e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

OBJECT DETECTION: YOLO VS FASTER R-CNN

Fiza Joiya[*1]

[*1]Research Student, Department Of Information Technology, B.K. Birla College Of Art, Science And Commerce (Autonomous), Kalyan, Thane, Maharashtra, India.

DOI : https://www.doi.org/10.56726/IRJMETS30226

ABSTRACT

Object detection is one of the unique abilities of computer vision that locates objects within an image or video. The field of Artificial Intelligence is built on Object detection techniques. Object Detection typically leverages machine learning and deep learning to produce meaningful and accurate results. It basically consists of classification and localization. In recent years there has been an advancement in the state-of-the-art algorithms used for real-time object detection. The objective of this research paper is to compare the state-of-the-art algorithms i.e. you only look once (YOLO) and faster region convolutional neural network (Faster R-CNN). These algorithms are representations of deep neural networks i.e. neural networks with many hidden layers. Both these algorithms are compared to check which one is better, although they both stand-out for their own uniqueness, this paper researches on the area that shows which of the either are more efficient to use even though they have the same core i.e. CNN (Convolutional Neural Networks).

Keywords: Object Detection, Computer Vision, Machine Learning, Deep Learning, State-Of-The-Art, You Only Look Once (YOLO), Faster Region Convolutional Neural Network (Faster R-CNN), Deep Neural Networks, Convolutional Neural Networks (CNN).

I. INTRODUCTION

As humans have a strong sense of visualization they easily detect and identify objects surrounding them, no matter what position or color the object has, but detecting objects is a bit complex and requires a lot of processing when it comes to computers. Computer vision is a field that deals with how computers gain high-level understanding from digital images or videos. Computer vision consists of Object Detection, Image classification, image Captioning and image recognition, etc. Object detection is basically the foundation upon which artificial intelligence is built. Convolutional Neural Network (CNN) are the most common deep learning technology that makes detection more accurate and instantaneous by applying multiple convolutional layers and convolutional computation. All object detection algorithms use Convolutional neural networks.

CNN is one of the artificial neural networks that uses convolutional layers along with other types of layers, such as nonlinear, pooling, and fully connected layers, to create a deep convolutional neural network. It uses backpropagation to train it's convolutional filters. In this research deep learning algorithms YOLO (you only look once) and R-CNN (Regional Convolutional Neural Networks are used for determining which works more accurately and efficiently.

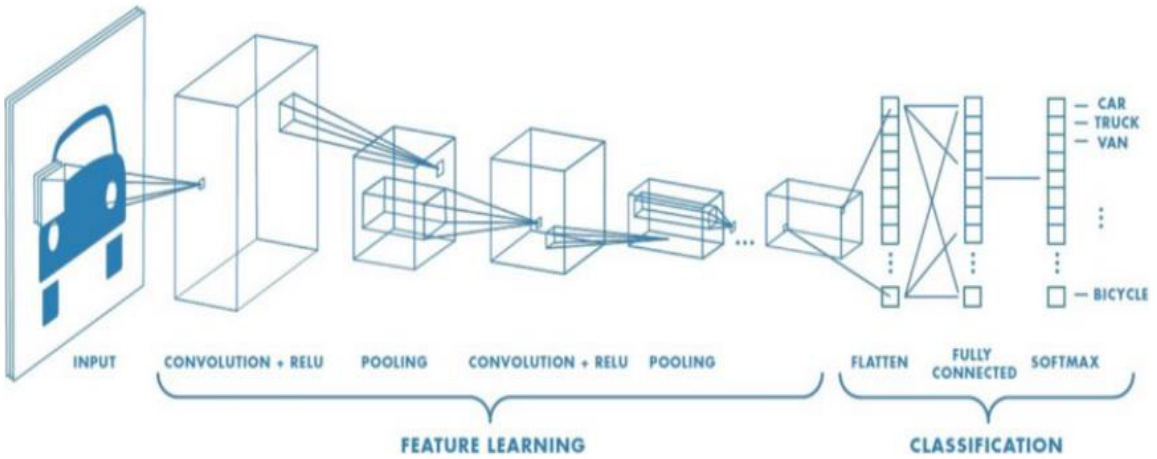

Figure 1: Convolutional Neural Network Architecture.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

Object Detection consists of two types of algorithms the first one is the one-step detection algorithm like You Only Look Once (YOLO) algorithm and the second one is two-step detection algorithm namely Faster Region-Based Convolutional Neural Network (Faster R-CNN) algorithm. The two steps involved are object classification, in which objects are classified based on the colors they have; and object localization, in which objects are located by drawing a bounding box around the detected object.[1]

II. YOLO

YOLO (You only look once) is a new algorithm which means that an image can predict the objects and their locations at one glance. It uses neural networks for real-time object detection. This algorithm has evolved over the years, it started with YOLO v1 (or unified) – It has several localization errors, Yolo v2, YOLO v3, YOLO v4. Currently, YOLOv3 is the state of art algorithm which is used for single stage object detection. YOLOv3 can basically achieve its real-time performance on a standard computer with graphics processing unit (GPU).[2]

The whole framework only needs to use a relatively simple structure of CNN to directly complete the regression of target detection to predict the position of the bounding box and the class of the candidate box.[3]

YOLO focuses on the entire image as a whole and predicts the bounding boxes and then calculates the class probability to label the boxes. It predicts limited number of bounding boxes to achieve its goals. It can classify objects up to 155 FPS (frames per second) in real time, achieving twice the mean average precision (mAP) of other object classifiers. It is a single convolutional network that simultaneously predicts multiple bounding boxes on multiple objects and then generates a class probability for that object.[4]

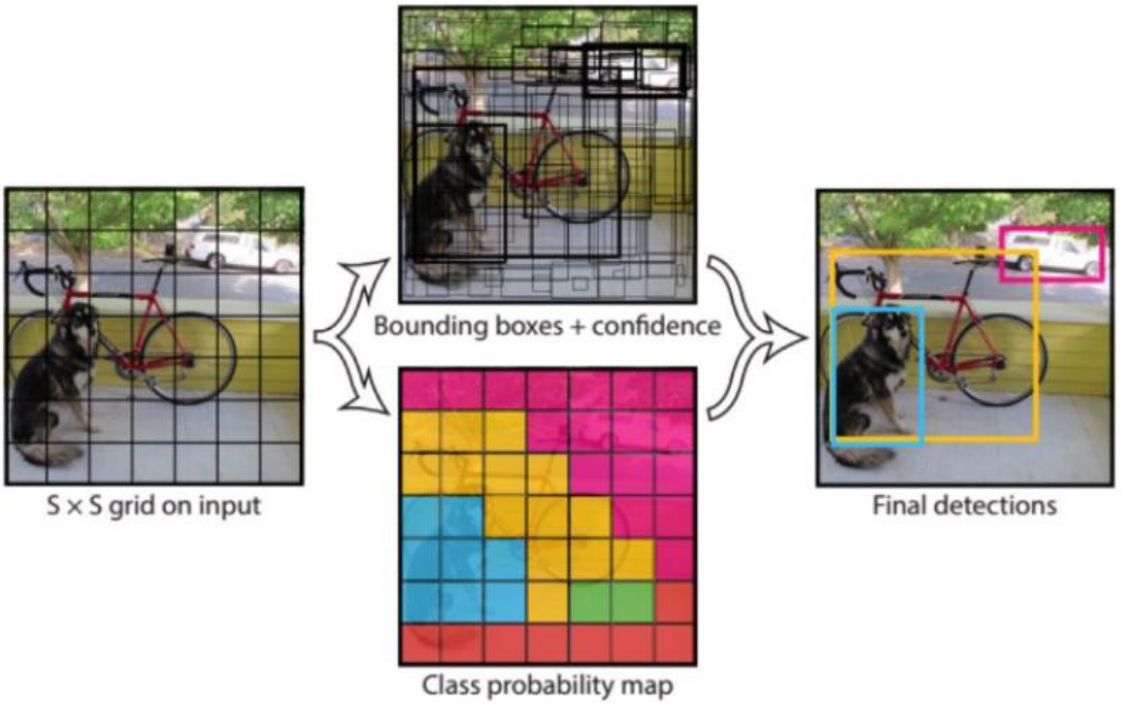

Figure 2: YOLO Bounding Box, Object detection and localization.

In YOLO: -

- The image is divided into M grids, each grid having equal dimensional regions P x P. Each of these grids are responsible for detecting and locating the objects present in it.
- These M grids predict their bounding box coordinates relative to the cell coordinates, along with the object label and the probability of it being present in the cell.
- This highly decreases the computation rate as the cells of the image handle both detection and recognition.
- Non-Max suppression is used to filter through all the boxes, and also eliminates overlapping boxes and duplicate predictions.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

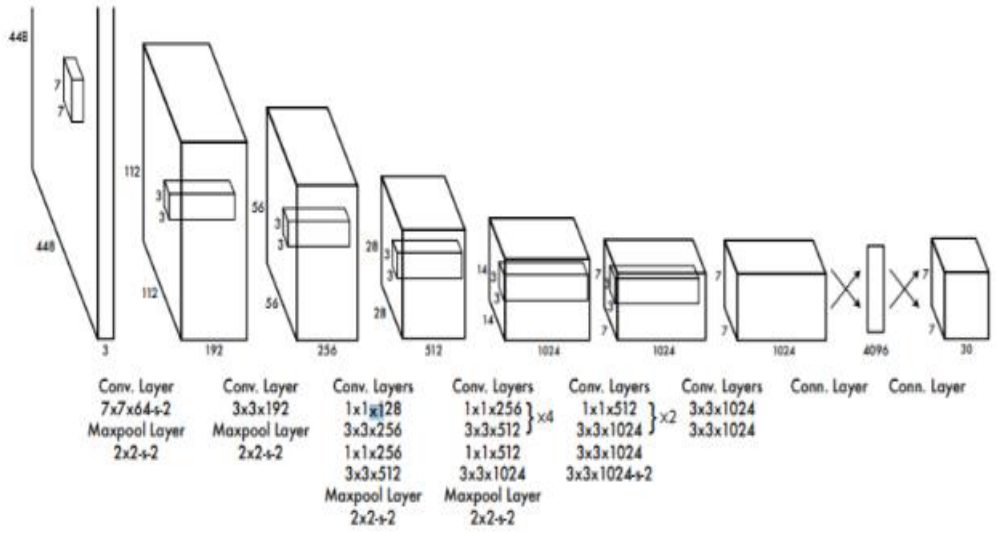

Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1×1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224×224 input image) and then double the resolution for detection.

Figure 3: YOLO architecture and working.

YOLO's architecture has 24 convolutional layers with 2 fully connected layers at the end of the structure.

III. FASTER R-CNN

Faster R-CNN is one of the most preferred and used version of the R-CNN family. It uses a particular selection of search algorithms for proposing regions, which take a few seconds (1or 2) per image and run on CPU computation. Faster R-CNN uses RPN's i.e Region Proposal Networks, which generate region proposals and reduces the time of generation from seconds to milliseconds per image.[5]

In Faster R-CNN,

- RPN is used to genarate bounding boxes i.e. a rectangular box that surrounds an object, that specifies its position, class (e.g.: car, person) and confidence (how likely it is to be at that location).
- In this stage, usually CNN is used to generate features of these objects. Region proposal is not done on the original image but the final feature image which will then be inpput into the ROI pooling (Region of Interest Pooling fixes image size requirement for object detection).
- The output from the ROI pooling layer has a size of (N, 7, 7, 512) where N is the number of proposals from the region proposal algorithm. After passing those ROI pooling outputs through two fully connected layers, the features are fed into the sibling classification and regression branches.
- A classification layer is present to determine which class the object belongs to.
- Finally, a regression layer is used to make the coordinates of the bounding boxes more precise leaving no gaps for errors.
- To deal with different scales and aspect ratios of the objects, anchors are introduced in RPN. An anchor is at each sliding location of the convolutional maps and thus at the center of each spatial window. Each anchor is associated with a scale and an aspect ratio.[6]

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:09/September-2022 Impact Factor- 6.752 www.irjmets.com

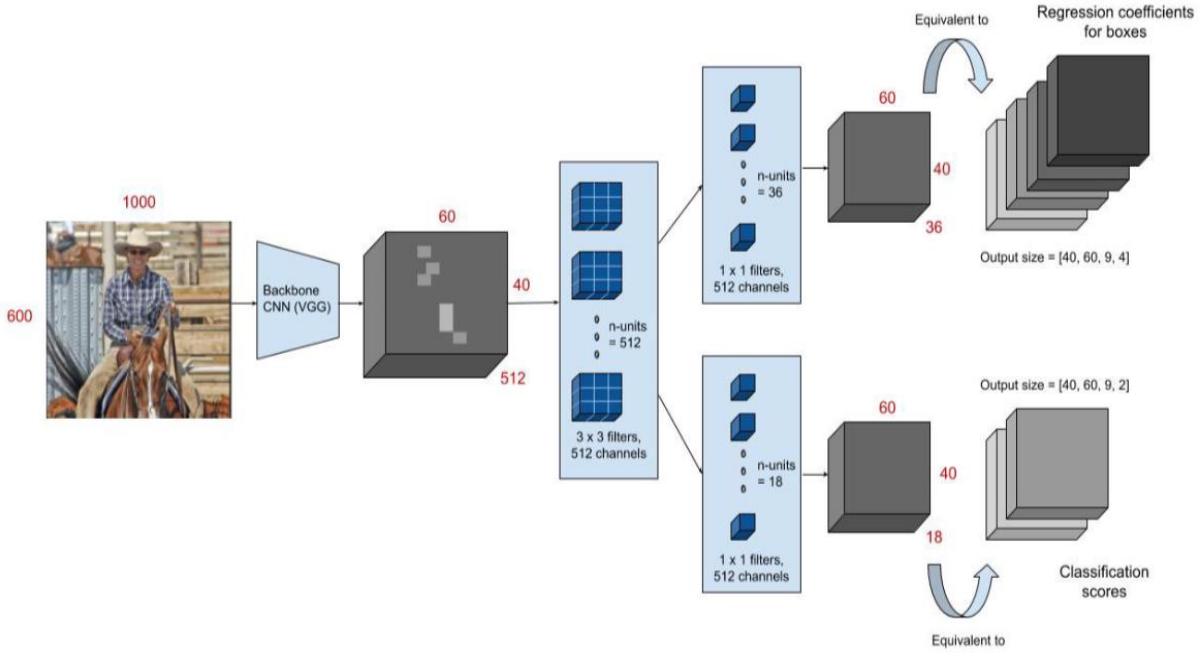

Figure 4: Faster R-CNN architecture.

IV. COMPARISON BETWEEN YOLO AND FASTER R-CNN

Even though both Faster R-CNN and YOLO use CNN as their core and their key purposes is to find a better way of dividing region proposals based on CNN, their frameworks are quite different from each other. Region proposal classification networks (e.g. Faster RCNN) perform detection on various region proposals and end up performing predictions multiple times of various regions of an image, on the other hand, YOLO architecture is more like a fully connected convolutional neural network, the image passes through the FCNN once and then the output gives the prediction. Faster R-CNN offers region of interest to perform convolution on it while YOLO does detection and classification at the same time. YOLO makes less than half the number of background errors as compared to Faster R-CNN. YOLO architecture enables end-to-end training and real-time speed while maintaining high average precision. Faster R-CNN offers end-to-end training as well but involves much more steps as compared to YOLO. Faster R-CNN must be used, if high-end GPUs are available on the deployed devices. Faster R-CNN focuses on speeding up the R-CNN framework by sharing computation and using neural networks to propose regions instead of Selective Search.[7] While YOLO offers promising speed and accuracy over Faster R-CNN, both still somewhere fall behind when it comes to real-time performance.

V. CONCLUSION

The most important part of this research paper is not about finding the best detector, as it lies on the preference of the users. The real question is which detector and what configurations give us the best balance of speed and accuracy that a particular application will require. As compared to Faster R-CNN, YOLO has more advanced applications.

YOLO proves to be a cleaner and more efficient for doing object detection since it provides end-to-end training. Both the algorithms are fairly accurate but, in some cases, YOLO outperforms Faster R-CNN in terms of accuracy, speed and efficiency. As YOLO performs single shot algorithms it is more preferable to be used in real time object detection whether it be in an image or a video. Its simple to construct and can train directly on full images. YOLO's better generalizing representation of objects as compared to Faster R-CNN makes it a more worthy, fast and robust algorithm to rely on. These bold advantages make this algorithm strongly recommended and stand out.

VI. REFERENCES

[1] A. M. A. ghani Abdulghani and G. G. Menekşe Dalveren, "Moving Object Detection in Video with Algorithms YOLO and Faster R-CNN in Different Conditions," European Journal of Science and Technology, Jan. 2022, doi: 10.31590/ejosat.1013049.

[2] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement." arXiv, Apr. 08, 2018. Accessed: Sep. 25, 2022. [Online]. Available: http://arxiv.org/abs/1804.02767

[3] L. Tan, T. Huangfu, L. Wu, and W. Chen, "Comparison of YOLO v3, Faster R-CNN, and SSD for Real-Time Pill Identification," In Review, preprint, Jul. 2021. doi: 10.21203/rs.3.rs-668895/v1.

[4] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, Jun. 2016, pp. 779–788. doi: 10.1109/CVPR.2016.91.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017, doi: 10.1109/TPAMI.2016.2577031.

[6] H. Jiang and E. Learned-Miller, "Face Detection with the Faster R-CNN." arXiv, Jun. 10, 2016. Accessed: Sep. 25, 2022. [Online]. Available: http://arxiv.org/abs/1606.03473

[7] J. Du, "Understanding of Object Detection Based on CNN Family and YOLO," J. Phys.: Conf. Ser., vol. 1004, p. 012029, Apr. 2018, doi: 10.1088/1742-6596/1004/1/012029.