Object Detection Techniques: A Comparison

Priyanka Malhotra
Chitkara University Institute of Engineering and Technology
Chitkara University, 140401, Punjab, India
priyanka.malhotra@chitkara.edu.in

Ekansh Garg
Chitkara University Institute of Engineering and Technology
Chitkara University, 140401, Punjab, India
ekansh24j@gmail.com

Abstract— **Computer vision is one of the technologies that aim at digitally perceiving the real world at a higher level through digital images and videos. Object detection, a subset to computer vision is one of the prominent techniques in this area of research..Object detection is basically an algorithm based on either machine learning or deep learning approaches employed for classification of elements in diverse classes and localization in the image. This paper provides a comparison among the three prominent approaches to achieve object detection. R-CNN, Fast R-CNN, YOLO are the techniques in the trend which facilitates the developer in accomplishing the task of detecting an object in the image. These techniques train and compute the parameters of the model in reduced hence increase performance as compared to the traditional object detection techniques.**

Keywords—Deep learning, object detection

I. INTRODUCTION

Object detection is a computer vision problem that deals with identification and localization of objects of certain classes in the image. Object Classification can be realized since every object has special features thus can be categorized according to its shape (circular, square, cylinder), size and color etc. A rectangular box referred to as the bounding box [1] is determined around the object by upper left and lower right x and y coordinates of the rectangle. Bounding boxes and marking pixels within the image are some of the common approaches to achieve localization.

Object detection algorithms are based on either machine learning or the deep learning approaches. In machine learning, firstly features in the image are defined using different methods like Harr features, scale invariant features transform, histogram of oriented gradients etc. After defining the features, generally support vector machine is employed for classification. In deep learning approaches, object detection is performed without defining the features and employ convolutional neural network (CNN). After machine learning, deep learning algorithms took the center stage for complex problems. With the introduction of deep learning, algorithms for computer vision started developing in 1980 with “Neocognitron” by Kunihiko Fukushima[2]. Unlike machine learning, deep learning proves to be a more compact process in which the classification and localization occurs in a single trail after the image is captured. The techniques like R-CNN, Fast R-CNN, YOLO have emerged eventually making the process faster and compact. R-CNN performs classification within regions of the image. Fast R-CNN technique generates a feature map of the image in one trail. YOLO is an algorithm which operates on the basis of probability of objects in bounding box and grids. The field of artificial intelligence based object detection suffers with complexity due to the need of high accuracy and precision, especially in modern applications like autonomous cars, face detection, traffic sensing etc.

The paper discusses the techniques employed for object detection. The section II of paper explains the three prominent techniques for object detection. Section III gives the different metrics employed for evaluating the performance of object detection algorithm. This section also explains the various datasets employed for training and testing of object detection algorithms. Section IV gives comparison between the different techniques of object detection. Section 5 concludes the paper.

II. OBJECT DETECTION TECHNIQUES

Deep neural network- Inspired by the human brain, deep learning has been contributing abundantly in computer programming techniques, to enable the machine to perform tasks nearly imitating human intelligence. Being a subset of machine learning in artificial intelligence, it is capable of both supervised and unsupervised learning from data available in bulk. Supervised learning uses labeled data for training the model, whereas unsupervised training uses data which is both unstructured and unlabeled. Deep learning algorithm employs convolutional neural networks (CNN) with deep layers to enable learning. The different layers of the convolutional neural network [3] perform different functions. The convolution layer together with pooling layer produces feature maps. These features further are fed to fully connected layer; it performs high-level reasoning in NN for linear classification. The soft max layer assigns the different objects on to different classes. The deep learning algorithm improves itself by performing the task repetitively, each time tweaking it a little to enhance the outcome. Fig.1 shows deep neural network architecture. The different object detection models based on deep learning has been proposed by researchers for detection of objects in various datasets. Several deep learning algorithms have been proposed by researchers in the literature. Each algorithm has its own number of layers and computed parameters. AlexNet is one of the basic deep learning algorithm proposed by Krizhevsky et al. [4] consists of 8 layers- five convolutional layers with max pooling layers and ReLu activation function, three fully connected layers and dropouts.

978-1-7281-7223-1/20/$31.00 ©2020 IEEE

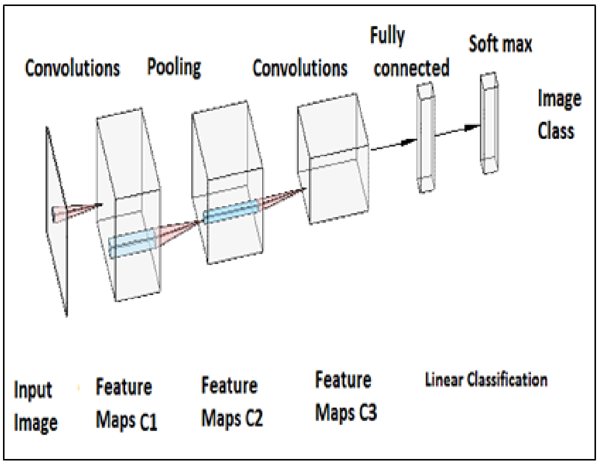

Fig.1 Convolutional Neural Network

Region with CNN (RCNN)- RCNN generates features in a region using CNN. The algorithm proposed by Girshick et al. [5] employed selective search to extract just 2000 regions from the entire input image. These regions are referred to as region proposals. The subsets of these regions are identified in the image and are employed for classification of the objects in the regions. Therefore, instead of classifying big number of regions, just 2000 regions (as shown in Fig. 2) can be worked with. All these regions are generated by employing the selective search algorithm. The last layer of the CNN consists of the features which are extracted from the entire image and these features are further fed to a classification algorithm like support vector machine. This algorithm classifies the objects lying within the region proposal network.

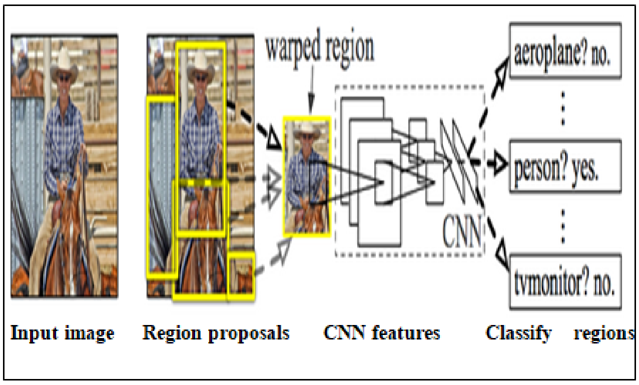

Fig.2 RCNN model [5]

Fast R-CNN R-CNN model took a huge amount of time to train the network. Girshick et al. [6] built another faster object detection algorithm known as Fast R-CNN to circumvent this problem. The input image is fed to the CNN algorithm which further generates convolutional feature map. The fig.3 represents Fast R-CNN architecture. The model takes the entire image as an input with a set of different object proposals. The network firstly processes the entire input image using more than one convolutional and max pooling layers so as to produce the convolution feature map for the entire image. The region proposals are generated using an algorithm such as Edge Boxes. Further, for each of the object proposals, a region of interest (RoI) pooling layer converts the object proposals into a feature vector of a particular size. These feature vectors are fed to a set of fully connected layers attached to the two output layers, one layer produces softmax probability estimate and the other output layer gives four real value numbers for the K object classes. The bounding box coordinates are refined for one of the K classes using SVM trained using CNN features.

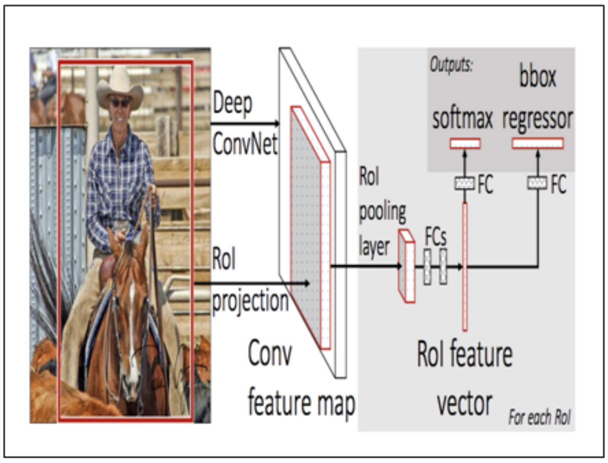

Fig 3 Fast RCNN architecture [6]

YOLO: In contrast to the previous object detection algorithms, instead of complete image, the network looks at the portions of the image possessing immense probabilities that the object is present. You only look once or YOLO is an algorithm for object detection, which has features much different from the previous algorithms. This algorithm deploys bounding boxes and classifies the image hence predicting possibilities for these boxes using a unit convolutional network.

YOLO [7] works on input image by splitting it into an SxS grid, in each of which "m" bounding boxes are marked. In these bounding boxes, the network predicts class probability and offset values. The bounding boxes exceeding the threshold value for class probability are selected and are instrumental in locating the object inside the image. YOLO processes the image as fast as 45 FPS than the rest algorithms for object detection.

YOLO algorithm has some constraints as it hustles when detecting small objects in the image, for example it may struggle with detection of flock of birds because of its spatial constraints. YOLO is acutely quick and precise in mAP. Speed and accuracy can be compensated for each other just by transforming the size of the model, requiring no retraining. Before detection, the systems repurpose the classifiers or localizers in order perform detection. The model is used at multiple locations in an image. Regions with high scores are considered to be detected regions in the image.

At the time of testing it glances at the entire image, therefore its forecasts are influenced by overall context of the image. Unlike R-CNN, it is able to predict with a unit network evaluation instead of thousands for one image. This results in the algorithm being 1000x and 100x faster than R-CNN and Fast R-CNN respectively. Section IV gives a detailed comparison of the three techniques.

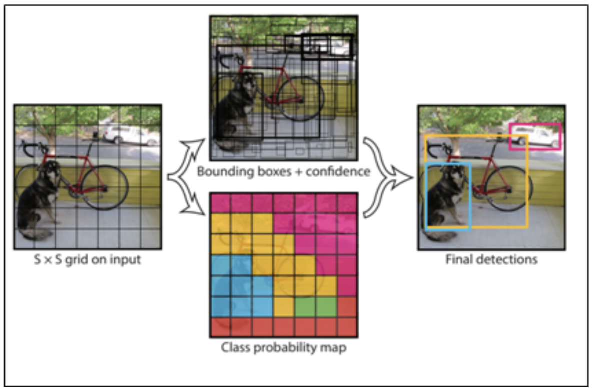

Fig.4 YOLO model [7]

III METRICS and DATASETS

Metrics: The performance of different object detection algorithm can be evaluated using different performance metrics:

MAP (mean Average Precision): It is a metric used for measuring the degree of precision and accuracy for object detection techniques. The metric gives the mean for average precision of object detection.

TABLE 1 APPLICATION OF RCNN, FAST-RCNN AND YOLO ON DIFFERENT DATASETS

STUDY	**OBJECT DETECTED**	**MODEL UTILIZED**	**DATASET USED**	**TYPE OF IMAGE**
Ross Girshick et al.	Different objects	RCNN	ILSVRC2012 ILSVRC2013 PASCAL VOC	Png/jpeg
Braun et al.	Car and pedestrian	RCNN	KITTI	jpeg
Xia et al.	Blood cell	Fast-RCNN	Microscopic images	png/jpeg
Ullah et al.	Pedestrian	Fast RCNN	Infrared Image dataset	jpeg
Ren et al.	20 objects	Fast RCNN	PASCAL VOC2007	png/jpeg
Shinde et al.	Human action	YOLO	Liris Human Activities dataset	jpeg
Redmon	Natural images	YOLO	ImageNet2012	jpeg

$$\textbf{MAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \qquad (1)$$

Where, AveP(q) represents the average precision for a given instance. Q represents the total number of instances.

Intersection over Union (IoU): It is a metric employed to measures the amount by which the predicted object boundary overlaps the given ground truth.

$$\textbf{IoU} = \frac{\textbf{area of overlap}}{\textbf{area of union}} \qquad (2)$$

Where, Area of overlap represents the pixels found in both the predicted image and the ground truth image. Area of union represents the pixels found in either predicted or ground truth image.

Datasets: There are different data sets employed for object detection task. The few of datasets are explained below:

ILSVRC2012: It is a subdivision of the vast Image Net dataset which is hand-labeled and has 1.2 million images with 1000 different object categories [5].
PASCAL VOC: Pascal VOC challenge is one of the famous datasets for algorithm building and evaluation for image classification, detecting objects, segmentation. It contains number of ".jpg" images as data [5].
ImageNet: It is a database of images standardized according to the WordNet ranking. Each node is depicted by images in numbers up till hundreds and thousands [13].
COCO: It is a data set for object detection, segmentation, and captioning. It has 1.5 million object instances and 80 different categories of objects.
Google's OpenImageV4: It is a dataset having diverse images and consists of complex scenes with various objects (on an average of 8.4 per image). It has labeled annotations on image level, bounding boxes around the objects, localized narratives, visual relationships, object segmentations.
Blood Cell Count Detection: This dataset consists of 12,500 augmented blood cell images (JPEG) which have labeled cell types (CSV). It has 4 different cell variations clubbed into 4 contrasting folders (according to cell type) for which it has 3,000 images for each.

IV COMPARISON

The three different object detection techniques discussed in this paper are compared in this section. Table 2 gives comparison between the three object detection techniques discussed in this paper. R-CNN, Fast RCNN and YOLO use different classification methods. YOLO can run in real time and is fast but the technique struggles to find small objects in a group like detecting a bird in the flock of birds.

Table 2 COMPARISONS BETWEEN DIFFERENT OBJECT DETECTION TECHNIQUES

R-CNN	**Fast RCNN**	**YOLO**
Based on classification	Based on classification	Based on regression
Uses 2000 conv Nets for each region	Uses single deep convent	Uses single convolution network for entire image
Uses selective search algorithm	Uses selective search algorithm	---
Needs 49 seconds to test one image	Needs 2.3 seconds to test one image	Needs less than 2 second to test on one image
Slow, cannot be implemented real time	Faster than RCNN	Can run real time
Uses SVM for classification	Uses softmax for classification	Uses regression for classification
Produces bounding boxes	Produces bounding Box regression head and classification head	Produces bounding box, prediction contextual concurrently
Can find small objects	Can find small objects	Struggles to find small objects that appear in groups

V CONCLUSION

RCNN, Fast RCNN and YOLO are the common techniques employed for object detection. RCNN and Fast RCNN are slower than YOLO but can detect small objects. YOLO is good at regression than classification. YOLO has difficulty in classifying small objects. Both RCNN and Fast RCNN fails to perform real time detection but YOLO can perform real time classification with good speed. The choice for the type of object classification algorithm employed depends on type of data set, type of images, amount of training/testing time and the application requiring detection of object and the type of object.

REFERENCES

1. He, Y., Zhu, C., Wang, J., Savvides, M. and Zhang, X., "Bounding box regression with uncertainty for accurate object detection" In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2888-2897.
2. Fukushima, K. and Miyake, S., 1982. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In Competition and cooperation in neural nets (pp. 267-285). Springer, Berlin, Heidelberg.
3. LeCun, Y., Bengio, Y. and Hinton, G., "Deep learning" 2015, Nature, V521(7553), pp.436-444.
4. Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pp. 1097-1105.
5. [5] Girshick, R., Donahue, J., Darrell, T. and Malik, J., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 580-587.
6. [6] Girshick, R., 2015. Fast r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 1440-1448.
7. [7] Redmon, J., Divvala, S., Girshick, R. and Farhadi, A., "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.

978-1-7281-7223-1/20/$31.00 ©2020 IEEE

8. [8] Braun, M., Rao, Q., Wang, Y. and Flohr, F., 2016, November. Pose-rcnn: Joint object detection and pose estimation using 3d object proposals. In 2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC) (pp. 1546-1551). IEEE.

9. [9] Xia, T., Jiang, R., Fu, Y.Q. and Jin, N., 2019, October. Automated Blood Cell Detection and Counting via Deep Learning for Microfluidic Point-of-Care Medical Devices. In IOP Conference Series: Materials Science and Engineering (Vol. 646, No. 1, p. 012048). IOP Publishing.

10. [10] Ullah, A., Xie, H., Farooq, M.O. and Sun, Z., 2018, November. "Pedestrian detection in infrared images using fast RCNN," in 2018 Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA), IEEE, 2018, pp. 1-6.

11. Ren, Y., Zhu, C. and Xiao, S., 2018. Object detection based on fast/faster. RCNN employing fully convolutional architectures. Mathematical Problems in Engineering, 2018.

12. Shinde, S., Kothari, A. and Gupta, V., 2018. YOLO based human action recognition and localization. Procedia computer science, 133, pp.831-838.

13. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. and Berg, A.C., 2015. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), pp.211-252

14. Vicente, S., Carreira, J., Agapito, L. and Batista, J., 2014. Reconstructing pascal voc. in Proceedings of the IEEE conference on computer vision and pattern recognition , pp. 41-48.

15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Duerig, T. and Ferrari, V. "The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale", 2018, arXiv preprint arXiv:1811.00982.

16. Wong, A., Famuori, M., Shafiee, M.J., Li, F., Chwyl, B. and Chung, J "YOLO nano: A highly compact you only look once convolutional neural network for object detection", 2019, arXiv preprint arXiv:1910.01271.