Journal of Theoretical and Applied Information Technology
15th June 2023. Vol.101. No 11

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A REAL-TIME HYBRID-YOLOV4 APPROACH FOR MULTI-CLASSIFICATION AND DETECTION OF OBJECTS

[1, *]SMITA RATH, [1, *]SUSHREE BIBHUPRADA B. PRIYADARSHINI, [2]DEEPAK KUMAR PATEL, [3]NARAYAN PATRA, [4]PRABHAT SAHU

[1, 1, *, 2,3,4]Assistant Professor, Siksha 'O' Anusandhan Deemed to be University, Computer Science & Information Technology, India

E-mail: [1,*]smitarath@soa.ac.in, [1,*]bimalabibhuprada@gmail.com,

[2]deepakpatel@soa.ac.in, [3]narayanpatra@soa.ac.in,[4]prabhatsahu@soa.ac.in

Corresponding Authors:

Smita Rath, smitarath@soa.ac.in,

Sushree Bibhuprada B. Priyadarshini, bimalabibhuprada@gmail.com

ABSTRACT

This paper improves the object detection accuracy for detecting objects in complex scenes and ensures real-time classification operations by planning a novel detection method called lightweight and efficient hybrid YOLOv4 model. In this context, Computational vision is one of the most useful and entertaining forms of artificial intelligence (AI) used in everyday life. Computer vision study focused on replacing intricate aspects of the human world with sophisticated AI and computers. Deep neural networks have recently become an essential part of several industries due to their renowned ability to handle visual input. One of the main directions that computer vision has taken is the domain of classification & tracking of objects employing neural networks, which are presently being employed by relevant trendsetting enterprises specializing in solving several arrays of predicaments such as security, health care, and agriculture. The main factors affecting the development of computer vision are the volume of data it generates, as well as the amount it utilizes to train and enhance it. In this paper, a method for categorizing and detecting objects utilizing an object detection algorithm namely hybrid-YOLOv4 is proposed. Convolutional neural networks provide extremely accurate object tracking and feature extraction out of the images. Strategies such as Bag-of-Specials and Bag-of-Freebies are used in item identification and DarkNet is used in the backbone that increases the feature exchange and reutilization. Thus, the improved network design maximizes both identification accuracy and speed. Additionally, two new extra blocks in the neck and backbone enhance feature extraction and reduce processing expenses. The model was compared with other object detections methods. According to the experimental findings, mean average precision (MPS) of YOLOv4-hybrid model was found to be 0.986 better than that of YOLOv4 and other object detection models.
Keywords: *Artificial Intelligence, Computer Vision, Faster RCNN, YOLOv4, RCNN*

1. INTRODUCTION

The branch of computer vision known as object detection is reviewed with the identification, placement, and categorization of things present in pictures and movies. It involves locating and identifying every occurrence of an object (such as vehicles, people, signs on street, etc.) in the specialty of vision. Typically, for specific classification methods, traditional detection techniques based on manually extracting features having six steps: pre-processing, window sliding, feature selection, feature extraction, feature classification, and post-processing. Small data sets, low portability, lack of pertinence, high time complexity, window redundancy, lack of robustness for diversity changes, and good performance only in certain simple situations are the primary drawbacks of traditional methods. Therefore, we can easily teach computers to detect and classify many items within a picture with high accuracy and availability of vast amounts of data, faster GPUs, and better models.

With this process of identification and localization, object detection is able to count the objects in a scene, spot and track their locations, and label each one of them precisely. Similar challenges have come up with a number of computer vision tasks, such as categorization, fragmentation, motion detection, scene perception, etc. The field of visual perception was altered by the arrival of convolutional neural networks (CNNs) for picture classification [1]. Currently, object detection is used for security, medical, self-driving cars, identity identification, and many other purposes. It has experienced an exponential expansion in recent years, along with the quick creation of new techniques and procedures.

Object tracking is the process of identifying and locating recognized objects in images that fall into a current set of classes. Specifically, object tracking indicates a computer vision strategy that aids in finding and locating objects in an image or video. We can locate these discovered objects within a given image by using a bounding rectangle that object detection produces around them. Image recognition assigns a label to a picture. The term "dog" refers to a snap about a dog. The term "dog" is employed to depict an image of two canines. On the contrary, object tracking envelops each dog in a box that reads "dog." The model forecasts the location of each object and the appropriate label. Furthermore, object tracking affords greater information about an image as compared to recognition. It assists in our comprehension conjointly with the investigation of scenes lying in photos or videos.

The challenges that an object detection works on are as follows:

• Localization and object classification are both given top priority. Model most frequently employ a multi-task loss function to address this problem by penalizing both classification and localization failures.

• It also needs to be incredibly quick at prediction time to meet the real-time needs of video processing

• Relevant objects may appear in a variety of sizes and aspect ratios for numerous applications of object detection. Practitioners use a variety of strategies to guarantee that detection algorithms can capture objects at different scales and angles.

• A variety of regions of interest can be produced by carefully selecting anchor boxes with different sizes and aspect ratios.

• Another significant barrier is the small quantity of annotated training data that is presently available for object detection.

The current article suggests a way for classifying and detecting objects employing a method for object tracking like YOLOv4 and existing CNN methods. A relatively high detection technique called You Only Look Once (YOLO) can accurately predict the position and characteristics of the object in an image. Convolutional neural networks are used to retrieve features out of photos and ensnare entities with hyper accuracy. In the early stages of automated object identification, a computer system can recognize, find, and identify an item from a provided image or video. Such object recognition strategies can be pre-trained and can be considered from the corresponding scratch, which represents a major part of computer vision technology.

To find out whether a split rectangular area retains an actual item, feature retrieval gets carried out in this context for every individual small segment produced in the picture and a large number of bounding rectangles that cover the entire picture. Boxes that superimpose get attached to create a bounding rectangle. The YOLO series type of model uses starting markers to identify the object region and forecast the category directly in order to completely perform object recognition without the use of RPN. As a result, although this type of framework is considerably rapid than two-stage models, it has lesser precision. One-stage models are more suitable in terms of speed. This study uses YOLOv4, a one-stage class model that can be used as the basis for solving the aforementioned three challenges, and adds certain modifications to it to fulfil the goal of real-time detection.

The semantic analysis of extracted features from low convolution layers seem to be typically relatively sparse but accurate at locating objects, whereas high-level schema is typically highly rich but imprecise in locating objects. We choose YOLOv4 because it inherits the advantage of YOLOv3, with the property to adopt combining and up-sampling operations to combine 3d feature maps, improving the precision of the algorithm for tiny objects.

The remaining portions of this research is arranged as follows: the subsequent section incorporates the related literature works done in this domain. Section 3 discusses the object detectors. Section 4 elaborates YOLOv4while analyzing the

results attained out of the experimentation. At Last, section 5 concludes the paper.

2. RELATED WORK

The R-CNN, Fast R-CNN, and YOLO architecture are being improved in this paper [2]. The work concentrated on extracting areas of interest from images by the help of the Region Proposals Network (RPN). Based on the object detection score, RPN generates a picture. In order to classify the output items, roll polling is used and it achieved a minimum GPU capability of 3.0. As their practical uses expand, the demand for lightweight versions that can be used on portable and integrated products will rise. In this paper [3], They offer a real-time compact metrics processing device that eliminates the need for image restoration. The system adopts pixel-wise coded exposure (PCE) and applies a deep learning method called You Only Look Once (YOLO) for object detection. Radio Detection and Ranging is referred to as radar which is used in [4]. It is an electromagnetic device with long-range object tracking, detection, and location capabilities. It operates by directing electromagnetic radiation towards so-called targets and then monitoring the echoes that are received. The targets may be moving cars, satellites, ships, planes, boats, spacecraft, birds, insects, rain, or any other object. Radars can monitor these items' presence, location, velocity, as well as their size and shape. Radars can too recognize far-off objects in adverse weather and precisely determine their ranges and distances.

While comparing to object detection only utilizing search approaches, Perez et al. proposed a method that provides adequate detection results for object detection tasks, with only 25% of the suggestions get graded to meet the reliability of selective search. This hybrid model is significantly more effective at categorizing images than normal methods because it provides greater information about the image's features [5]. CNN is used to predict key points rather than classifying and sectioning a man like an object, the desired result might be accomplished [6]. Recent developments in deep learning for removing objects of interest from photos are covered in the paper. R-CNN, as well as Fast R-CNN, are two of the approaches included in Najva and Bijoy's survey [7]. This thesis examined the performance of work area-based deep learning networks for semantic segmentation and object identity while using several core configurations. The program picks out images of tigers and provides them to wildlife activists so they may analyze and plan the best strategies to save the threatened species.

The FR-CNN is a cutting-edge method of tiger detection. It can be improved to give even better results, just like all other models created, including its predecessor, the Fast R-CNN[8]. The authors of such a study developed an object recognition and classification system that uses state-of-the-art machine learning techniques including BING and PCANet. BING can faster choose the target classes of things after being trained with several pictures containing the desired objects[9]. R-FCN conjointly with the Feature Pyramid Network demonstrates the deployment of position sensitivity in a fully-convolutional network as well as the transfer of relatively high meaning to the network's lower tiers.

The context-independent item identification layers may be introduced as the network depth increases and there are too new layers above it[10]. In this study, identifiable objects in films are identified using a special classification algorithm that makes use of deep and precise neural networks. The current challenge is aimed at identifying and classifying the presence of a person and a car in a video clip. The research outcomes are tracked and assessed by varying the whole count of hidden layers and the sum of neurons associated with every hidden layer. Auto-encoders get used to building DNN, and the effectiveness of the classifier is assessed [11-13]. Fig 1 provides an insight to various issues arising during tracking of object.

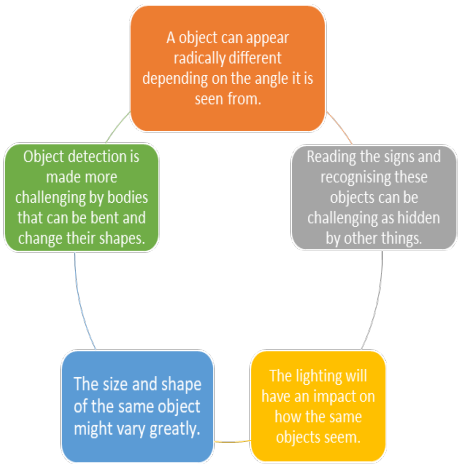


Fig. 1. Ultimatums Encountered in Object Tracking

2.1 BACKGROUND OVERVIEW

The object classification technique comes before the object detection procedure and focuses solely on identifying the objects in an image. Using axis-aligned boxes, object detection seeks out all occurrences of the predefined classes in the picture and finds all examples of those cases. All cases of the object classes are able to be recognized by the detector, and a bounding box should be drawn around them. It is typically regarded as difficult with supervised learning. Modern object detection algorithms may train on vast collections of annotated pictures and are assessed against a variety of canonical standards.

2.2 KEY ASPECTS IN OBJECT DETECTION

Some of the major difficulties that networks encounter in practical applications are as follows:

- **Variety of classifications:** It is a difficult problem to tackle because of the large set of object classes that are to be classified. In addition, there is a requirement for increasingly difficult to obtain high-quality labeled examples.A future work question is how to run a detector with fewer examples.
- **Inter-class variability:** In nature, there is a fair amount of intraclass variation among instances of the same object. Numerous factors, such as occlusion, illumination, position, viewpoint, etc., could be to blame for this difference. These unrestricted exterior elements have a significant impact on how an object seems [14]. The objects are anticipated to deform non-rigidly, and they rotate, scale, or become fuzzy. Some things could be surrounded by unnoticeable elements, making extraction challenging.
- **Efficiency:** Modern models need a fast processor for effective detecting capabilities. Reliable object monitors are vital for continuing research in the area of computer vision.

2.3 BACKBONE ARCHITECTURES

The backbone structures of the object detector are among its most important components. These networks take various aspects of the input image taken within the model.

2.3.1 AlexNet

AlexNet [15], a CNN-based model for image classification that was designed by Krizhevsky et al. Around 15 million high-resolution pictures and 22,000 categories create the ImageNet dataset. Compared to modern models. Further, it secured a significantly higher accuracy (more than 26%). Three fully connected layers with five convolutional layers make up AlexNet's eight learnable layers. The final level of the fully connected layer is paired to an N-way softmax classifier.

2.3.2 GoogleNet

In 2014, Google researchers introduced Google Net in the research article namely "Going Deeper with Convolutions" with the help of several institutions. It bagged a prize in 2014 ILSVRC image classification competition. Compared to prior winners, it has offered a significantly lower mistake rate. AlexNet. The entire pattern consists of 22 layers. The pattern was produced using efficient computation. The model can be used with limited processing resources[16].

2.3.3 EfficientNet

Tan et al. conducted a thorough investigation on network scalability and how it affects model accuracy in [17]. Altering model features such as width, depth ,and resolution could affect accuracy. They demonstrated the expense associated with scaling any factor independently.

Richer and more complex features can be captured by a network with more depth. But due to the vanishing gradient issue, they are difficult to train. Similar to this, narrowing the network will allow it to be simpler to collect fine-grained characteristics but more challenging to acquire high-level information. The advantages of growing image resolution, such as depth and width, become saturated as the model scales. A straightforward and effective architecture is EfficientNet. It operated faster and more accurately than current models despite being significantly smaller. It may start a new era in the study of effective networks by offering a massive gain in efficiency.

2.3.4 CSPNet

When using the Cross Stage Partial Network (CSPNet) method to DenseNet, we create CSPDenseNet, a convolutional neural network, and object detection backbone. The CSPNet divides the basic keyframe feature map into two sections, merging the two using a cross-stage hierarchy [18]. A split-and-merge technique promotes greater gradient flow throughout the network.

2.3.5 DarkNet

An open-access neural network system is known as Darknet[19]. It is a quick and extremely reliable framework for real-time object detection based on training data, epochs, and batch size, and can be used for images.

3. OBJECT DETECTORS

3.1 REGION-BASED-CNN

A three-module object-detecting system is used. The first produces regional ideas that are independent of category. The candidate detections that our detector is capable of detecting are defined by these ideas. A neural network can extract a stationary feature vector from every segment. A group of class-specific linear SVMs makes up the third component.

Fig. 2 illustrates the procedure used by R-CNN[20] to categorize the entities by creating a bounding rectangle with the concerned picture. A class-independent region proposal module is created using R-CNN. This module employs selective search to pinpoint regions of the image where finding an item is more likely. The CNN network then warps and propagates these candidates, extracting a 4096-dimension feature vector for each proposition.

3.2 FRAMEWORK OF FAST R-CNN

The requirement to train various systems independently was among the main problems with R-CNN/SPP-Net. This was resolved by Fast R-CNN by developing a single end-to-end system. The network receives an input as an image. Then object suggestions are mapped to the feature maps that are processed through a series of convolution layers. The pyramidal framework of the SPP-net was replaced by a single spatial bin known as the RoI pooling layer by Girshick et al[13]. This layer is passed through two completely connected layers prior to proceeding out into a bounding box layer with a fully connected layer and an N+1-class SoftMax layer.

The Fast R-CNN detector investigates the whole picture in contrast to the concerned R-CNN detector, which reduces and scales up such suggestions; Fast R-CNN considers CNN properties concerning every regional plan, while an R-CNN detector must classify every field. Fast R-CNN was introduced as a supplement to R-CNN and as a speed improvement (146x) over R-CNN. It eliminated pyramidal pooling, thus, streamlined the training process, and introduced a new loss function. Without the region proposal network, the object detector recorded with high accuracy and close to the real-time speed (Fig. 3).

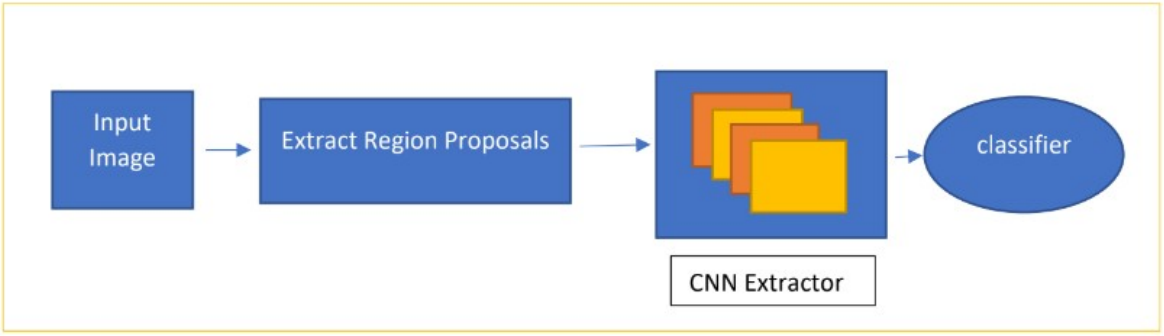


Fig.2. Steps of pictures employing R-CNN

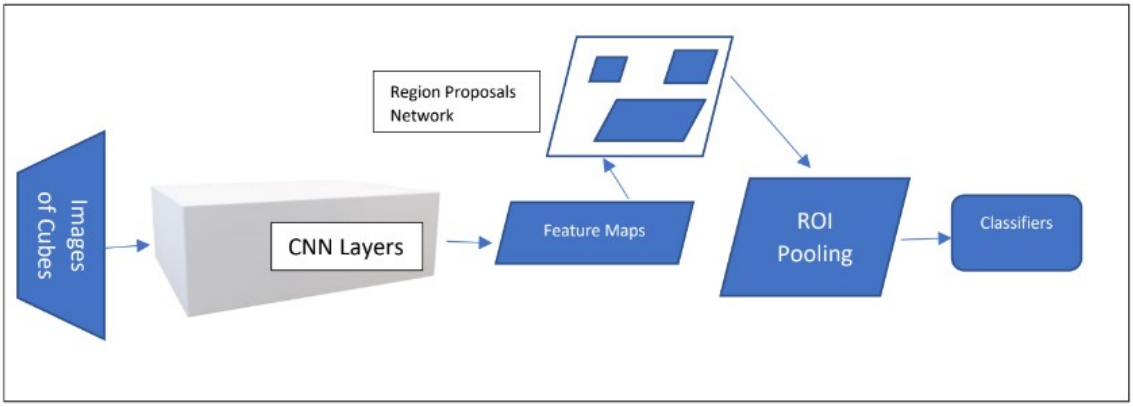


Fig.3. Fast-RCNN Framework

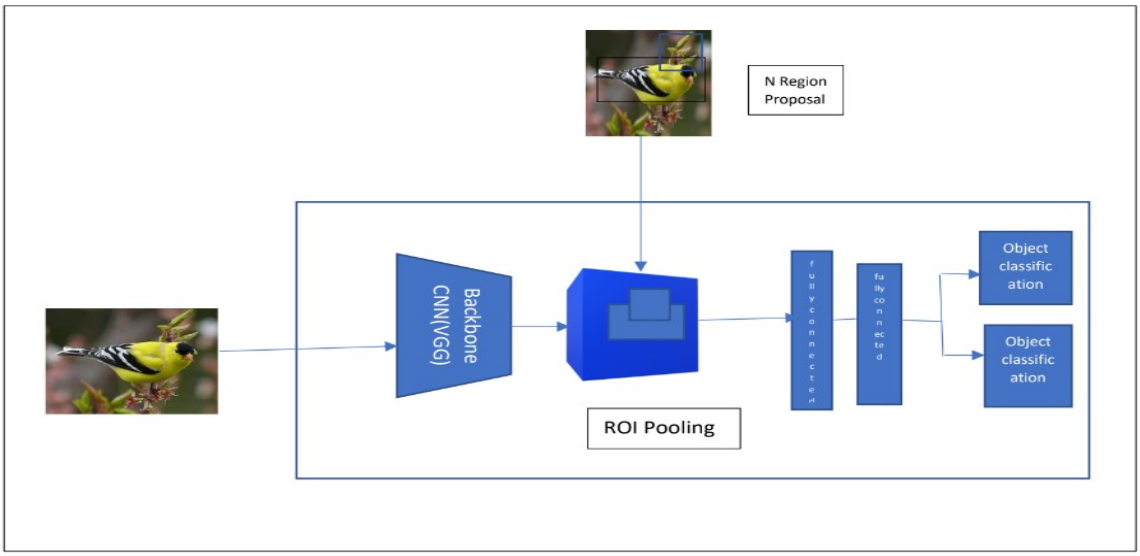
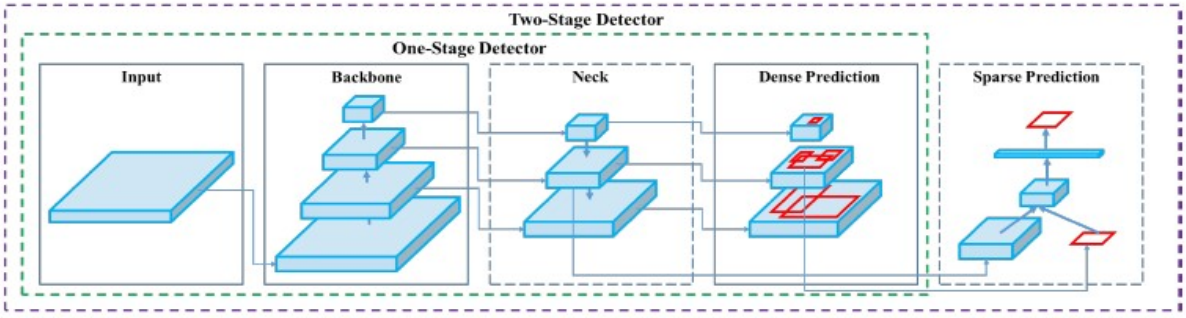


Fig.4. Segments in FR-CNN



Input: { Image, Patches, Image Pyramid, … }

Backbone: { VGG16 [68], ResNet-50 [26], ResNeXt-101 [86], Darknet53 [63], … }

Neck: { FPN [44], PANet [49], Bi-FPN [77], … }

Head:

Dense Prediction: { RPN [64], YOLO [61, 62, 63], SSD [50], RetinaNet [45], FCOS [78], … }

Sparse Prediction: { Faster R-CNN [64], R-FCN [9], … }

Fig. 5. Working Model of YOLOv4[24]

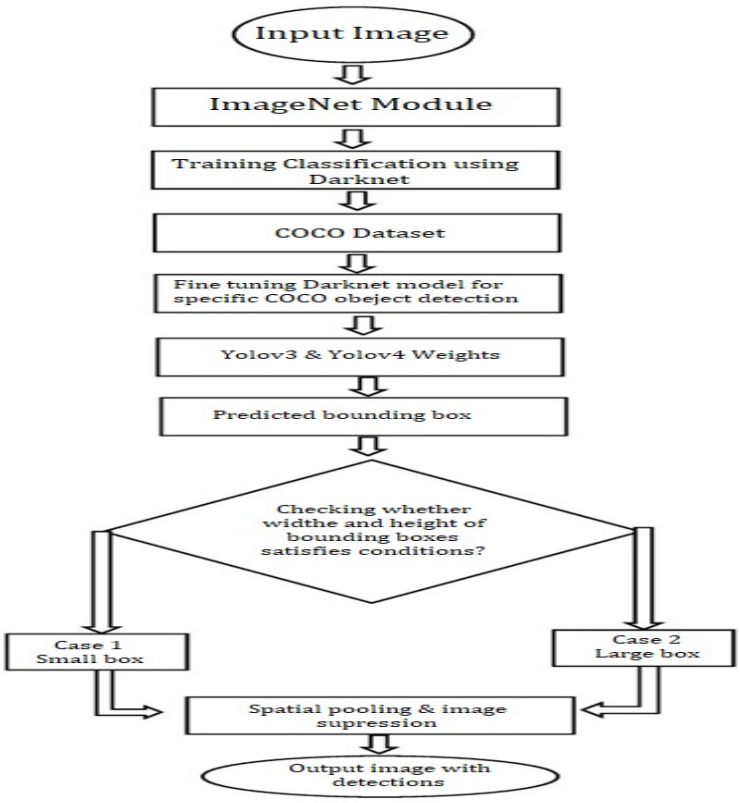


Fig..6 . Working Steps of Object Detection

Sequence of Steps involved in Object Detection:

- Import GPU processing components (CUDA, CUDNN, LIBSO)
- Import Darknet repository from AlexiyAB's
- Fit the data using COCO classes and model for dataset.
- Importing COCO (yolov4-csp, coco-data, yolov4-weights) weights.
- Importing and customizing Darknet functions for Object detection.
- The JavaScript function will resize the image before the CNN algorithm has a chance to work on the image, this will improve the performance of the whole Object Detection Model
- Optimize the COCO dataset and weights using helper functions.
- Using JavaScript plug in webcam to view and capture input images.
- Capture image will run through the darknet model and pass-through COCO weights and form bonding boxes.
- Display the analysed image form the webcam and save and display output using JavaScript helper function.

Fig. 7. Flow Of Steps Involved In Object Detection

3.3 FRAMEWORK OF FASTER R-CNN(FR-CNN)

In place of deploying an outer strategy such as Edge Boxes, the FR-CNN detector incorporates an area proposal network (RPN) to build zonal proposals in the network. Moreover, anchor boxes get applied through the RPN for ensuring various entities. The network forms zonal suggestions more rapidly as well as efficiently for concerned data. An FR-CNN entity tracker learns by employing the Faster RCNN Object tracker. FR-CNN Object Detector that will find out segments in a snap is the method's outcome. Using an RPN network, FR-CNN analyses the image in the stages shown in Fig. 4. Because of the shared layers among two models that carry out very distinct tasks, the training of the model is more complicated[21] by more than 3% while reducing inference time by a factor of scale. It ran at 5 frames per second in almost real-time after removing the bottleneck caused by the slow region idea. A CNN in-region idea had the added benefit of being able to improve over time, which would increase accuracy.

3.4 SSD

The Single Shot MultiBox Detector (SSD) [22] forms the initial single-stage detector like Faster R-CNN whereas saving real-time speed. Using extra auxiliary structures to boost performance, SSD was constructed on VGG-16 and the dimension of these additional convolution layers progressively decreases. When the picture characteristics are not sufficiently basic, the more complex layers handle the offset of the default boxes while SSD detects smaller items in the network more quickly.

4 PERFORMANCE ASSESSMENT

.

Experts in computer science and AI regularly use the COCO dataset associated with a variety of research problems. This Object Detection Model gets investigated by applying the Microsoft-advocated MS COCO data, a huge object recognition, and segment, conjointly with the labeling dataset. The image dataset COCO, which represents Common Objects in our domain, was chosen to emphasize the goal of improving image identification. The COCO dataset contains sophisticated, top-notch visual computer vision datasets that are largely applied by neural networks. The Darknet is one such C and CUDA-based open-source neural network architecture that serves as the basis of YOLO.

It is quick, easy to set up, and handles both CPU and GPU computation. Darknet acts as YOLO's training ground, laying the foundation for the network's organization. You Only Look Once is abbreviated as "YOLO," and YOLOv4 represents the fourth member of the group of YOLO object-detecting devices. This foundational framework helped in building YOLO's prominence and reputation in the vision community. This got recognized through the notion of bag of specials (BoS) and bag of freebies (BoF) strategies for enhancing the quality of the model. Modern one-stage recognizer YOLOv4 beats existing detectors in both speed and accuracy.

4.1 Working of YOLOv4

An image and the actual numbers for the bounding boxes are provided as input. The complete input image is divided into a square grid, and each object is detected in the grid cell that contains its center. The B bounding boxes and the corresponding confidence ratings will be predicted for each grid cell. This rating shows how accurate the box is and how certain the model is that the item is inside the box. The confidence number is the IoU of the predicted values and actual values of the boxes. YOLO [23] replans it as a case of regression and successfully predicted both the bounding box characteristics of the picture and the status of the object.

The framework comprises of various divisions as shown in Fig.5. - The input set of training images will be fed to the network – then they are trained in batches in parallel by the GPU. The Backbone (CSPDarknet53) and the Neck(SPP and PAN) perform the grouping and extraction of features. Both the Neck and Head together forms the Object Detector and finally, the Head does the detection/prediction. The Head is responsible for the detection (both localization and classification). The DenseNet architecture is the foundation of CSPDarkNet53. Before entering the dense layers, it combines the prior inputs with the current input; this is known as the Dense connectivity pattern. The CSPDarknet53 makes two layers Convolutional Base Layer and CSP i.e.Cross Stage Partial Block.

The S x S grid in YOLO decomposes the image into cells, that contains the center of object as the object's detection point. Each prediction array for a grid cell's predicted bounding boxes has five elements: the box's center, its dimensions (w and h), and its confidence level as shown in Fig 5. As seen in Figure 5, the Yolo method divides the whole image into cells that can only be 19 by 19.

After then, each cell will be in charge of forecasting its bounding rectangle, which offers a better level of precision. A non-max suppression will combine the bounding rectangle by calculating the joining surface and forecasting the bounding rectangle to classify recognizable images.

Libraries are imported initially (base64, NumPy, cv2, JavaScript, and Ipython. display). Fig. 6 depicts the flowchart model to detect object detection. Fig. 7 illustrates the pseudo-code of steps involved in object detection. An object detector that could execute in current production systems was designed by YOLOv4 [24] by incorporating many exciting ideas. It uses a "bag of freebies," or techniques, that takes a long training time but have no impact on inference time. To enhance training, YOLOv4 engages a variety of strategies, including class label smoothing, CIoU-loss, Cross mini-Batch Normalization and self-adversarial training.

The network is also expanded with techniques known as "Bag of Specials," which only affect inference time. These techniques include Mish activation [25], Cross-stage partial connections (CSP) [26], SPP-Block [27], PAN path aggregated block [28],etc. Additionally, it searches for hyper-parameters using a genetic approach. The issues of imbalanced background categorization and significant imbalance between positive and negative data in a one-stage model can be resolved using Focal loss. It lessens the importance of many simple negative samples in training, which is also referred to as a challenging form of sample mining. Focal loss can be defined as in equation (1)

$$L_{fl} = \begin{cases} -\alpha(1-y')^{\beta} log y', when\ y=1 \\ -(1-\alpha)y'^{\beta} log(1-y'), when\ y=0 \end{cases} \quad (1)$$

where the purpose of the "factor" β is to have the loss function focusing more on the challenging and incorrectly categorized samples while reducing the loss of easily identified examples. To correct for the uneven distribution of positive and negative samples, the equilibrium factor α is added. When this is only taken into account and the percentage of positive samples is low, that can be calculated as 0.75. The purpose is to prevent the classification loss function from becoming too tiny and thus can be taken to be 0.25 along with regulating to achieve a relative balance of β and α factors.

4.2 Outputs and Discussion

The open-source platform Darknet framework was used to develop the technique in this paper. The configuration of the computing workstation included 1080 Ti GPUs,4 GeForce GTX an Intel Xeon E5-2620v4 8-core, 128 G RAM and 2.1 GHz, 20 M.

Equations (2) and (3) report the analytical representation associated with precision as well as recall.

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP+FN} \quad (3)$$

TP = No of positive counts
FP = No of false positive counts

Detection Speed: Another important evaluation metric for detecting an object detection in real-time detection .

Frames Per Second (FPS): FPS is an important parameter to measure the detection speed, and it indicates the count of objects that the algorithm is capable to notice in a second. Following the investigation of various Objection Models associated with our recent hardware lacuna, Yolov4 was affording us with a required quantity of average precision as well as FPS.

Table 1:Performance Of Various Tracking Models

Models	Frames per second	Mean Average Precision	Detection Speed
F- RCNN	3.7	0.925	275ms
YOLO-v4	60.2	0.976	16.47ms
YOLO-v4-hybrid	110.56	0.986	16.01ms
YOLO-v3	61.3	0.974	19.67ms
SSD	5.8	0.883	178.6ms

A picture comprising the several classes are associated with the MS COCO dataset. Vehicles as well as tiny objects such as handbags, books, Laptops, etc. have been employed as the training data for assessing the YOLOv4. Further, the outcomes associated with the test undertaken are outlined below. Individual sections of the test snaps got investigated for gathering the TP, TN, FP, FN required to find out the correctness and MAP in the concerned framework. The investigation about the snaps ensnared through our webcam got recognized

in Fig. 8 and Fig. 9 and as well as classes conjointly with the reliance values were allotted.

The following are the fundamental training methods for YOLOv4 and hybrid YOLOv4-F: Both the two models were trained using a Cn series. Both the learning rates were originally set to 0. With the modification technique having 1000 iterations, the learning rate would increase to 0.001 and then stay the same. The learning rate dropped by 10 times when there were 30,000 iterations. After reaching 46,000, the learning rate dropped by further 10 times. The dropout rate was put to 5%, the confidence level was 0.5, the momentum value was 0.8, the weight degradation regular term was 0.0015, the size of the batch was 64, the batch division was 16, the count of rounds was 50,000.

Fig. 8. Outcomes Of Images Employing Yolov4

Fig. 9. Object Ensnared Within Human Images

By observing the images, this is obvious that the model was quite capable of capturing the object classes out of distinct angles and distances with maximal confidence. Further, the tests were carried out several times by affording the trapped image taken out of the webcam employing the MS COCO dataset via JavaScript's functions.

Table 1 illustrates the Map and FPS and detection speed of various object tracking models conjointly with the concerned outcomes in the concerned framework comprising of 8 Gb RAM along with Nvidia 1650 Ti GPU. Even while CSPDarknet53 exceeds the network size, the SPP block escalates the count of associations between the pixel point and the final actuation up to the network size while making it possible to observe the object's context. It slightly slows down the network speed while greatly expanding the receptive region and thus isolating the most important context elements. Model have attained parameters from distinct backbone levels for different detector levels while employing the PANet strategy.

5. CONCLUSIONS

A novel object identification framework called YOLOv4-hybrid is put forth in this work for real-time object recognition. This framework uses optimized techniques for the loss function and enhancement of the objects based on YOLOv4. The effectiveness of YOLOv4-FPM is assessed in comparison to the current models. The outcomes demonstrate that our model is capable of overcoming the difficulty posed by a tangled background, and the detection processing speed is sufficient to meet the real-time performance of the classifier.

Object detection employs both classification and localization tasks to analyze more realistic scenarios where multiple items may be visible in a photo. CSPDarknet53 acts as backbone of YOLOv4. A CSP block in CSPDarknet53 uses Cross-stage hierarchy to divide the feature map in the base layer into two sections and merge them, while allowing additional gradient to flow through the layers and conjointly addressing the dreaded "Vanishing Gradient" issue. In a nutshell, YOLOv4 is a condensed version of a broad range of computer vision methods for object detection. The best real-time object detector in the game is made from these tested and enhanced methods, and it is portable and simple to use. Still the proposed approach can be further extended to focus on

detecting very close objects as each grid can propose two bounding boxes.

ACKNOWLEDGEMENT

The authors are highly grateful to the Computer Science and Information Technology department, Siksha 'O' Anusandhan University for creating this exploration outstanding.

REFERENCES:

[1] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Communications of the ACM, 2017, 60(6), pp. 84-90.

[2] S. M. Abbas and S. N Singh, "Region-based object detection and classification using faster R-CNN", 4th International Conference on Computational Intelligence & Communication Technology (CICT), IEEE, 2018, pp. 1-6.

[3] C. Kwan, et al., "Real-time and deep learning-based vehicle detection and classification using pixel-wise code exposure measurements", Electronics, 2020, 9(6), pp. 10-14.

[4] V. Bamane, J. Sapkale, A. Pawar, P. G. Chilveri, N. Akhter, N., A. A. B. Raj, "A Review on AI Based Target Classification Advanced Techniques", 2022, 10(4), pp: 88-99.

[5] R . Pérez, et al., "Deep-learning radar object detection and classification for urban automotive scenarios", Kleinheubach Conference, IEEE, 2019, pp. 1-4.

[6] T. Chen, et al., "Road marking detection and classification using machine learning algorithms", IEEE Intelligent Vehicles Symposium (IV) , IEEE, 2015, pp. 617-621.

[7] J. Redmon, et al., "You only look once: Unified, real-time object detection", In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779-788.

[8] N. Najva, K.E. Bijoy, "SIFT and tensor-based object detection and classification in videos using deep neural-networks", Procedia Computer Science, 2016, pp. 351-358.

[9] J. Hung, A. Carpenter, "Applying faster R-CNN for object detection on malaria images", In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2017, pp. 56-61.

[10] D. Yi, et al., "Probabilistic faster R-CNN with stochastic region proposing: Towards object detection and recognition in remote sensing imagery", Neurocomputing, 2021, 459, pp. 290-301.

[11] B. Liu, et al., "Study of object detection based on Faster-R-CNN", 2017 Chinese Automation Congress, 2017, pp. 6233-6236.

[12] R. Gavrilescu, et al., "Faster R-CNN:an Approach to Real-Time Object Detection", International Conference and Exposition on Electrical And Power Engineering (EPE), 2018, pp. 0165-0168.

[13] R. Girshick, et al., "Rich feature hierarchies for accurate object- detection and semantic segmentation", In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580-587.

[14] J. Zhang, et al., "Multi-class object detection using faster R-CNN and estimation of shaking locations for automated shake-and-catch apple harvesting", Computers and Electronics in Agriculture, 2020, pp. 53-84.

[15] S. H. Lee, C. H. Yeh, T. W. Hou, C. S. Yang, "A lightweight neural network based on AlexNet-SSD model for garbage detection", In Proceedings of the 2019 3rd High Performance Computing and Cluster Technologies Conference, 2019, pp. 274-278.

[16] K. Chaitanya, G. Maragatham, G., "Object and obstacle detection for self-driving cars using GoogLeNet and deep learning", In: Artificial Intelligence Techniques for Advanced Computing Applications: Proceedings of ICACT, Springer Singapore, 2021, pp.315-322.

[17] M. Tan, Q.V. Le, "EfficientNet: rethinking model scaling for convolutional neural networks", arXiv:1905.11946.

[18] C. Y. Wang, H.Y. M. Liao, Y. H. Wu, P. Y. Chen, J. W. & I. H. Yeh, , " CSPNet: A new backbone that can enhance learning capability of CNN", In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, 2020, pp. 390-391.

[19] M. Mahrishi, S. Morwal, A. W. Muzaffar, S. Bhatia, P. Dadheech, M. K. I. Rahmani, Video index point detection and extraction framework using custom YoloV4 Darknet object detection model. IEEE Access, 9, 2021, pp. 143378-143391.

[20] J. Dai, Y. Li, K. He, J. Sun, "R-fcn: Object detection via region-based fully convolutional networks", Advances in neural information processing systems, 2016, 29.

[21] R. Girshick, "Fast R-CNN", in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 1440–1448.
[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, A. C. Berg, "Ssd: Single shot multibox detector. In Computer Vision–ECCV", 14th European Conference, Amsterdam, The Netherlands, Springer International Publishing, pp. 21-37, 2016.
[23] T. Diwan, G. Anirudh, J. V. Tembhurne, "Object detection using YOLO: Challenges, architectural successors, datasets and applications", Multimedia Tools and Applications, 2022, pp. 1-33.
[24] A. Bochkovskiy, C. Y. Wang, H. Liao, "Yolov4: Optimal speed and accuracy of object detection", 2020.
[25] D. Misra, "Mish: A self regularized non-monotonic activation function", arXiv preprint arXiv:1908.0868, 2019
[26] Y. F. Lu, Q. Yu, J. W. Gao, Y. Li, J. C. Zou, H.Qiao, "Cross stage partial connections based weighted Bi-directional feature pyramid and enhanced spatial transformation network for robust object detection", Neurocomputing, 2022, 513, 70-82.
[27] A. M. Roy, R. Bose, J. Bhaduri, J., "A fast accurate fine-grain object detection model based on YOLOv4 deep neural network", Neural Computing and Applications, pp. 1-27, 2022.
[28] P. Kalgaonkar, M. El-Sharkawy, M., "NextDet: Efficient Sparse-to-Dense Object Detection with Attentive Feature Aggregation. Future Internet", 2022, 355, 14(12).