Comparison of YOLOv3 and SSD Algorithms

Dr. S. V. Viraktamath
Dept. of Electronics and Communication
SDM College of Engineering and Technology
Dharwad, India

Ambika Neelopant
Dept. of Electronics and Communication
SDM College of Engineering and Technology
Dharwad, India

Pratiksha Navalgi
Dept. of Electronics and Communication
SDM College of Engineering and Technology
Dharwad, India

***Abstract*— Object recognition is an advancement associated with computer vision and imaging, which manages to recognize and locate cases in computerized images and recordings for semantic artifacts of a particular class (such as persons, structures, or vehicles) in automated objects and observations. Continuous object identification and following is a huge, energetic yet uncertain and complex region of PC vision. It has end up being a noticeable module for various significant applications like video reconnaissance, self-sufficient driving, face identification; and so forth. As a feature of the overview, the theme investigated incorporate different calculations, quality measurements, speed/size trade-offs and preparing approaches. This paper centers around the two kinds of object identification YOLO (You Only Look Once) and SSD (Single Shot multi-box Detector) class of single step indicators and the Faster R-CNN class of two stage locators [1] and applications of the same.**

Keywords—YOLO, SSD, R-CNN.

I. INTRODUCTION

Object detection is identification of an object inside the image that is close by its confinement and order. Object limitation alludes to distinguish the circumstance of at least 1 object in an image and drawing a bounding box around their degree. Image ordering might be a strategy that is used to group or anticipate the classification of a chosen object in an image. Object detection joins these two procedures and limits and arranges at least one object during an image. Detection is further developed, which conveys what the "primary subject" of the image is though object detection can discover numerous objects, arrange them, and find where they are in the image.

Each object class has its own uncommon highlights that help in arranging the class – for instance all circles are round. Object class detection utilizes these extraordinary highlights. Bounding boxes are predicted by the object detection models. The model predicts bounding boxes and classification probabilities for each object. It is natural for object detection to anticipate too many bounding boxes. In addition, each box has a confidence score that indicates how likely the model actually thinks the picture contains an object. As a final, all boxes whose score falls below a specific edge are removed (called non-maximum suppression).
There are generally two kinds of object detectors, two-stage detectors, such as Faster R-CNN or Mask R-CNN area proposal network to create first stage regions of interest and submit region of proposals down the pipeline for object classification and regression of bounding boxes. These models achieve the highest accuracy rates, but are usually slower. The problem of R-CNN is that it still takes a lot of time to train the network as it has to classify 2000 regional proposals for each image, and hence real time implementation is not possible [1]. Therefore, no learning is required. This may contribute to the development of poor proposals for regions. Then again, there are single-stage proposals such as YOLO and SSD, which treat artifacts detection as a simple issue of regression by taking a picture and learning probabilities of the class and bounding box co-ordinates such models achieve lower rates of accuracy, but they are substantially quicker than two-stage object detectors.

II. COMPARISON OF SINGLE STAGE AND TWO STAGED OBJECT DETECTOR

A. Two Staged Detection

This strategy prioritizes detection accuracy. The two-stage approach separates the detection and postures assessment steps. After object detection, the identified objects are cropped and processed by a separate network for present assessment. This requires resampling the image at any rate multiple times: once for region proposals, once for detection, and once for present assessment. That works well but on the other hand it's delayed as it requires running the detection and classification portion of the model on various occasions. The basic models are Fast R-CNN, Faster R-CNN and Mask R-CNN.

B. Single Staged Detection

This strategy prioritizes inference speed. The proposed technique, on the other hand, does not require a re-sampling of the image, but relies on convolutions to recognise the object and its position in a single forward propogation. This offers a large acceleration in the light of the fact that the image is not re-sampled and the calculation for detection and position assessment is shared. That is a lot faster and significantly more appropriate for cell phones. The most well-known instances of one-stage object detectors are YOLO, SSD, SqueezeDet and Detect Net [2]. The most popular benchmark is the MSCOCO dataset. Models are commonly assessed according to a Mean Average Precision metric.

III. DRAWBACKS OF TWO STAGED OBJECT DETECTION

- It actually sets aside a large measure of effort to train the network needed to group 2000 regional proposals per image.

- As it takes around 47 seconds for each test image, it cannot be actualized in real time.
- The specific search algorithm is a fixed and hence learning is not required at that stage.
- They totally lose all their internal information about the position and the orientation of the object and they route all the information to similar neurons that will most likely be unable to manage this sort of information [3].
- A CNN makes predictions by taking look at an image and afterwards verifying whether certain parts are present in that image or not. In the event that they are, it orders the image accordingly. R-CNN, Fast R-CNN and Faster R-CNN were conceived to address obstructions. In terms of accuracy and training time, the Faster R-CNN was the best algorithm out of all the above by testing its output with the COCO data-set.
- In order to overcome this limitation, Faster R-impediment CNN's was developed. YOLO error analysis against Quick R-CNN reveals that YOLO makes several locale errors [4]. That compromised the accuracy of the SSD compared to the Faster R-CNN, however. In addition, YOLO object detection algorithms have been established using the darknet frames; in terms of accuracy and inferences time, the latest version of, for example, the V3 from YOLO has overrun the Faster R-CNN and SSD [5].

IV. SINGLE STAGED OBJECT DETECTION

A. *YOLO*

YOLO utilizes a totally unexpected tactic, for the object detection in real-time, YOLO is a CNN. The algorithm applies the complete image to a solitary neural network and then isolates the image into regions, predicting bounding boxes and probabilities for each region. These bounding boxes would evaluate the intended probabilities with high accuracy while still being able to run in real time, YOLO is famous for requiring only one advance propagation over the neural network [6]. After non-max suppression (which ensures recognition of each object exactly once) it returns acknowledged objects with bounding boxes. YOLO works by accepting an image as information and dividing it into a S X S grid, taking m bounding boxes inside each grid.

For every bounding box, the network gives a yield 'a' class probability and counterbalance esteems for each bounding box formed. The bounding box with the probability class above the threshold value is selected and used to further find the object within the picture. By order of sizes (45 FPS) YOLO is faster than other object detection algorithms present. The limiting and disadvantage aspect of the YOLO algorithm is, for example, that it faces difficulties when distinguishing a smaller object [7], it is because of the YOLO algorithm's spatial constraints. In an image, YOLO acknowledges artifacts very well unlike sliding window and area proposalbased approaches as it is used to see the whole image during training and testing time so that it gets every insight into the entire image and object and its appearance. The algorithm divides the image into grids and on each of the grid cell runs the algorithm for image classification and limitation. It forecasts N bounding boxes and scores on all grids. The certitude score reflects the exactness of that class's bounding box. As several of these boxes have poor safety values, unwanted bounding boxes or items can be evaded by setting a threshold.

B. *SSD*

The SSD architecture adopts an algorithm for the detection of various object classes in a picture by providing confidence scores associated with the presence of any category of objects. In addition, it creates changes to the shape of the objects in the boxes. This is suitable for real-time applications as it does not re-evaluate bounding box assumptions (like in Faster R-CNN)[8].The SSD architecture is CNN-based and for detecting the target classes of objects it follows two stages: extract the feature maps, and apply convolutional filters to detect the objects. Detection of objects is still an issue in pc vision and recognition of patterns. The key image classification challenges, such as noise robustness, transformations, and obstacles are inherited, in addition new challenges, such as detection of various artifacts, overlapping images, identifying their positions within a picture are also added. A better harmony between quickness and accuracy is achieved by SSD. It only runs a traditional network once an image is inputted and displays a function diagram.

V. DIFFERENCE BETWEEN YOLOV3 AND SSD

Sl.no.	Parameters	YOLO	SSD
1	Model name	You Only Look Once	Single Shot Multi-Box Detector
2	Speed	Low	High
3	Accuracy	80.3% High	72.1% Low
4	Time	0.84~0.9 sec/frame	0.17~0.23 sec/frame
5	Frame per second	45	59
6	Mean Average precision	0.358	0.251

Table 1: Difference between YOLO and SSD

The table 1 shows comparison between YOLO and SSD as regards to speed, accuracy, time, frame per second (FPS) [8], Mean Average Precision (mAP) [11], and whether they can be used for real time applications or not. The table above shows clearly that YOLO is better than the low accuracy and higher FPS SSD algorithm [10]. At 416 X 416 YOLOv3 runs in 29 ms at 31.0 mAP almost as accurate as SSD but approximately 2.2 times faster that SSD [3]. It can be seen clearly that a precise compromise was made to achieve this speed. Even after a low mAP, YOLO has an appropriate mAP to be used in real time applications, and it becomes obvious that it is the best algorithms in its class when used alongside high FPS accurate information.

VI. APPLICATIONS

A. *Advertising Detection*

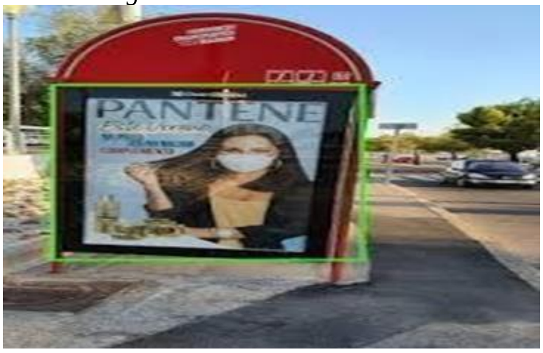

Fig 1: Advertising Detection

Detection of picture advertisement boards in both actual and virtual worlds provides essential applications. For instance, Google Street View could utilizes it to refresh or personalize the advertising that appears on street images.

B. *Animal Detection*

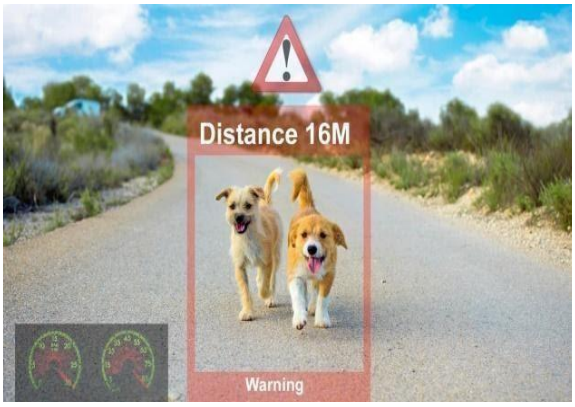

Fig 2 : Animal Detection

For various kinds of creature detection we can use the YOLO model. YOLO model is fit for identifying horse, sheep, cow, elephant, bear and zebra, giraffe from images and real time camera feed and recordings.

C. *Object Detection*

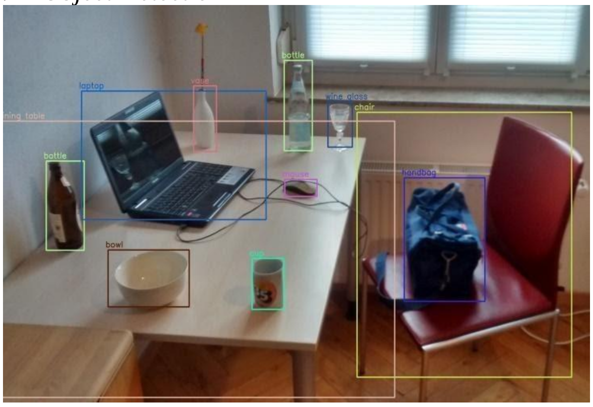

Fig 3 : Object Detection

Object detection is the mechanism by which a variable number of things in a picture are detected and characterised. The main difference is the part that is "variable" . The yield of object detection is variable in comparison with problems such as classification because the distinguishing number of objects will vary from picture to picture[10]. Different objects can be classified using the YOLO model.

D. *Activity Recognition*

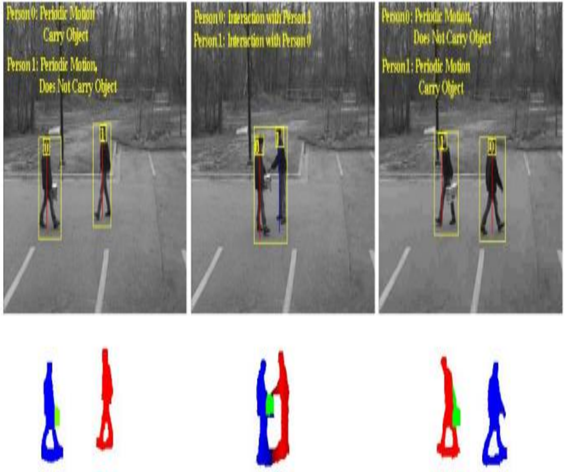
Person 0: Periodic Motion
Carry Object
Person 1: Periodic Motion,
Does Not Carry Object
Person 0: Interaction with Person 1
Person 1: Interaction with Person 0
Person 0 : Periodic Motion,
Does Not Carry Object
Person 1 : Periodic Motion
Carry Object
0
1

Fig 4 : Activity Recognition

The aim of activity recognition is to recognise, from a collection of observations of specialist activity and environmental conditions, the actions and objectives of one or more operators. This area of research has taken account of many informatics networks, as it provides tailored support for different applications and its relation to a number of fields, for example, the relationship between the human machine and the humanistic method.

E. *People Counting*

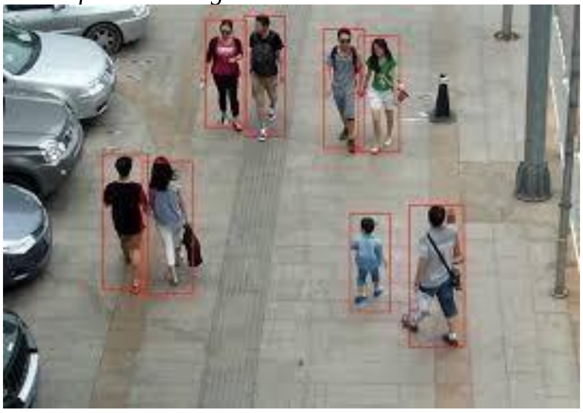

Fig 5 : People Counting

Object detection can be likewise utilized for individuals checking, it is utilized for breaking down store performance or crowd measurements during celebrations. These will be in general more troublesome as individuals move out of the frame rapidly.

F. *Others*

Some other real applications, including logo detection and video object detection. Logo detection in web-based business systems is an important research subject. The logo event with a clear non-rigid transformation is much smaller compared to generic detection.

CONCLUSION

In recent years, deep learning based object detection has been a major focus in research due to its high learning ability and interest in handling constraint, scale transformation, and context switches. So from the above discussion, we can state that the use of the YOLO Model in real life can greatly benefit many organizations. Just as we are probably aware that Yolo would make a marvelous impact in commercial and industrial sectors, as one of the most promising models, this algorithm is generalized to outperform various strategies between natural and various fields from object detection. The purpose of the algorithm is to classify artifacts that use a solitary neural network. The algorithm can be easily rendered and directly trained on a complete image. Above discussed regional strategies restrict the classifier to one region. In predicting borders, YOLO hits the entire picture. Moreover, in backgrounds, it predicts less constructive outcomes. This algorithm is much easier and simpler to use in real time than other classifier algorithms.

REFERENCES

[1] Kanishk Wadhwa, Jay Kumar Behera, “Accurate Real-Time Object Detection using SSD” SRM Institute of Science and Technology, Chennai,2020

[2] Zhong-Qiu Zhao, Member, IEEE, Peng Zheng, Shou-tao Xu, and Xindong Wu, Fellow, IEEE, “Object Detection with Deep Learning: A Review”.

[3] Joseph Redmon, Ali Farhadi, “Yolov3:An incremental improvement”, arXiv preprint arXiv:1804.02767, 2018.

[4] Joseph Redmon, Ali Farhad, “YOLO9000: Better, Faster, Stronger”, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[5] J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, “Selective search for object recognition”, IJCV, 2013.

[6] Erhan, D., Szegedy, C., Toshev, A., Anguelov, D, “Scalable object detection using deep neural networks”. In: CVPR (2014).

[7] Guei-Sian Peng “Performance and Accuracy Analysis in Object Detection” CALIFORNIA STATE UNIVERSITY SAN MARCOS.

[8] Redmon, J., Divvala, S., Girshick, R., & Farhadi, “You Only Look Once: Unified, Real-Time Object Detection.” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). doi:10.1109/cvpr.2016.91

[9] Lin, T.Y., Marie, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.; “Microsoft COCO: Common Objects in Context”. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, September 2014

[10] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C; “SSD: Single Shot MultiBox Detector”. In Proceedings of the European Conference on Computer Vision, Amsterdam, the Netherlands, 11–14 October 2016.

[11] K. Simonyan, A. Zisserman, “Very deep convolutional networks for large-scale image recognition Computer Vision and Pattern Recognition”, 2014.