast detection time, YOLO predicts accurate results and provide minimal background errors and the disadvantages are they are considering only value of weights for

1000 epochs to avoid overfitting. Only 5% increase is seen in the mAP value from 1000 epochs to avoid overfitting. Some false positives and false negatives are still identified.

[2]In the paper object detection system based on SSD(Single shot detector algorithm) written by the authors Qianlun Shuai, Xingwen Wu and published on 24th November 2020 the algorithm used is SSD(single shot detector). The advantages are SSD model accurately detects objects at different scales. It predicts the different feature mapping. SSD can obtain feature maps at different scales and the disadvantage is the model needs to be improved with the introduction of Convolution block attention model.

[3]In the paper pothole and object detection for an autonomous vehicle using YOLO written by the authors Kavitha R, Nivetha S and published on 26 may 2021 the method used was as follows. Initialize the package/library in openCV . Then load the network model from labelled dataset . Then initialize the parameter containing weight, model, threshold. Later camera captures the object. Then read the frame from the captured scene. After that predict the object from the frame . Then extract the feature of the object and compare it with YOLO dataset. Later object detection happens. Then if object matches with the dataset then boundary box with class name and confidence value is displayed as output else if object dosen't match then go back to predict object from frame step and continue further steps until it matches. COCO and VOC dataset were used. The advantages are Region based CNN (RCNN) is too slow. When compared to RCNN, Fast RCNN and Faster RCNN are better in terms of speed. When compared to these YOLO and SSD are having highest speed and simpler architecture and the disadvantages are Even though Fast RCNN and Faster RCNN are slow when compared to YOLO and SSD, Fast RCNN and Faster RCNN are better in terms of accuracy. In this paper to carry out object detection they have specifically used YOLOv3, its speed is lesser than YOLOv5.

[4]In the paper a pedestrian detection method based on YOLOv3 model and image enhanced by Retinex written by the authors Hongquan Qu, Tongyang Yuan, Zhiyong Sheng, Yuan Zhang and published on 4 february 2019 the method used was carring out

localization first then classification but one extra thing done here was they compared two models with and without image enhancement. YOLOv3 object detection algorithm is used. darknet-53 in the CNN algorithm used. For enhancing the image quality retinex is used. Instead of retinex they could have used other image enhancement techniques like histogram eqilization. But what histogram equilization does is it expands gray scale, different gray scale will be trying to scale down to one. Hence details are lost. Using high pass filter, low pass filter to improve image quality is also not a good idea as it only smoothen or sharpen image. So retinex is used to improve image quality. Retinex is made out of two words retina and cortex. It's mechanism is similar to human eye method which focuses on brightness and reflection. So the two models were tested and compared. Model without image quality enhancement had 90% detection rate and 5% false alarm rate and model with image quality enhancement had 94% detection rate and 2% false alarm rate. The advantages are since retinex theory is used to enhance the image quality, accuracy is increased. YOLOv3 is used, its detection speed and accuracy is more than SSD and the disadvantage is YOLOv3 is having lesser accuracy than YOLOv5, if YOLOv5 would be used it would have given much more accuracy.

[5] The paper Multiple Real-time object identification using Single shot Multi-Box detection was written by authors Kanimozhi S , Gayathri G and Mala T and got published in the year 2019. The method used in this paper was Singleshot Multi-Box detection. The main advantages of using this method are High speed and accuracy of SSD using relatively low resolution images. The disadvantages are The multi-output layers at different resolutions have impacted the performance hugely, in fact, even removal of few layers resulted in a decrease in the accuracy by 12%.

[6]In the paper object detection based on SSD-ResNet, SSD is employed as basic network structure and the inside VGG16 is replaced with The ResNet101 network. This paper was written by the authors Xin lu,Xin kang,Shun Nishide and Fuji Ren and got published in the year 2019.The advantages of implementing this method are, authors in

this paper used SSD model which is more suitable for solving multi-classification problems.Compared to original SSD model the accuracy is increased by 17%.The disadvantage is SSD-ResNet is better than original model in accuracy but the amount of calculation is increased.

[7] In the paper face detection bassed on single shot detection and camshift written by the authors xizhi hu and bingyu huang was published in the year 2020. The method used was single shot detection and camshift . The advantages are combining of ssd network and improved camshift algorithm improves the detection efficiency which has a strong robustnesss to the effect of light loss and the disadvantage is it includes huge storage requirements.

[8] In the paper face detection based on Object detection and tracking using YOLO written by the authors N Muralikrishna, Ramidi Yashwanth Reddy, Gaikwad Sudham was published in the year 2021. The method used was object detection using YOLO. YOLO has multiple advantages compared to RCNN which is the conventional technique used for object tracking. and the disadvantages are comparitevely low recall and more localization error compared to faster RCNN.

1.2 Methodology of SSD algorithm

Object detection has two parts object localization and object classification. Object localization locates an object in an image. Object classification classifies located object in appropriate category. SSD is the object detection algorithm which carries out both localization and classification in one step that means single shot.

1.3 SSD Algorithm Architecture:

As shown in fig 1 the steps are as follows:

Step 1:[11]First an image is passed with two objects in it. A condition must be satisfied before sending an image as input in object detection that is the object must have ground

truth box around it. Ground truth box will further help us locate the object. Step 2:The image is passed through the classification algorithm i.e MobileNet, there feature map is extracted. The feature is compared with the feature in dataset. There it is defined as a particular object has these salient features then its a dog, else if a specific object has certain other set of salient features then it is a cat. So accordingly the features are compared with the dataset and it classifies that object to a class with maximum similarity of salient features with the dataset.

Step 3:[10] Next we encounter 6 convolutional layers which help us to locate the object .

What these 6 convolutional layers does is it gives 8732 predictions for each object. That means each object has 8732 bounding boxes as prediction around it.

Step 4: The image reaches object detection. By the time the image reaches the object detection the objects have 8732 predictions each, which is a large number, so the non max supression removes all the duplicate predictions.

Step 5: So the next step is sending the image to non max suppression. Yet there is a large number of prediction so those predictions are filtered out with more than 50% overlap of bounding box(prediction) over ground truth box. The top 200 predictions are taken based on confidence score. [5]Here the bounding boxes with less confidence score are eliminated. SSD makes use of IOU(Intersection Over Union) to find a bounding box(prediction with maximum overlap with ground truth box.

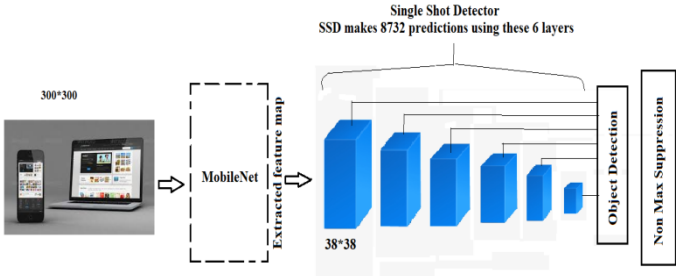

Fig 1 : SSD Architecture

1.4 Methodology of YOLO algorithm

Any idea implemented on test set gives good results then the idea is good. Using rigorous evaluation proved that this idea is better. Speaking about bounding boxes, sometimes square boxes locate the object, but sometimes a smaller or bigger square bounding boxes might be required, sometimes rectangle shaped bounding boxes might be required to locate the object and sometimes elongated rectangle might be required not horizontal. This way problem keeps increasing. We need to check for all these shape and size bounding boxes, it takes time. So the solution is object detection as a regression problem. What exactly happens here is a specific shaped and sized bounding box is placed at one edge of image and runs the classifier. It checks if object is there that is it takes the features and compare with the dataset if the features dosen't match with the feature set of any object in dataset then the box moves to next position and same thing repeats until object is found. This is sliding window based algorithm .Now the solution according to YOLO algorithm is that they made a more generalised training set is formed. Here there are 4 parameters for bounding box they are bx,by,bw,bh, then object is there or not, and three classification parameters c1,c2,c3. Here the bounding box is a real number i.e bounding box parameters dosen't tell whether it belongs to class 1 or class 2 or class 3 so only classification is not enough . So YOLO is declared as a regression problem. Regression is a supervised machine learning algorithm which predicts a specific output based on the set of features the only difference between classification and regression is that in classification the output is a label and in regression the output is a real number. Example of regression is based on size of the house the cost of the house is predicted. As size of the house increases cost of the house also increases. Here the size of the house is independent variable and the cost of the house will be dependent variable. As size of the house increases cost also increases i.e continuous and cost which is output is a real number. That is in regression the output will be real values or continuous nature. Another example is during winter sales of sweater increases.In YOLO along with classification localization must also be done so the

bounding box parameters are needed. The bounding box parameters give real number as output so YOLO is said to be regression problem and not only classification problem.

1.5 YOLO algorithm

YOLO algorithm is a single neural network (only one layer of input nodes). Here the process is simplified i.e input is passed through a deep learning architectire and object will be detected. [1][8]YOLO is faster than R-CNNs and Fast R-CNNs. Only YOLO will do all detection work in self driving cars i.e no need to use many algorithms. In self driving cars an object is detected but by the time its recognized what exactly it is, it might have travelled with 60km/hr speed and come closer to object. This delay happens with R-CNN, Fast R-CNN . Hence in self driving cars we can't use R-CNN, Fast R-CNN.This problem is solved in YOLO the delay is reduced i.e while training we train the model using real objects but still it recognizes what object it is. YOLO generalizes in a better way i.e even for paintings it detects object.[3] Accuracy of YOLO is slightly lesser than R-CNN, Fast R-CNN.YOLO divides the image into N grids. [4]YOLO detects in single neural network only.

As shown in the fig 2 the steps are as follows:

Step 1: Image will be classified based on similarities of their salient features. The salient features will be compared with the dataset and accordingly will be classified.

Step 2: [9]It passes the image through several convolutional layers and detects the object.

Step 3: The image is divided into several grids and then it will be located. In the fig 2 we can notice that there are many grid cells of equal dimension.

Step 4: Every grid cell will detect objects that appear within them.

Step 5: Bounding box regression helps in locating.

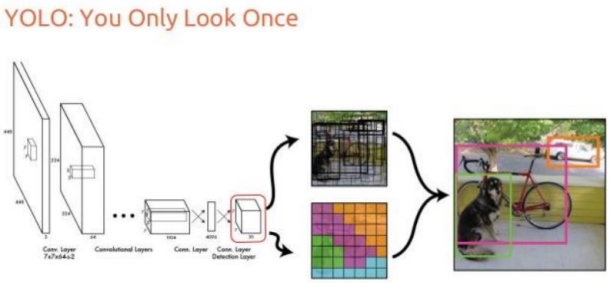

Fig 2 : YOLO Architecture

2.1 Implementation of SSD

In this project we tried to detect objects using mobile net ssd pre trained deep learning models and opencv using python as programming language. Using these pre trained models we classified labels with the help of training data.Based on salient features we will classify the image to which category a image belongs to. Deep learning algorithm that we will be using for image classification is MobileNet. MobileNet ssd algorithm gives good speed as well as accuracy.Using ssd we can detect multiple objects in single shot.So localization and classification is done at the same time unlike other methods which require two or more shots to detect. First we need to import cv2 that is opencv,this library is used to perform image processing. Then we need to import matplotlib.It is used for plotting. There are many deep learning architectures and already pre trained models in tensor flow.We use open cv to load the models.The most recent model is mobilenet ssd version 3.Then we need to have configuration file where we set default values like input size etc. So we load the configuration file. Now we have a create a model that is cv2.dnn_DetectionModel.Then we pass frozen model and configuration file as parameters to this model.This is the main model which detects the objects.Then we create a text file which has names of all labels of coco data set.Coco data set consists of total 80 classes.So we get 80 labels for 80 classes and we copy all these labels in a text file.We need these

labels to check whether we got correct output or not. Then using python list we store these labels.

If we print the class labels we can see the list of labels of all 80 classes and if we check the length it shows 80. Once we load the model the next important thing is to setup configuration. Then as shown in fig 3 we read an image using command cv2.imread and display the image on screen using plt.show. By default tensor flow displays image in BGR format. We need to convert the image to RGB format using cv2.color convert. After loading the model we have to setup the configuration. We gave default input size 320×320 , median Grey scale level and set input swabRb as true. Now the model gives class index ,bbox and confidence as output. we need to set font size and font type for labels. Then we need to have a loop for class index , confidence in a zip. As shown in fig 4 the model will successfully detect the objects in the image with labels and bounding boxes . Similarly we also deployed the code for video and Webcam and efficiently detected the objects .

Fig 3 : Reading the Image

Fig 4 : Detection and labelling the image using SSD

2.2 Implementation of YOLO

First we have to import opencv and numpy libraries and define the input and mention the detection confidence threshold and non-max suppression threshold.Then read the classes from the coco dataset file . For training the model we have used convolutional weights that are pre-trained on Imagenet. Now initialize the weights and configuration files.using OpenCV darknet module read the model configuration file and model weight file. we need to generate the label files that Darknet supports. Darknet requires a .txt file for each image with a line for each ground truth object. The features learned by the convolutional layers are passed onto a classifier which makes the detection prediction.The prediction is based on a convolutional layer that uses 1×1 convolutions.The size of the prediction map is exactly the size of the feature map before it. Then we create a function to find the objects and assign the bounding box thresholds and using the OpenCV rectangle function we draw the bounding boxes and using putText function we add the labels to the output.Then for testing the setup, we will write a small code inside the while loop. Prediction features include the classification loss, loss function, objectness score .We run the loop and the model will detect the objects in the image and draw bounding box around it. Each bounding box has an x, y, w, h coordinates and box confidence score value. The confidence score is the value of how probable a class is contained by that box, as well as how accurate that bounding box is.Confidence score helps us to check weather the output

we got is correct or not. The class confidence score for each final boundary box used as a positive prediction is equal to the box confidence score multiplied by the conditional class probability. The conditional class probability here is the probability that the detected object is part of some class. So yolo v3 prediction has 3 values of height , width and depth. YOLO by default only detects and displays objects with a confidence of .25 or higher. In the fig 5 we can notice that the horse is been detected correctly.

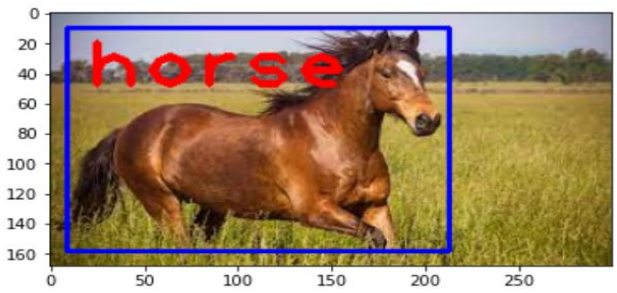

Fig 5 : Detection and labelling the image In YOLO

3. RESULT

After implementing Object detection using SSD Model with OpenCV we understood how exactly SSD model works and also we implemented object detection using pre trained model using the COCO dataset for images and videos with higher accuracies.

We obtained an accuracy of 72% with the SSD algorithm but when comparison with YOLO Algorithm we obtained an higher accuracy of 79% . As the new versions of YOLO keep emerging we will be obtaining higher accuracy hence YOLO is a comparatively better algorithm in comparison with SSD.

4. CONCLUSIONS

We can carry out object detection using many other algorithms like Region based Convolutional Neural Network which is RCNN, fast RCNN, faster RCNN. We got to

know that object detection using RCNN, fast RCNN, faster RCNN is having higher accuracy than object detection using algorithms like SSD and YOLO. But SSD and YOLO are having higher speed and also simpler architecture when compared to RCNN, fast RCNN, faster RCNN. These days people expect fast results and hence using YOLO and SSD for object detection would be a better option. Amongst SSD and YOLO it is better to use YOLO as it has better accuracy than SSD. As new versions of YOLO are emerging we can see the accuracy is improved tremendously. And no doubt that in future we can see more new versions of YOLO with much more accuracy. We can notice that these days object detection has a huge number of applications in many different sectors. Few of the applications of object detection are it is used to track the objects, count people, automated CCTV surveillance, person detection, vehicle detection and many more. Looking at the increasing number of applications we can definitely say that the field of object detection has a huge scope in the future. It will provide lot of job opportunities to people.

REFERENCES

[1] Abhishek Sarda, A. S. (2021). Object Detection for Autonomous Driving using YOLO. 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (p. 5). Tirunelveli, India: IEEE.

[2] Qianjun Shuai, X. W. (2020). Object detection system based on SSD algorithm. 2020 International Conference on Culture-oriented Science & Technology (ICCST) (p. 4). Beijing, China: IEEE.

[3] Kavitha R, N. S. (2021). Pothole and Object Detection for an Autonomous. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (p. 5). Madurai, India: IEEE.

[4] Hongquan Qu, T. Y. (2019). A Pedestrian Detection Method Based on YOLOv3 Model and Image Enhanced by Retinex. *2018 11th International Congress on Image and Signal*

Processing, BioMedical Engineering and Informatics (CISP-BMEI) (p. 5). Beijing, China: IEEE.

[5] Kanimozhi S, G. G. (2019). Multiple Real-time object identification using Single. *2019 International Conference on Computational Intelligence in Data Science (ICCIDS)* (p. 5). Chennai, India: IEEE.

[6] Xin lu,Xin kang,Shun Nishide,Fuji Ren (2019). Object detection based on SSD-ResNet *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)* (p. 5). IEEE.

[7] Xizhi Hu, B. H. (2021). Face Detection based on SSD and CamShift. 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC) (p. 5). Chongqing, China: IEEE.

[8] N. Murali Krishna, R. Y. (2021). Object Detection and Tracking Using Yolo. 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA) (p. 7). Coimbatore, India: IEEE.

[9] shakhadri. (2021, June 11). *Implementation of YOLOv3: Simplified.* Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/06/implementation-of-yolov3-simplified/

[10] Retrieved from pngkit: https://www.pngkit.com/view/u2w7w7r5e6e6r5y3_text-images-music-video-3d-shape-cuboid/

[11] Candacefaber . Retrieved from Candacefaber: https://www.candacefaber.com/free-laptop-mockup/