IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

Object Detection

Ritvik Bhadola
B.Tech Student, Department of Computer Science and Engineering
Dronacharya College of Engineering, Gurgaon, Haryana, India

Abstract: *Target identification, one of the most important functions in machine learning, has been a research hotspot for the last 20 years and is widely used. Its goal is to quickly and accurately detect and reveal a huge number of elements in a given image that correspond to specific categories. Based on the model learning approach, the algorithms are classified into two types: single-stage detection algorithms and two-stage detection algorithms. The standard algorithms for each level are then described in this work. Following that, numerous sample methodologies are reviewed and contrasted in this domain, as are open and special datasets often used in target identification. Finally, challenges that may emerge while identifying targets are discussed.*

Keywords: Neural Networks, Object Detections, CNN, SPP, YOLO, mAP, R-CNN, recogniseds

I. INTRODUCTION

In fields like machine learning, deep learning, and artificial intelligence, object recognition is a popular topic. It is needed for operations liks target tracking, event detection, behaviour analysis, and scene semantic understanding. Its purpose is to locate the target of interest in an image, properly classify it, and calculate the bounding box for each target. Vehicle robotics, video and picture recovery, smart surveillance, medical image analysis, industrial inspection, and other industries have all made use of it. Pre-processing, window sliding, feature extraction, feature selection, feature classification, and post-processing are the six processes in traditional feature extraction methods, which are developed for specialised recognition tasks. Limited data capacity, low portability, lack of pertinence, high temporal complexity, window duplication, diversity resistance, and amazing performance only under a few fundamental circumstances are some of its major faults. Krizhevsjy and colleagues introduced the Alex Net convolutional neural network (CNN) image categorization model in 2012. Using the test procedure, they outperformed the second-place scorer in the ImageNet picture classification competition by 11%. A variety of potential approaches have been proposed by several scientists who have begun to employ deep convolutional neural networks to address recognition problems. There are two types of identification techniques: single-stage detection based on region proposal and two-stage detection regression models.

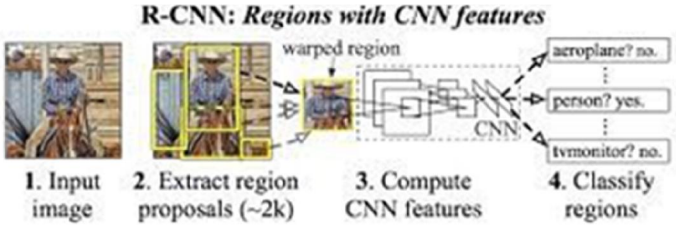
II. REVIEW OF LITERATURE

2.1 Framework for Two-Stage Target Detection

A. R-CNN

In 2014, Girshick announced the R-CNN approach, which would be the first realistic convolutional neural infrastructure target, identification model. The mAP of the modified R-CNN model is approximately 66%. The model utilizes Selective Search to generate around 2000 area suggestions for each image to be identified, as illustrated in Figure 1. The retrieved image features are then categorised using the SVM classifier after this recovered suggestion is evenly scaled to a corrected feature vector. To complete the bounding box regression process, a linear regression model is constructed. When compared to the traditional detection method, the R-CNN significantly improves accuracy; nevertheless, the quantity of processing required is substantial, and the calculating efficiency is low. There's also the question of size to consider. When compared to the traditional detection approach, the R-CNN dramatically improves accuracy; unfortunately, the degree of computational resources is enormous, and the computation efficiency is low. Second, immediately converting a region's recommendations to a corrected feature vector may induce object distortion.

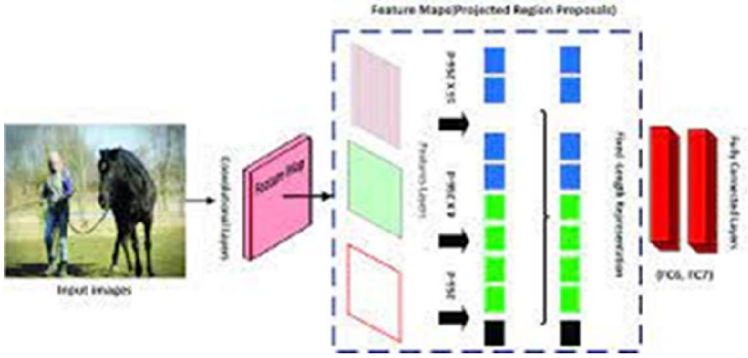
IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252



B. SPP-Net

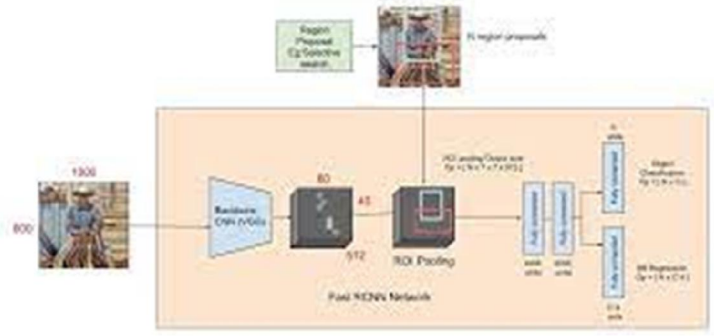
He proposed the Spatial Pyramid Pooling (SPP) model in 2015 to address concerns regarding R-low CNN detection rate and the necessity for fixed input size picture blocks. This technique extracts features of the area's proposition on the features mAP after the original image has gone through the convolution, and all convolution calculations are done just once. Simultaneously, the spatial pyramid aggregating layer is inserted after final convolution operation, and the feature of the area proposal is routed through it to extract the feature vector of predefined size. Unlike R-CNN, SPP-Net just conducts feature extraction on the full picture once, eliminating needless calculations. This does, however, have had the same disadvantages as R-CNN:

1. Mastering multi-step training methodologies is challenging.
2. Separate SVM classifiers, as well as extra regressors, must be trained.



C. Fast R-CNN

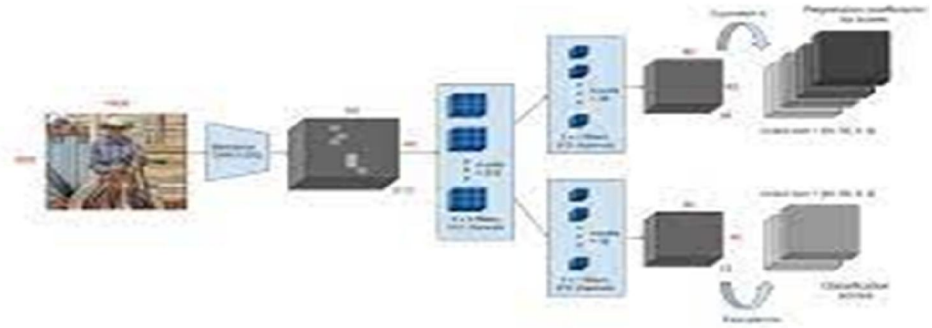
Girshick proposed the Fast R-CNN model in 2015. The mAP of the combined VOC2007 and VOC2012 datasets is 70.0 percent. Figure 2 depicts its structure. Fast R-CNN is distinguished from R-CNN in three ways. For classification, the SoftMax function was utilised instead of the SVM used in R-CNN. Second, to convert the candidate box's feature into a feature mAP with a defined size for access to the entire connection layer, the model uses the pooling layer from the SPP-pyramid Net and replaces the final pooling layer in the convolutional layer with the area of interest pooling layer. Finally, two parallel fully linked layers replace the CNN network's final SoftMax classification layer. It is, however, insufficient for real-time detection.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

D. Faster R-CNN

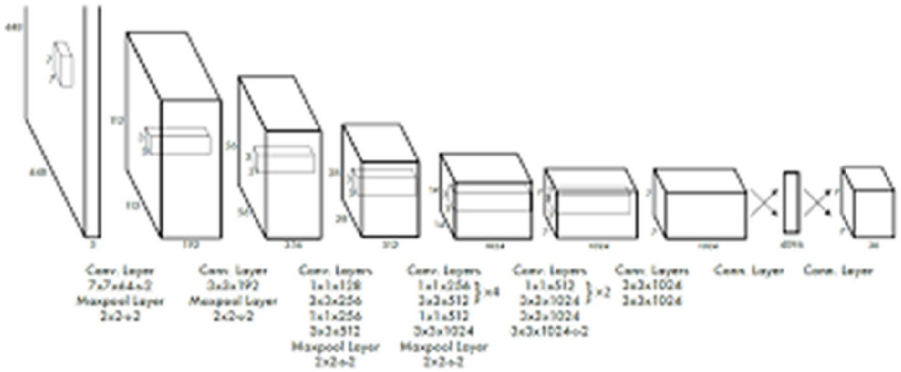
Instead of utilising the prior Selective Search technique, Ren's Faster R-CNN model offers region recommendations using region proposal networks. The model is divided into two parts: an all-region proposal created with a fully convolutional neural network and the Fast R-CNN detection approach. A set of convolutional layers is maintained by these two modules. The input image is sent through the CNN network and onto the Shared convolutional layer. The picture is transmitted forward to a particular convolutional layer to produce a higher dimensional feature mAP, and the feature mAP for the RPN network's input is acquired on the other hand. Despite its great detection accuracy, faster R-CNN is not capable of real-time detection.



2.2. Algorithms for One-Stage Target Detection

A. YOLOv1

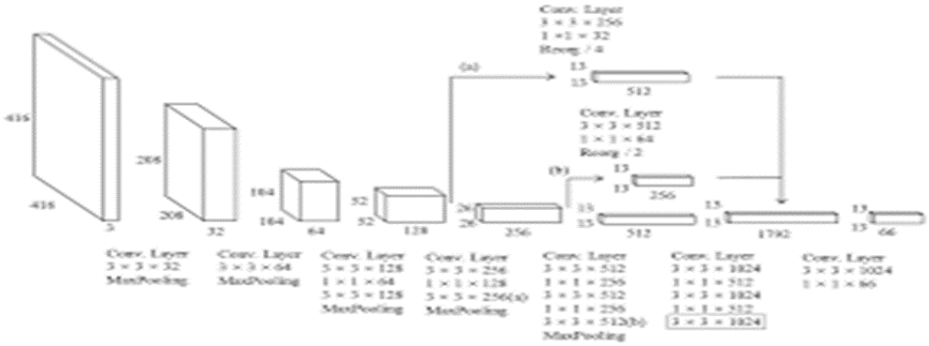
The YOLOv1 object recognition algorithm had first reported by Joseph Redmon in 2016. The YOLOv1 detection algorithm does not require the region proposal extraction procedure. The entire detection procedure may be reduced to a CNN network structure in its most basic form. The key concept is to feed the network the whole graph as input and deliver the bounding box's position and category directly to the output layer. Each picture is divided into a S*S grid, with each cell predicting the B bounding box as well as its confidence score. To put it another way, each cell predicts a total of four B*(4+1) values. A single TitanX can detect at 45 frames per second, enabling for complete real-time detection.



B. YOLOv2

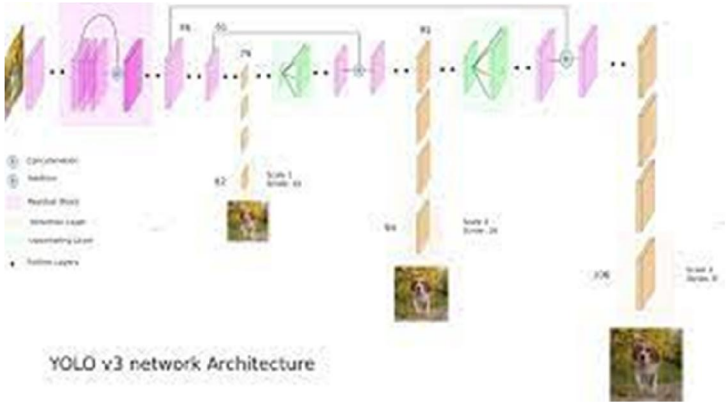
Redmon first introduced the YOLOv2 idea in 2016. The main goal is to improve memory and location while maintaining classification accuracy. In YOLOv2, Darknet-19 is employed, a new fully convolutional extraction of features network with 19 convolutional layers and 5 maximum pooling layers. Applying a batch homogenisation layer to the cnn model and decreasing dropout, as well as incorporating an anchor box mechanism, k-means clustering on the training images bounding box, and multi-scale training to improve recall and accuracy. On the other hand, identifying targets with the a lot of overlap and smaller targets has yet to be improved.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252



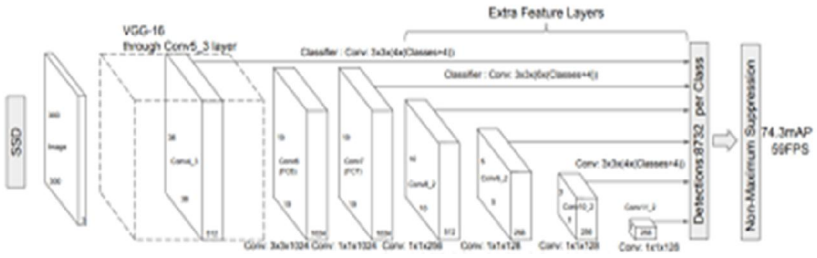
C. YOLOv3

Redmon's YOLOv3 is by far the best-balanced object detection model in terms of detection sensitivity and precision. The main objective of YOLOv3 in terms of category prediction is to convert single-label classifications to multi-label classifications and replace the original SoftMax layer for single-label multi-classification with either a logistic regression layer for a number of co multi-classification. Similarly, the model predictions on many scales. It employs a similar process to FPN's open merging and then combines three subscales to boost the detection impact of tiny things considerably. The network topology of this model was influenced by the Darknet-53 higher extraction of features network. The YOLOv3 model can improve detection speed and tiny object interaction, albeit at the expense of accuracy.



D. SSD

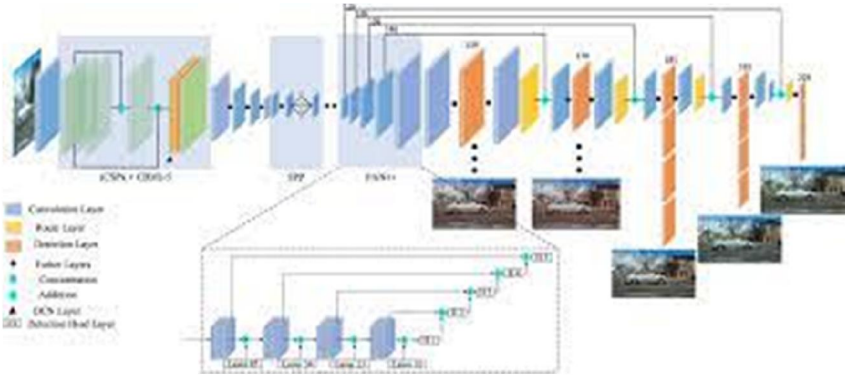
Liu proposed the SSD concept in 2016. The model combines the regression notion of the YOLO approach with the anchor box concept of the Faster R-CNN detection model. The SSD model recommends using both bottom and high-level feature mAPs for detection to maximise the effect of multi-scale object detection. The base is the VGG architecture, with convolutional layers replacing the last two fully linked layers. SSD makes use of the RPN network's anchor mechanism. On VOC2007, SSD gets 74.3 percent mAP at 59 frames per second on an Nvidia Titan X. However, for small targets, the SSD classification result is poor, and feature mAPs of different scales are independent, resulting in the identification of the same object by boxes of varying sizes at multiple scales at the same time.



IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

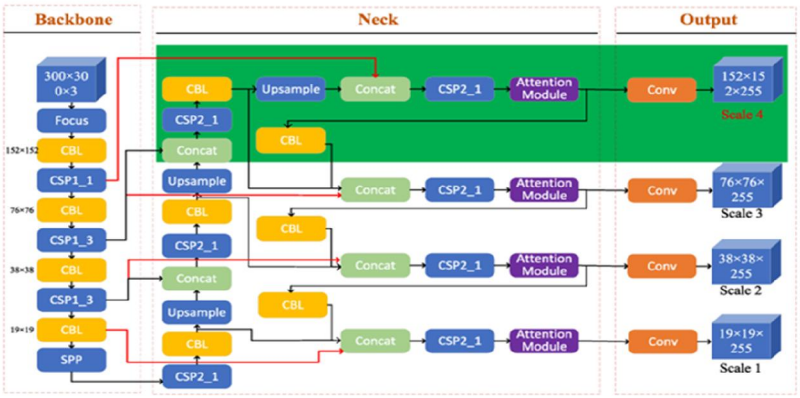
E. YOLOv4

In 2016, Liu developed the SSD technology. The model combines the YOLO approach's regression idea with the Accelerated R-CNN detection model's anchor box concept. To maximise the effect of multi-scale object detection, the SSD model advises integrating both bottom and high-level element mAPs for detection. The VGG architecture serves as the foundation, with convolutional layers replacing the last two fully connected layers. The RPN network's anchor technology is used by SSD. On an Nvidia Titan X, SSD scores 74.3 percent mAP at 59 frames per second resulting in the same item being classified by boxes of different sizes at required enhanced VOC2007. The SSD classification result



F. YOLOv5

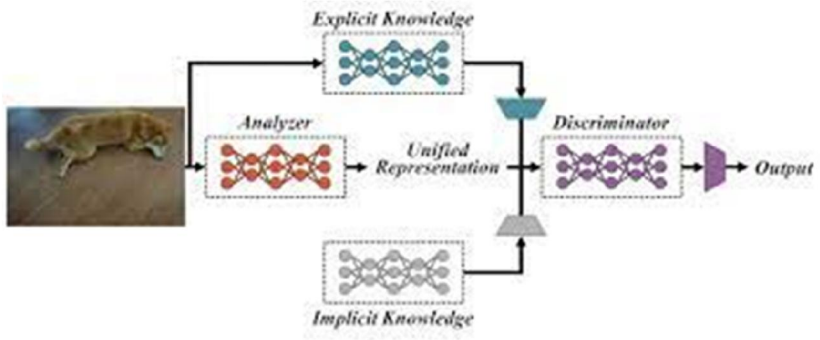
YOLOv5 increases data online by passing training data through a data loaders with each training batch. Scaling, colour space alterations, and mosaic enhancement are the three sorts of augmentations performed by the data loader. The tile data augmentation technology stands out when it combines four pictures into four random ratio tiles. Previously, the mosaic data loader could only be accessed through the YOLOv3 PyTorch repo, but it is now also available through the YOLOv5 repo. Mosaic is especially significant for the COCO object identification benchmark because it helps the model overcome the well-known "small object issue," which arises when little things are not identified as accurately as larger ones. Experimenting with your own set of modifications to increase efficiency on your particular mission might be useful. YOLOv5 generates model settings in yaml, as opposed to .cfg files on Darknet. The key difference between the two is that the .yaml file is cut to only show the network's specific levels, then multiplied by the block's layer count.



G. YOLOR:

The new YOLOR algorithm intends to complete projects for a sixth of the cost of competing algorithms. As a result, YOLOR is a unified network capable of processing implicit and explicit information at the same time and delivering a refined generic representation. Using cutting-edge techniques, the YOLOR achieved object recognition accuracy comparable to the Scaled YOLOv4 while increasing inference speed by 88 percent. As a result, YOLOR is one of the most efficient object recognition algorithms currently available. On the MS COCO dataset, YOLOR's mean average accuracy is 3.8 percent greater than PP at the same inference speed

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252



2.3 Various Algorithms Performance Comparison and Datasets

A. Dataset

In 1956, the term "artificial intelligence" was initially introduced. Artificial intelligence, on the other hand, took until 2012 to reach a tipping point. The introduction of machine learning techniques, as well as increased data volume and computer capabilities, are all contributing to this. The expansion of data volume is tightly linked to the progress of monitoring systems. This is related to the dataset's necessity in performance analysis and algorithm analysis, as well as in the progress of detection methodologies research. Table 1 lists the characteristics of common public data sets.

TABLE I. PUBLIC DATA SET AND ITS PARAMETERS

Dataset	Amount	Sort	Size/Pixel	Year
Caltech101[18]	9145	101	300×200	2004
PASCAL VOC 2007	9963	20	375×500	2005
PASCAL VOC 2012	11540	20	470×380	2005
Tiny Images[19]	80 million	53464	32×32	2006
Scenes15	4485	15	256×256	2006
Caltech256	30607	256	300×200	2007
ImageNet	14197122	21841	500×400	2009
SUN[16]	131072	908	500×300	2010
MS COCO[17]	328000	91	640×480	2014
Places[20]	More than10 million	434	256×256	2014
Open Images	More than 9 million	More than 60 million	Different size	2017

B. A Comparison of the Performance of Several Algorithms

The single-stage and two-stage detection approaches are compared and contrasted in Table II.

TABLE II. COMPARISON OF OBJECT DETECTION ALGORITHMS

Method	Backbone	Size/Pixel	Test	mAP/%	fps
YOLOv1	VGG16	448×448	VOC 2007	66.4	45
SSD	VGG16	300×300	VOC 2007	77.2	46
YOLOv2	Darknet-19	544×544	VOC 2007	78.6	40
YOLOv3	Darknet-53	608×608	MS COCO	33	51
YOLOv4	CSP Darknet-53	608×608	MS COCO	43.5	65.7
R-CNN	VGG16	1000×600	VOC2007	66	0.5
SPP-Net	ZF-5	1000×600	VOC2007	54.2	-
Fast R-CNN	VGG16	1000×600	VOC2007	70.0	7
Faster R-CNN	ResNet-101	1000×600	VOC2007	76.4	5

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 2, Issue 5, June 2022
Impact Factor: 6.252

III. CONCLUSION

In recent years, object identification has received considerable attention as one of the most fundamental and difficult subjects in computer vision. Deep learning spotting techniques are widely applied in a range of sectors, so they still face a few challenges:

1. Minimizing your reliance on data is a decent start.
2. To improve the speed with which little things may be recognised.
3. Object detection with multiple categorizations is now available

ACKNOWLEDGMENT

I would like to thank my teachers Mrs. Ashima Mehta and Mr. Ashwani Katoch who gave me this opportunity to work on this project. At last, I would like to extend my heartfelt thanks to my parents because without their help this project would not have been successful. Finally, I would like to thank my dear friends who have been with me all the time.

REFERENCES

[1]. https://pjreddie.com/darknet/yolo/
[2]. https://www.v7labs.com/blog/yolo-object-detection/
[3]. https://viso.ai/deep-learning/object-detection/
[4]. https://medium.com/ml-research-lab/what-is-object-detection-51f9d872ece7/

BIOGRAPHY

My name is Ritvik Bhadola residing in New Delhi. Currently pursuing my B.Tech from Dronacharya College of Engineering and working as Machine learning intern in Think Future Technologies. Presently I am assigned in the field of object detection which led me to the curiosity about how object detection evolved which led me to this research paper.