Designing A Deep Learning Model to Detect Objects

CHANDAN VISHWAKARMA[1], SWAPNIL SONAWANE[2], NIKHIL MAHADIK[3], DR. J. W. BAKAL[4]

[1, 2, 3] *Student, Dept. of Information Technology Engineering, SSJCOE, Maharashtra, India*

[4] *Principal, Shivajirao S. Jondhale College of Engineering, Maharashtra, India*

Abstract- Object detection is to identify objects in the image along with its localization and classification. This paper deals in the area of computer vision, mainly for the application of deep learning in the object detection task. On the one hand, there's a straightforward summary of the dataset and deep learning algorithms commonly utilized in computer vision. There's a literature survey of papers containing different current approaches like faster r-cnn, Yolo, etc. A literature survey is represented in tabular form with our inference.

Indexed Terms- Object Detection, Dataset, Convolutional Neural Network, Computer Vision.

I. INTRODUCTION

Predicting the class of the object together with the location is called object detection. A Convolutional Neural Network is a Deep Learning algorithm that can be used to assign importance to different objects in the image. It can be able to differentiate one from the other. CNN requires less pre-processing than other classification algorithms. Algorithms inspired by the structure and performance of the brain is called deep learning.

A. Dataset

Dataset is one of the foundations of deep learning, some commonly used datasets in computer vision are mentioned.

a. ImageNet

The ImageNet dataset has over 14 million images covering over 20,000 categories. There are over 1,000,000 pictures with explicit class annotations and annotations of object locations within the image. The ImageNet dataset is one amongst the foremost widely used datasets within the field of deep learning. Most of the research work like image classification, location, and detection is predicated on this dataset. There is a well-known challenge called "ImageNet International Computer Vision Challenge" (ILSVRC) supported the ImageNet dataset.

b. PASCAL VOC

The PASCAL VOC (pattern analysis, statistical modelling and computational learning visual object classes) provides standardized image data sets for object class recognition and provides a common set of tools for accessing the data sets and annotations.

c. COCO

COCO (Common Objects in Context) could be a new image recognition, segmentation, and captioning dataset, sponsored by Microsoft. COCO dataset has over 300,000 images covering 80 object categories.

B. Approaches

There various approaches available in deep learning for object detection

a. R-CNN

R-CNN is the convolution neural network based on the region proposal brought up in 2014 by Girshick who came up with the concept of region proposal for the first time. The principle of R-CNN is that it utilizes the region segmentation method of selective search to extract the region proposals in the image, which include the possible object candidates, and loads them into convolution neural network to extract the feature vectors. Later, the classifier SVM will be used to classify the feature vectors to obtain the classification results in each region proposal. After merging by non-maximal suppression (NMS), the model outputs the precise object classifications and object bounding boxes to achieve object detection.

b. SPP-net

SPP-net is a deep neural network based on the spatial pyramid pooling proposed by MSRA He in 2014. The spatial pyramid pooling layer can get rid of the crop/warp operation on the input image in the former method. And it enables the input images of different sizes to connect with the full connection layer with the feature vector of the same dimension after passing the convolution layer. Although SPP-net solves the problems of object image incompleteness and object deformation, it is still of colossal computation and poor real-time because its image processing is similar to that of R-CNN.

c. Fast R-CNN

Fast R-CNN is the upgrade of R-CNN proposed by Girshick and has the capability to solve the repetitive calculation problem of the 2000 region proposals passing through the convolution neural network in turn. The improvement of Fast R-CNN compared to R-CNN lies in that it maps the region proposal extracted by selective search algorithm in input image to the feature layer of convolution neural network and conducts the pooling on the mapped region proposal of feature layer by ROI pooling. The ROI pooling can help Fast R-CNN obtain the feature vector of fixed sizes, which is necessary to successfully connect with the full connection. The role of ROI pooling is just like the spatial pyramid pooling of SPP-net. The method of mapping region proposal of input image to the feature layer in Fast R-CNN shares the convolution computation, which substantially reduces the calculation. In addition, in order to decrease the parameters of full connection, Fast R-CNN adopts truncated SVD to enable that the single fully connected layer corresponding to weight matrix is replaced by two small fully connected layers, which further lessens the network calculation.

d. Faster R-CNN

Faster R-CNN, proposed by Ren, He, Girshick, et al., is the upgrade version of the Fast-CNN. Faster R-CNN employs the region proposal network (RPN) to solve the issues of huge computation and poor real-time caused by the selective search method in R-CNN and Fast R-CNN. And Faster R-CNN is an end to end framework which can train the model easier. The function of RPN in Faster R-CNN is to replace the role of selective search in obtaining region proposals. RPN could divide the feature layer into n×n regions and obtain the feature regions of various scales and aspect ratios that are centered on the region, and the method is called anchors mechanism. The anchors in RPN are used to produce object proposals and then the proposals are sent to the rear classification and regression networks for object recognition and location.

e. R-FCN

R-FCN proposed by Dai, is a full convolution neural network based on regions, having solved the problem that RoI can’t share the computation. The object detection framework of R-FCN also adopts RPN to generate candidate RoIs. With the position-sensitive score maps (k*k*(C+1) dimensional convolution layer), R-FCN can record the response of every object in different locations. R-RCN defines the feature vector (C+1-dimension column vector) by voting according the RoIs and adopts softmax classification to classify the feature vectors in order to achieve the object recognition. Moreover, the object location could be realized by appending a 4 * k * k dimension convolutional layer to the above position-sensitive score maps and defining the feature vector (4-dimension column vector that represents the coordinates and width-height (tx, ty, tw, th) in the RoI region) by voting according to the RoIs. B. Models based on regression at present, the object detection methods based on deep learning using region proposal gets satisfactory achievements, but the object detection is still of poor real-time that can't satisfy the application requirement.

f. YOLO

YOLO came up with by Redmon, Divvala, Girshick, et al., is a convolution neural network for real-time object detection and can accomplish end to end training. Because of the cancel of RoI module, YOLO won’t extract the object region proposal any more. The front end of YOLO connects a convolution neural network for feature extraction and the rear end connects two full connected layers for classification and Regression in the grid regions. YOLO divides the input image scale into 7*7 grids, each of which will produce two bounding boxes. The bounding box will output a 4-dimnesional vector of coordinate information and the object confidence. Meanwhile, each grid also outputs 20 category probabilities, thus each grid produces a 30-dimentional vector including recognition information and location information. During the detection, YOLO filters the object proposals with low confidence by setting the threshold and wipes off the redundant object proposals to gain the detection results.

g. SSD

SSD is the single shot multi-box detector proposed by Liu Wei. The design of SSD has integrated YOLO's regression idea and Faster R-CNN's anchors mechanism. With the regression idea of YOLO, SSD simplifies the computation complexity of the neural network to guarantee the real-time. With the anchors mechanism, SSD can extract the features of different scales and aspect ratios to guarantee the detection accuracy. And the local feature extraction method of SSD is more reasonable and effective compared with the general feature extraction method of YOLO. What's more, because the feature representations in different scales are different, the method of multi-scale feature extraction has been applied in SSD, which contributes to promoting the detection robustness of different-scale objects.

C. Problem Definition

Many problems in computer vision were struggling with their accuracy before a decade. However, with the increase of deep learning techniques, the accuracy of those problems drastically improved. One of the major problems was that of image classification, which is defined as predicting the class of the image. one in all the key problems was that of image classification, which is defined as predicting the category of the image. a rather complicated problem is that of image localization, where the image contains one object, and also the system should predict the category of the placement of the thing within the image. The input to the system is going to be an image, and also the output is going to be a bounding box around all the objects within the image, together with the category of the object in each box.

II. LITERATURE SURVEY

We have done exhaustive literature survey over 7 papers. These papers focus on deep learning methods to perform object detection. Result of literature survey is tabulated as follow.

Table 1: Literature Survey

Literature Survey	
Sr. No.	Papers
1.	Girshick, J. Donahue, T. Darrell and et al have proposed RCNN method in "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". • Inference: 1. The principle of R-CNN is that it utilizes the region segmentation method of selective search to extract the region proposals in the image, which include the possible object candidates, and loads them into convolution neural network to extract the feature vectors. 2. Later, the classifier SVM will be used to classify the feature vectors to obtain the classification results in each region proposal. After merging by non-maximal suppression (NMS), the model outputs the precise object classifications and object bounding boxes to achieve object detection. • Advantages: 1. VOC2007 test dataset, the mAP of R-CNN object detection reaches 58.5% that is considerably lifted up compared with the former methods. • Disadvantages: 1. real-time detection is poor 2. accuracy is less than SPP-net
2.	K. He, X. Zhang, S. Ren and et al have proposed SPP method in "Spatial pyramid pooling in deep convolutional networks for visual recognition".

	• Inference: 1. The spatial pyramid pooling layer can get rid of the crop/warp operation on the input image in the former method. 2. It enables the input images of different sizes to connect with the full connection layer with the feature vector of the same dimension after passing the convolution layer. 3. The crop/warp operation reshapes the sizes of input convolution neural network to the fixed size, which will lead to the incompleteness of object image and object deformation • Advantages: 1. SPP-net solves the problems of object image incompleteness and object deformation. • Disadvantages: 1. It is still of colossal computation and poor real-time because its image processing is similar to that of R-CNN.
3.	R. Girshick proposed Fast RCNN method in “Fast r-cnn” paper in 2015. • Inference: 1. Fast R-CNN maps the region proposal extracted by a selective search algorithm in input image to the feature layer of convolution neural network and conducts the pooling on the mapped region proposal of feature layer by ROI pooling. 2. The role of ROI pooling is just like the spatial pyramid pooling of SPP-net. 3. The method of mapping region proposal of input image to the feature layer in Fast R-CNN shares the convolution computation, which substantially reduces the calculation. 4. Fast R-CNN adopts truncated SVD to enable that the single fully connected layer corresponding to weight matrix is replaced by two small fully connected layers, which further lessens the network calculation. • Advantages: 1. In the training stage, the speed of Fast R-CNN is 8.8 times that of R-CNN and 2.58 times that of SPP-net. 2. In the test stage, the speed of Fast RCNN is 146 times that of R-CNN and 7 times that of SPP-net 3. Speed is faster than SPP and RCNN. • Disadvantages: 1. Poor real-time performance.
4	S. Ren, K. He, R. Girshick and et al have proposed faster rcnn method in “Faster r-cnn: Towards real-time object detection with region proposal networks” in 2015. • Inference: 1. Selective Search method is replaced by region proposal network 2. Faster R-CNN employs the region proposal network (RPN) to solve the issues of huge computation and poor real-time caused by the selective search method in R-CNN and Fast R-CNN.

	3. Faster R-CNN is an end to end framework which can train the model easier. The function of RPN in Faster R-CNN is to replace the role of selective search in obtaining region proposals. 4. RPN could divide the feature layer into n×n regions and obtain the feature regions of various scales and aspect ratios that are centered on the region, and the method is called anchors mechanism. 5. The anchors in RPN are used to produce object proposals and then the proposals are sent to the rear classification and regression networks for the object recognition and location. • Advantages: 1. Faster RCNN adopts the RPN, the region proposals are reduced from 2000 (by selective search) to 300. 2. The mAPs of Faster R-CNN in VOC2007 and VOC2012 dataset tests have been raised by 2% to 3% to reach 69.9% and 67.0% respectively compared with those of Fast R-CNN. • Disadvantages: 1. Still, it cannot be used as real-time.
5.	Y. Li, K. He and J. Sun have proposed R-FCN method in “R-FCN: Object detection via region-based fully convolutional networks” paper in 2016. • Inference: 1. R-FCN could do the recognition and location simultaneously to achieve object detection 2. R-FCN can record the response of every object in different locations. 3. R-FCN defines the feature vector (C+1-dimension column vector) by voting according the RoIs and adopts softmax classification to classify the feature vectors in order to achieve the object recognition. • Advantages: 1. Test speed of R-FCN is 2.5 times that of the Faster R-CNN. • Disadvantages: 1. The accuracy of R-FCN is similar to that of Faster R-CNN.
6.	J. Redmon, S. Divvala, R. Girshick and et al have proposed YOLO method in “You only look once: Unified, real-time object detection” paper in 2016. • Inference: 1. The primary reason of the YOLO’s accuracy decline is the cancel of region proposal. • Advantages: 1. YOLO’s detection speed is 45 fps, 2. YOLO’s can reach 155 fps to make the real-time detection possible • Disadvantages: 1. Accuracy of YOLO is 66.4%, while the accuracy of Faster R-CNN is 73.2%.
7.	W. Liu, D. Anguelov, D. Erhan and et al have proposed SSD method in “SSD: Single shot multibox detector” paper in 2016.

	• Inference: 1. The design of SSD has integrated YOLO's regression idea and Faster R-CNN's anchors mechanism. 2. SSD simplifies the computation complexity of the neural network to guarantee the real-time performance with the anchor mechanism. 3. SSD can extract the features of different scales and aspect ratios to guarantee the detection accuracy. • Advantages: 1. The detection speed of SSD is 59 fps. its accuracy is 74.3%. Satisfies real-time requirements. • Disadvantages: 1. Weak detection capacity to the small objects.

III. IMPLEMENTATION

Here we have focused on YOLO method to detect images. We have experimented with several implementations of YOLO. User will give an input image to system. Image is passed to convolutional layer. Convolution layer predicts bounding boxes, class probabilities and extracts feature from image. The front end of YOLO connects a convolution neural network for feature extraction and the rear end connects two full connected layers for classification and regression in the grid regions.

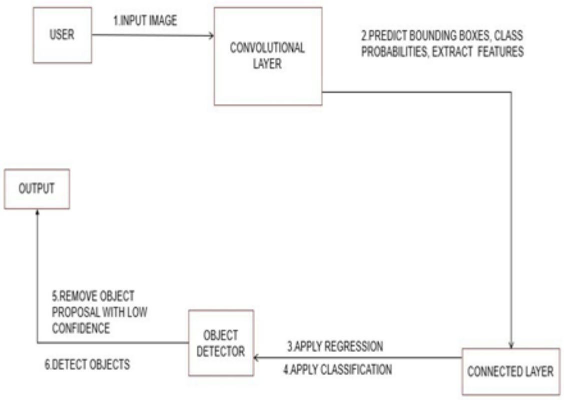

Fig -1: Block diagram of object detection

The front end of YOLO connects a convolution neural network for feature extraction and the rear end connects two full connected layers for classification and regression in the grid regions. The front end of YOLO connects a convolution neural network for feature extraction and the rear end connects two full connected layers for classification and regression in the grid regions. During the detection, YOLO filters the object proposals with low confidence by setting the threshold and wipes off the redundant object proposals to gain the detection results.

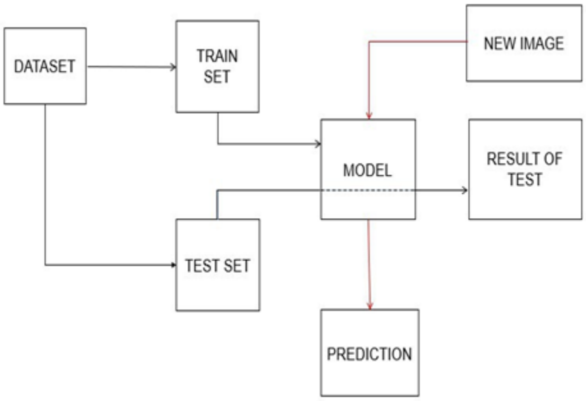

Fig -2: Workflow of object detection.

Above figure shows the workflow of any object detection project. We need a deep learning model to detect objects. We have chosen yolo model for the same. Yolo model is trained on coco dataset. We tested this model and presented output in result section.

IV. RESULT

We have used python programming language and OpenCV python library to implement YOLO model.

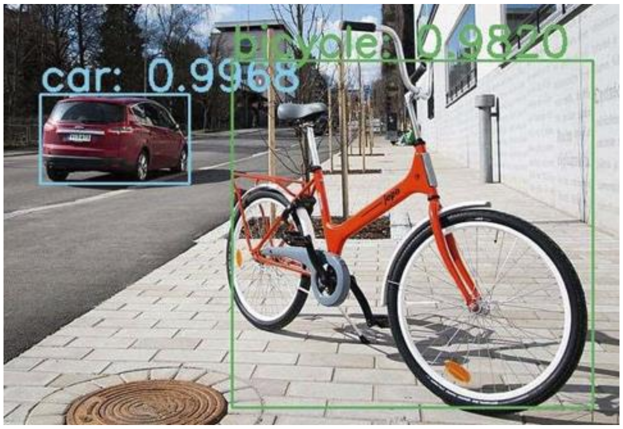

Fig -3: Detecting bicycle and car

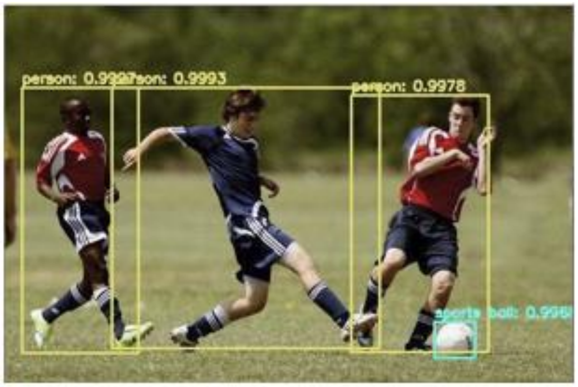

Fig -4: Detecting players while playing

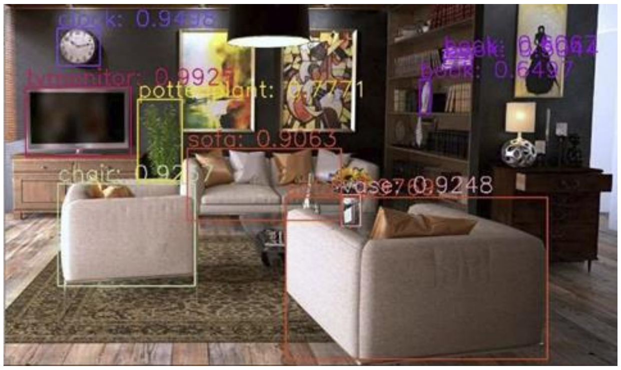

Fig -5: Detecting objects in living room

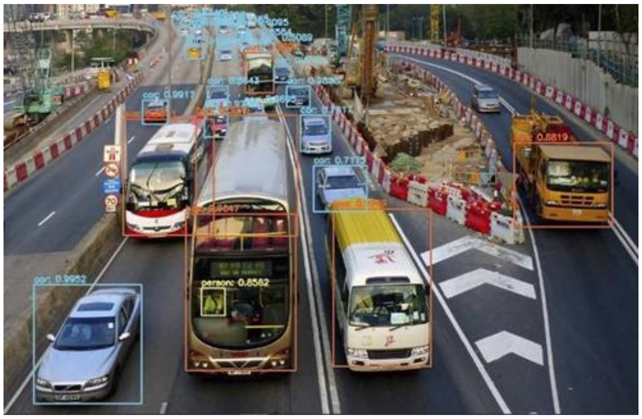

Fig -6: Detecting vehicles in traffic

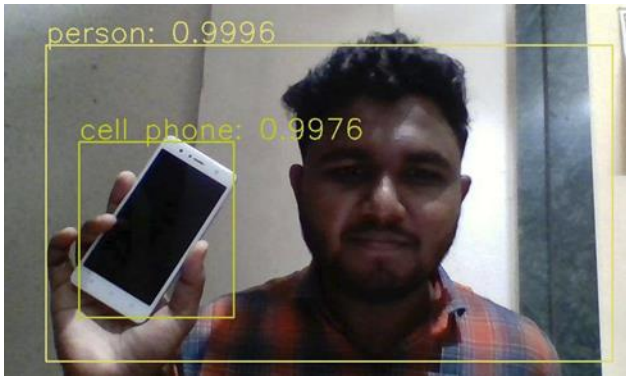

Fig -7: Testing on webcam

Table 2: Time analysis on intel i3 processor

Fig. no.	Time taken to detect objects (seconds)
3	1.843729
4	1.921862
5	1.843751
6	1.859379
7	1.953133

CONCLUSION

YOLO is faster than other previous approaches. Regression technique used in YOLO makes it fast. Although performance of YOLO is restricted to device specification. If a client server application is built where server side device does all the detection operation and has a high specification, then clients may get satisfy with the speed. Here performance of detection will only depend on server side and not on the client devices. But Yolo lags to detect small objects in the image. It can be used in real-time detection.

REFERENCES

[1] Girshick, J. Donahue, T. Darrell, et al, "Rich feature hierarchies for accurate object detection and semantic segmentation," IEEE Conference on Computer Vision and Pattern Recognition. 2014, pp.580-587

[2] K. He, X. Zhang, S. Ren, et al, "Spatial pyramid pooling in deep convolutional networks for visual recognition," European Conference on Computer Vision, 2014, pp.346-361

[3] R. Girshick. "Fast r-cnn," 2015 IEEE International Conference on Computer Vision, 2015, pp. 1440-1448

[4] S. Ren, K. He, R. Girshick, et al, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, 2015, pp.91-99

[5] Y. Li, K. He, J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," Advances in Neural Information Processing Systems, 2016, pp.387-397

[6] J. Redmon, S. Divvala, R. Girshick, et al, "You only look once: Unified, real-time object detection," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp.779-784

[7] W. Liu, D. Anguelov, D. Erhan, et al, "SSD: Single shot multibox detector," European Conference on Computer Vision, 2016, pp.21-37