Multiple Object Detection in Images and Video using Deep Learning Models

Madhusudhan Reddy.K[1], S.Muni Kumar[2].

[1]*Student, Department of Computer Applications, KMM Institute of Postgraduate Studies, Tirupati*

[2]*Assistant Professor, Department of Computer Applications, KMM Institute of Postgraduate Studies, Tirupati*

Abstract- **Real-time object detection and recognition is a vast, vibrant yet inconclusive and complex area of computer vision and is also an important task in image processing and computer vision. It is concerned with determining the identity of an object in an image or a Real-time video surveillance from a set of known labels. Humans can recognize any object in the real world easily without any effort or difficulty. But computerized recognition of an object in an image is not an easy task as it involves processing of data (images and videos).In this paper, you will come across the most used deep learning models for object detection.**

Index Terms- **Object Detection, Multiple Object Detection, Deep Learning, Object detection in image, Computer vision.**

I. INTRODUCTION

Humans glance at an image and instantly know what objects are in the image, where they are, and how they interact. The human visual system is fast and accurate, allowing us to perform complex tasks like driving with little conscious thought. Fast, accurate algorithms for object detection would allow computers to drive cars without specialized sensors, enable assistive devices to convey real-time scene information to human users, and unlock the potential for general purpose, responsive robotic systems.

Object detection and recognition in digital images and video have become one of the most important applications for industries to ease user, save time and to achieve parallelism. This is not a new technique but improvement in object detection is still required in order to achieve the targeted objective more efficiently and accurately. By using a computer and later on develop a system that reduces human efforts. The process of object detection analysis is to determine the number, location, size, position of the objects in the input image. Object detection is the basic concept for tracking and recognition of objects, which affects the efficiency and accuracy of object recognition. The common object detection method was the color-based approach, detecting objects based on their color values. Template matching and shape-based approaches which detect objects based on templates and shapes. Since the researchers have found many other techniques as the technology has evolved to detect and recognize the objects in images and in the live video feed, the most popular technology that is used today is the deep learning or the neural network for object detection.

When it comes to deep learning-based object detection R-CNN's are likely the most heard and used methods for object detection and recognition using deep learning. By the end of this paper, you will understand the applications of object detection and how deep learning is used in object detection and recognition, and how each R-CNN methods are different from one another.

II. R-CNN

R-CNN or Region-based Convolutional neural network is one of the first and oldest models in convolutional neural networks. The goal of this method is to identify the objects in an image by bounding a box around the object in an image. In brief, it first takes an image as an input and scans the image by using a selective search algorithm to generate the region proposals and runs a convolutional neural network on each of the regions generated. Finally the output of the CNN is provided as input to the SVM (Support Vector Machine) for classification of regions.

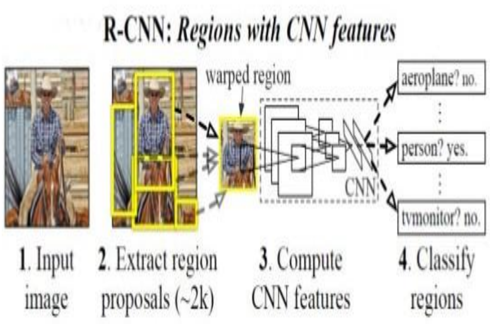

Figure 1. Architecture of R-CNN

III. FAST R-CNN

Fast R-CNN is the Advanced Version of R-CNN. Fast R-CNN is similar to R-CNN, but when compared to R-CNN Fast R-CNN was quick in detection of objects.it was made possible to detect objects quicker by making the following enhancements to R-CNN: Feature extraction of image was done before proposing the regions so that we can run CNN over the image only once instead of 2000 times over 2000 overlapping regions. Softmax layer was introduced instead of SVM (Support Vector Machine).

When compared to R-CNN, Fast R-CNN was 25x faster.

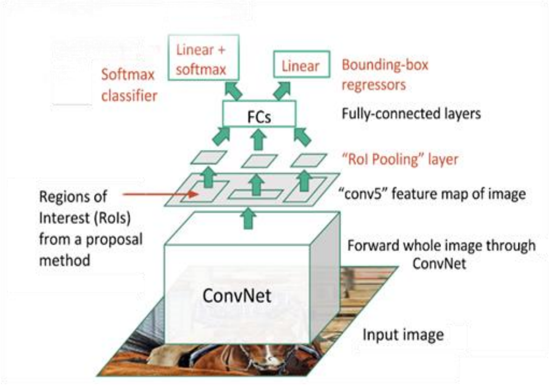

Figure 2.Architecture of Fast R-CNN

IV. FASTER R-CNN

The objective of Faster R-CNN was to replace the selective search algorithm which was very slow with the fast neural network by bringing in the RPN (Region Proposal Network).

3.1 Working of RPN

In the last layer of first CNN, a 3x3 sliding window is moved across the feature map and maps it into lower (256-d) dimension.

For every sliding window location multiple possible regions are generated based on fixed -ratio *k* anchor boxes.

Every region proposal consists of the score "objectness" for the region and the bounding box of the region represented by 4 coordinates.

In simple words, we look at every location in our feature map and consider k different boxes centered around it: a wide box, a large box, etc. for each of these boxes, we output whether or not we think it contains an object, and what the coordinates for that box are.

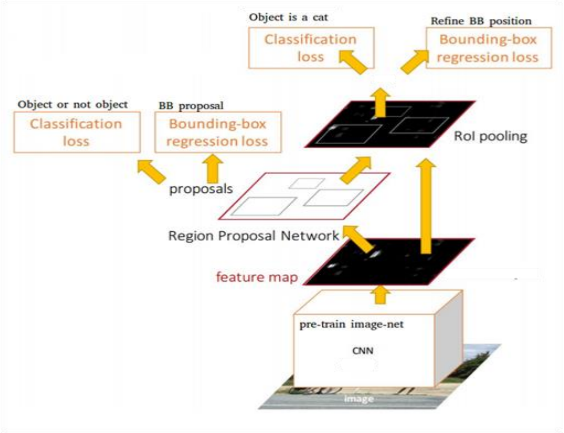

Figure 3.Architecture of Faster R-CNN

Faster R-CNN takes 0.2 seconds of test time per image which is 250x faster than previous models.

V. R-FCN

R-FCN stands for Region-based Fully Convolutional Network. The previous models such as Fast R-CNN and Faster R-CNN shares the computation of repeated convolutional features for object classification and region proposal to save time. But, many unshared fully connected layers in Faster R-CNN were present that must be computed. The problem of opposing needs for classification and localization leads to unnatural insertion of region proposals between convolutional layers.

R-FCN works as follows:

1. First, we run a convolutional neural network on the input image (ResNet in this case).
2. We will add a fully convoluted layer for generating the score bank. There should be k²(C+1) score maps, with k² representing the number of relative positions to divide an object (e.g. 3² for a 3 by 3 grid) and C+1 representing the number of classes plus the background.
3. To generate the region of interest (ROI's), run a fully convolutional region proposal network (RPN).
4. Every ROI is divided into k² "bins" or subregions as the score maps.
5. Check the score bank to find out if that bin matches the corresponding position of the same object for each bin.
6. Once each of the k² bins has an "object match" value for each class, average the bins to get a single score per class.
7. Finally, classify the ROI with a Softmax over the remaining C+`+1 dimensional vector.

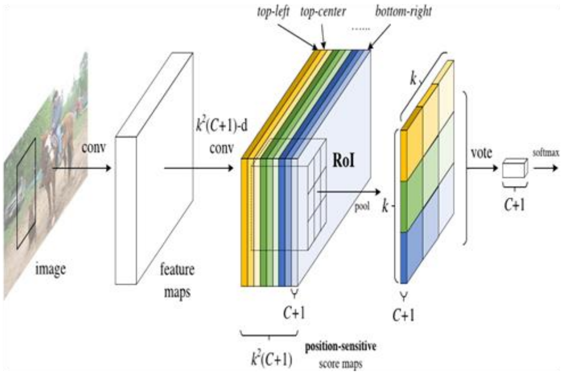

Figure 4.Architecture of R-FCN

When compared to Faster R-CNN, R-FCN is 2.5-20x faster with comparable accuracy.

VI. SSD

Our final model is SSD.SSD stands for Single Shot Detector. Similar to R-FCN it provides great speed over Faster R-CNN, but in a different manner.

The First two models performed Region proposals and classifications in two separate steps. Initially, they used RPN (Region proposal network) for generating ROI (Region of Interest). Then, they used fully-connected layers or position-sensitive convolutional layers to classify those regions. The single shot detector (SSD) does these two steps in a single shot and can predict the bounding boxes and the class as it processes the image simultaneously.

SSD does the following:

1. SSD passes the image through the series of convolutional layers which yields several sets of feature maps at different sizes.
2. For every location in each of these feature maps, uses a 3x3 convolutional filter to evaluate a small set of default bounding boxes.
3. Simultaneously predict the bounding box offset and the class probabilities.
4. Match the ground truth box with these predicted boxes based on IoU during training.

SSD sounds straightforward, but training it has a unique challenge. With the previous two models, the region proposal network ensured that everything we tried to classify had some minimum probability of being an "object." With SSD, however, we skip that filtering step. We classify and draw bounding boxes from every single position in the image, using multiple different shapes, at several different scales. As a result, we generate a much greater number of bounding boxes than the other models, and nearly all of them are negative examples.

To fix this imbalance, SSD does two things. Firstly, it uses non-maximum suppression to group together highly-overlapping boxes into a single box. In other words, if four boxes of similar shapes, sizes, etc. contain the same dog, NMS would keep the one with the highest confidence and discard the rest. Secondly, the model uses a technique called hard negative mining to balance classes during training. In hard negative mining, only a subset of the negative examples with the highest training loss (i.e. false positives) are used at each iteration of training. SSD keeps a 3:1 ratio of negatives to positives.

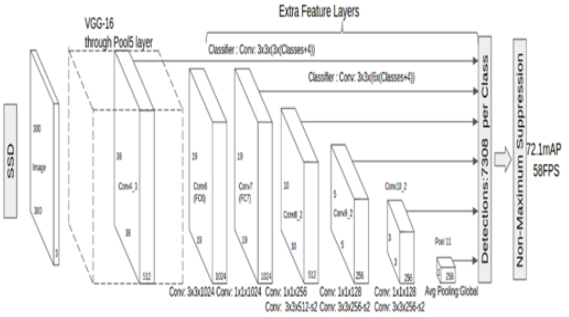

Figure 5.Architecture of SSD

VII. APPLICATIONS OF OBJECT DETECTION

A. Face detection

The popular application of object detection is the face detection. We have noticed this feature in the social network website "Facebook" where when we upload the photo the face will be detected to tag the person in the particular photo. This is the simple application of object detection in the day to day life.

B. Vehicle detection

Vehicle detection is another popular application of the object detection. Object detection with tracking can prove effective in estimating the speed of the object. And also the type of the vehicle entering into the premises can be recognized by object detection.

C. Manufacturing industry

Object detection is also used in the manufacturing industries for identification of products. If u want your machine to only detect circular or rectangular objects the object detection system is used to identify such shaped objects.

D. Online Images

Object detection can be used to classify the images found in the internet. Offensive images are usually filtered out using object detection.

E. Video Surveillance

Video surveillance is one of the popular application of object detection.by using object detection in video surveillance we can track the unauthorized people or vehicles entering the premises.

F. Self-Driving Cars

Self-driving cars use the object detection technology to detection the oncoming traffic, objects and symbols present on the roads. Using this object detection system the self-driving cars navigate from one place to another place.

VIII. CONCLUSION

In recent years, the technologies of deep learning in object detection have achieved great success. Methods such as R-CNN, Fast R-CNN, Faster R-CNN, R-FCN and SSD are most widely used object detection models. By using these models we can achieve greater heights in fields like automobiles industries and manufacturing industries and in security surveillance.

REFERENCES

[1] Ross Girshick "Fast R-CNN", (arvix), 2015.

[2] Shaoqing Ren, kaiming He, Ross Girshick, Jian Sun "Faster R-CNN: Towords Real-Time Object Detection with Region Proposal Networks" (NIPS), 2015.

[3] Jifeng Dai, Yi Li, Kaiming He, Jian Sun "R-FCN: Object detection Via Region-based Fully Convolutional Networks", (NIPS), 2016.

[4] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Redd, Cheng-Yang Fu, Alexander C.Berg "SSD: Single shot MultiBox Detector",(arvix),2016.

[5] Khushboo Khurana, Reetu Aswathi "Techniques for Object Recognition in Images and Multi-Object Detection", (IJARCET), 2013.

[6] www.towardsdatascience.com, Joyce Xu "Deep learning for object detection: A Comprehensive Review" 2017.

[7] Bhumika Gupta, Ashish Chaube, ashish Negi, Umang Goel " Study on Object Detection using Open CV-Python", (IJCA) volume 162-No 8,2017.

[8] joseph Redmon, Santosh Divvala, Ross Girshick, ali farhadi "You only Look Once: Unified Real-Time Object Detection"(Arvix), 2016.