Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

Object Detection using YOLO And Mobilenet SSD: A Comparative Study

Sabina N.
CSE Department
MGM College of Engineering and Pharmaceutical Sciences
Valanchery, India

Aneesa M.P.
CSE Department
MGM College of Engineering and Pharmaceutical Sciences
Valanchery, India

Haseena P.V.
Assistant Professor
CSE Department
MGM College of Engineering and Pharmaceutical Sciences
Valanchery, India

Abstract— Object detection refers to a computer vision technology that deals with detecting instances of semantic objects of explicit category in digital images and videos. The main purpose of object detection is to identify and spot one or more effective targets from still images or video data. It comprehensively includes a vital range of techniques like image processing, pattern recognition, artificial intelligence, and machine learning. Face detection, self-driving cars, vehicle detection and a few other technologies use object detection. Real-time object detection being a vivacious and complex area of computer vision needs faster computation power to identify the object at that specific time. The accuracy of object detection has increased tremendously with the advancement of deep learning techniques. In this work, two single-stage object detection models namely YOLO and MobileNet SSD are analysed based on their performances in different scenarios. Both models use Convolutional Neural networks for object detection. Different parameters used to determine the accuracy in detecting objects include loss function (LP), mean average precision (MAP), frames per second (FPS), etc.

Keywords- ObjectDetection,CNN,YOLO,MobileNet SSD,Detection Accuracy

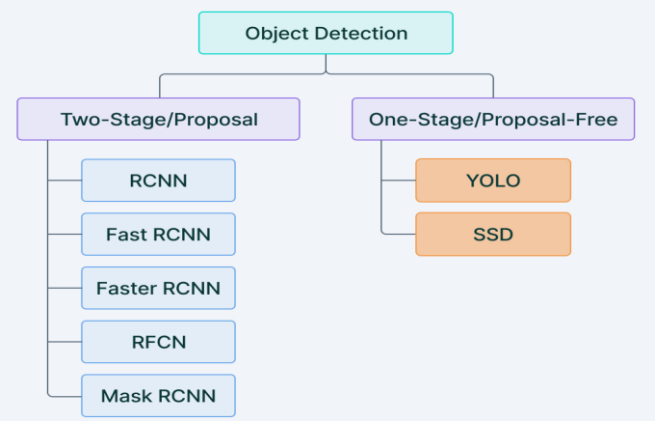
I. INTRODUCTION

For object detection, artificial neurons are used in deep neural networks which are similar to humans composed of neurons. Object detection thus refers to the detection and localization of objects in an image that belong to a predefined set of classes. Tasks like detection, recognition, or localization find widespread applicability in real-world scenarios, making object detection (also referred to as object recognition) a very important subdomain of Computer Vision.

Generally object detection can be categorized in to two as,

1. Two Stage Detector - where the detection completes in two steps. The first step uses a Region Proposal Networks to generate regions of interests that have high probability of being an object. The second step is the object detection which performs the final classification and bounding box regression of objects. RCNN, Fast RCNN, SPPNET, Faster RCNN etc are some of the two stage detectors.

2. One stage Detector - where the object detection is a simple regression problem that takes an input and learns the class probabilities and bounding box coordinates. YOLO, YOLO v2, SSD, RetinaNet etc comes under the one stage detector. Object detection is an advanced form of image classification where a neural network predicts objects in an image and points them out in the form of bounding boxes.



The main purpose of our analysis is to compare the operational performance and accuracy of the object detection techniques YOLO and MobileNet SSD in different aspects and feature a portion of the notable elements that make this study stand out.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

II. YOLO (YOU ONLY LOOK ONCE)

A. *What is YOLO exactly?*

The term YOLO refers to You Only Look Once. YOLO algorithm performs real-time object detection using covolutional neural nework(CNN). As the name suggests, the algorithm solely needs a single forward propagation through a neural network to detect objects. This means that prediction in the entire image is performed in a single algorithm run. The CNN is used to predict various class probabilities and bounding boxes simultaneously. The major features of YOLO are speed, high accuracy and its learning capability.

- Speed - This algorithm has an improved speed for real-time object detection.
- High Accuracy - This technique provides accurate results with minimal background errors.
- Learning Capability - The excellent learning capabilities of this algorithm enable it to learn the representations of objects and apply them in object detection.

B. *How does YOLO works?*

The working of YOLO Algorithm can be explained through three techniques as follows:

1. Residual blocks

First, the image is divided into different grids. Each grid is of the dimension S x S. The procedure of converting an input image into grids is represented in the following image.

Every grid cell will detect objects that appear within them.

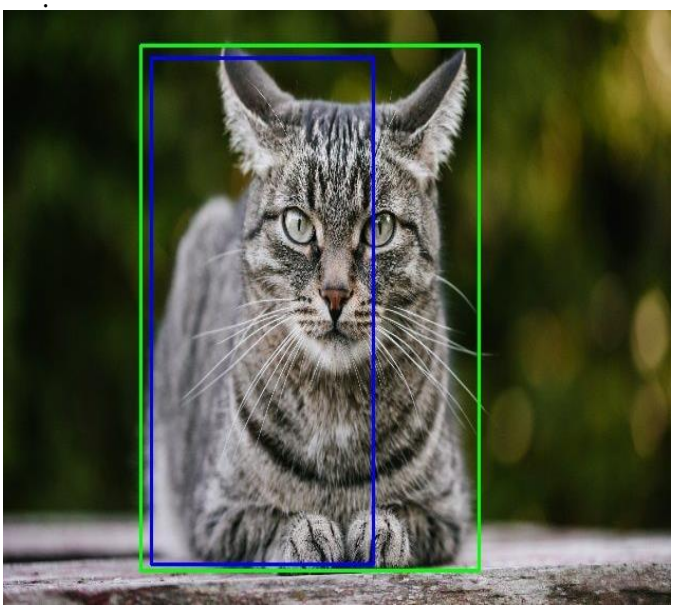
2. Bounding box regression

A bounding box is an outline that highlights an object in an image with some attributes such as Width (bw), Height (bh), Class (for example, person, car, traffic light, etc.),this is represented by the letter c. Bounding box centre (bx,by). YOLO uses a single bounding box regression to predict the height, width, centre, and class of objects.

3. Intersection over Union (IOU)

Intersection over Union (IOU) is a mechanism in object detection that describes how boxes overlap. YOLO uses IOU to surround perfect output boxes for the objects perfectly. Each cell in the grid is responsible for predicting the bounding boxes and their confidence scores. If the predicted bounding box is the same as the real box, then the IOU is equal to 1. This technique can eliminate the bounding boxes that are not equal to the real box.

.



In the image given above, among the two bounding boxes, the blue box represents the predicted box while the green box represents the real box. YOLO algorithms make sure that the two bounding boxes are equal.

C. *Through the history of YOLO*

YOLO has been first introduced by Joseph Redmon, a graduate of the University of Washington in 2016. It was a milestone in object detection research because of its capability of detecting objects in real-time with better accuracy. The main implementation of Redmon's YOLO is based on Darknet, which is a very flexible research framework written in low-level languages and has produced a series of the best real-time object detectors in computer vision: YOLO, YOLOv2, YOLOv3, and YOLOv4.

YOLOv2: YOLOv2 was released in 2017, and its architecture made a number of iterative improvements on top of YOLO including BatchNorm, higher resolution, and anchor boxes..

YOLOv3: YOLOv3 was released in 2018 which is built upon previous models by adding an objectness score to bounding box prediction, adding connections to the backbone network layers, and making predictions at three separate levels to improve performance on smaller objects.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

YOLOv4, YOLOv5, and Much More.....

After the release of YOLOv3, Joseph Redmon discontinued computer vision researches. Researchers like Alexey Bochkovskiy and innovators like Glenn Jocher started to open source their advancements in computer vision research.

YOLOv4: YOLOv4 was released in April 2020 by Alexey Bochkovskiy, which introduced improvements like improved feature aggregation, a "bag of freebies" (with augmentations), miss activation, and more.

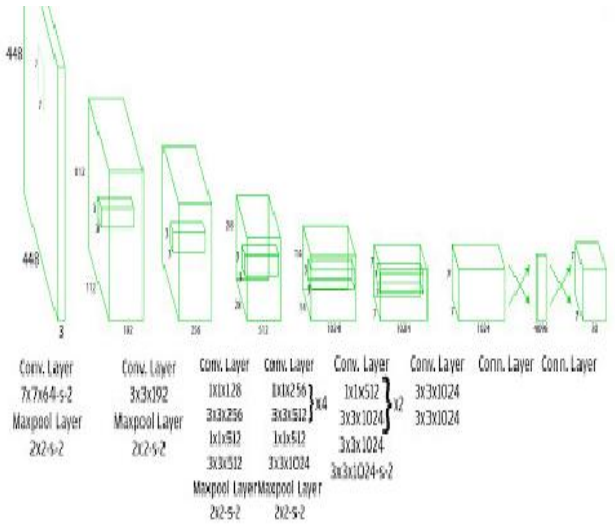
YOLOv5: YOLOv5 was released June in 2020 by Glenn Jocher, which is different from all other prior releases, as this is a PyTorch implementation rather than a fork from the original Darknet. Like YOLO v4, the YOLO v5 has a CSP backbone and PA-NET neck. The major improvements include mosaic data augmentation and auto-learning bounding box anchors.

PP-YOLO: It was released in August 2020 by Baidu, which is based on the YOLO v3 model. The major goal of PP-YOLO is to implement an object detector with relatively balanced effectiveness and efficiency that can be directly applied in actual application scenarios, rather than propose a novel detection model.

Scaled YOLOv4: It came out in November 2020 by Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. The model takes advantage of Cross Stage Partial networks to scale up the size of the network while maintaining both the accuracy and speed of YOLOv4.

PP-YOLOv2: Again authored by the Baidu team, this was released in April 2021which and made minor tweaks to PP-YOLO to achieve improved performance, including adding the miss activation functions and Path Aggregation Network.

D. YOLO Architecture:



This architecture takes an input image and resizes it to 448*448 by retaining the same aspect ratio and performing a technique called padding. This image is then passed to the CNN network. This particular model has 24 convolution layers, 4 max-pooling layers followed by 2 fully connected layers.

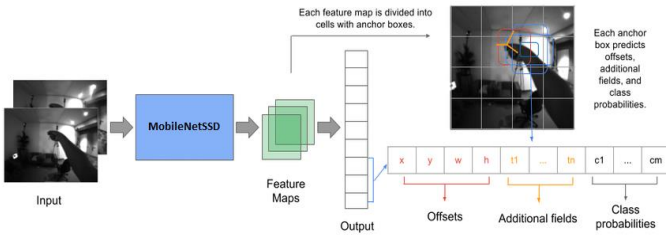
III. MOBILENET SSD

A. What is MobilenetSSD?

Convolutional neural networks are used to develop a model which consists of multiple layers to classify the given objects into any of the defined classes. These objects are detected by making use of higher resolution feature maps and are possible because of the recent advancement in deep learning with image processing. Mobilenet SSD is an object detection model that computes the output bounding box and class of an object from an input image. This Single Shot Detector (SSD) object detection model uses Mobilenet as the backbone and can achieve fast object detection optimized for mobile devices.

B. SSD

The term SSD stands for Single Shot Detector. The SSD technique is based on a feed-forward convolutional network that generates a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes and is followed by a non-maximum suppression step to produce the final detections[9]. Boxes contain offset values (cx,cy,w,h) from the default box. Scores contain confidence values for the presence of each of the object categories, the value 0 is reserved for the background.



SSD introduces multi-reference and multi-resolution detection techniques. Multi-reference techniques define a set of anchor boxes of different sizes and aspect ratios at different locations of an image, and then predict the detection box based on these references. Multi-resolution techniques allow detecting objects at several scales and at different layers of the network. A SSD network implements an algorithm for detecting multiple object classes in images by generating confidence scores related to the presence of any object category in each default box. It also produces adjustments in boxes to better match the object shapes. This network is suited for real-time applications since it does not resample features for bounding box hypotheses. The SSD

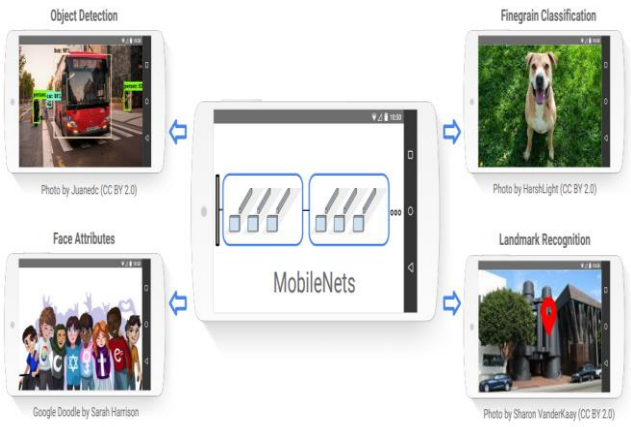
(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

architecture is CNN-based and for detecting the target classes of objects it follows two stages: (1) extract the feature maps, and (2) apply convolutional filters to detect the objects. SSD uses VGG16 to extract feature maps. Then, it detects objects using the Conv4_3 layer of VGG16. Each prediction is composed of a bounding box and 21 scores for each class (one extra class for no object); the class with highest score is selected as the one for the bounded object [3]. The major objective during the training is to get a high class confidence score and this can be attained by matching the default boxes with the ground truth boxes.

C. MobileNet

MobileNet is a class of efficient models called for mobile and embedded vision applications. This class of models is based on a streamlined architecture that uses depthwise separable convolutions to build lightweight deep neural networks. The MobileNet model is based on depthwise separable convolutions which is a form of factorized convolutions. It factorizes a standard convolution into a depthwise convolution and a 1×1 convolution known as a pointwise convolution. The depthwise convolution applies a single filter to each input channel in the case of MobileNets. The pointwise convolution then generates a 1×1 convolution to combine the outputs of the depthwise convolution. A standard convolution has a single step for both filtering and combining inputs into a new set of outputs. But the depthwise separable convolution splits this into two layers, a separate layer for filtering and a separate layer for combining. This factorization reduces computation and model size drastically [8].



MobileNet models can be applied to various recognition tasks for efficient on-device intelligence.

IV. COMPARATIVE ANALYSIS

Different metrics have been proposed to measure object localization accuracy. The Intersection over Union (IoU) which is also called Jaccard Index, is commonly used to evaluate the accuracy of detections. It can be calculated as the area of overlap between a predicted detection and its corresponding ground-truth divided by the area of the union between the predicted detection and the ground truth. The mean IoU for an image is computed by taking the IoU of each class and averaging them, for the binary or multi-class detection problems. This can be applied to all the images of the test dataset to have an average IoU value. Another related detection metric is the F1-score (also called Dice Coefficient), which is calculated as two times by the area of overlap divided by the total number of pixels contained in the detected and the ground truth regions. This measure can be represented in terms of Precision and Recall metrics. It also can be applied to all the target objects present in an image and we can compute the average F1 score for all images of the test dataset. The IoU and F1-score metrics are related and positively correlated for given fixed ground truth. This means that, while comparing two models using IoU if the first model is better than the second one using this metric, it will also be better using the F1 score. When taking the average score over a set of detections in images, the IoU metric has a tendency to penalize quantitatively single “inaccurate” detections more than the F1-score even when both of them can predict a given object instance is badly detected [3].

The standard metrics normally used for analyzing object detection accuracy and speed include recall, precision, F1 score (F1), mean average precision (MAP), and frames per second (FPS). In the target detection process, precision is the ratio of correctly detected targets to the number of all detected targets and recall is the ratio of the number of accurately detected targets to all targets in the sample set. F1 represents the weighted harmonic average of precision and recall. Average precision (AP) is the precision across all elements of a category of pills, as defined in the formula given below:

$$AP = \int_0^1 p(r)\, dr$$

Numerically, MAP is the average value of the AP sum across all categories, and this value is used to evaluate the overall performance of the model.

$$MAP = \frac{1}{n}\sum_{i=1}^{n} AP_i$$

FPS is an indicator that is commonly used for evaluating the speed of model detection. The number of images that can be processed per second is referred to as FPS.

Both detectors can produce acceptable results for different object sizes, illumination conditions, image perspective, partial occlusion, complex background and multiple objects in scenes. One of the major strengths of SSD model is the almost elimination of FP cases which is preferable in applications related to the analysis. On the other side, YOLOv3 produces better average results. YOLO struggles to localize objects properly, but SSD is quicker than the previous progressive for single-shot detectors.

For real-time purposes, speed and accuracy are determining factors for smooth functioning. YOLO variants (especially up to YOLOv3) provide excellent accuracy but require computation-intensive hardware. For such devices, this model

(This work is licensed under a Creative Commons Attribution 4.0 International License.)

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 11 Issue 06, June-2022

would suffice the speed requirement. MobileNet-SSD V2 also provides a somewhat similar speed to that of YOLOv5s, but it just lacks in the accuracy. SSD could be a higher choice when we have a tendency to square measurable to run it on a video and therefore the truth trade-off is extremely modest. YOLO is a better option when exactness is considered than you want to go super quick. So, either of the models can be chosen depending on the requirement of various applications.

CONCLUSIONS

Real-time object detection and tracking on video streams is a crucial topic of surveillance systems in many field applications. The objective of our paper is to make a comparative study on two object recognition systems using CNN to identify the objects in the images. We studied and analyzed the YOLO object detection model and MobileNet SSD model for performance evaluation in different scenarios. Each of the compared models has its own unique properties and is successful in its respective applications. YOLO provides better accuracy compared to MobileNet SSD, which provides more detection speed.

REFERENCES

[1] Ashwani Kumar , Sonam Srivastava ,"Object Detection System Based on Convolution Neural Networks Using Single Shot Multi-Box Detector",Third International Conference on Computing and Network Communications (CoCoNet'19.)

[2] Alexey Bochkovskiy , Chien-Yao Wang,,Hong-Yuan Mark Liao,"YOLOv4: Optimal Speed and Accuracy of Object Detection",arXiv:2004.10934v1,23 April 2020.

[3] Ángel Morera , Ángel Sánchez , A. Belén Moreno , Ángel D. Sappa and José F. Vélez "SSD vs. YOLO for Detection of Outdoor Urban Advertising Panels under Multiple Variabilities", Sensors 2020, 20, 4587; doi:10.3390/s20164587.

[4] Mark Sandler, Andrew Howard , Menglong Zhu , Andrey Zhmoginov and Liang-Chieh Chen ,"MobileNetV2: Inverted Residuals and Linear Bottlenecks",arXiv1801.04381v4.

[5] Mohit Phadtare , Varad Choudhari , Rushal Pedram and Sohan Vartak, "Comparison between YOLO and SSD Mobile Net for Object Detection in a Surveillance Drone", IJSREM, 2021.

[6] Harshal Honmote , Pranav Katta , Shreyas Gadekar and Prof. Madhavi Kulkarni,"Real Time Object Detection and Recognition using MobileNet-SSD with OpenCV", International Journal of Engineering Research & Technology (IJERT), ISSN: 2278-0181, Vol. 11 Issue 01, January-2022.

[7] Lu Tan, Tianran Huangfu, Liyao Wu and Wenying Chen,"Comparison of RetinaNet, SSD, and YOLO v3 for real-time pill identifcation", Tan et al. BMC Medical Informatics and Decision Making (2021) 21:324 https://doi.org/10.1186/s12911-021-01691.

[8] Andrew G. Howard, Menglong Zhu, Bo Chen,Dmitry Kalenichenko Weijun Wang,Tobias Weyand, Marco Andreetto and Hartwig Adam, MobileNets: Efficient Convolutional Neural Networks for Mobile VisionApplications,arXiv:1704.04861v1.

[9] Wei Liu , Dragomir Anguelov , Dumitru Erhan, Christian Szegedy , ScottReed,Cheng-YangFu, andAlexanderC.Berg,"SSD:SingleShotMultiBoxDetector",arXiv:1512.02325v5.

(This work is licensed under a Creative Commons Attribution 4.0 International License.)