Object Detection

Srishti Sharma[1] and Bhumiti Gohel[2]

Abstract— **Object detection is the task of detecting and identifying semantic objects from a set of classes as captured from an image or a video stream. Training newer models for solving the object detection and classification problem of a particular domain is difficult due to the computational complexity and the time it takes in training a new model. Transfer learning is the process of making use of a pre trained model for training the dataset used in the current domain. In this assignment, we have made use of Faster R-CNN and Mask R-CNN region proposal based pre trained deep neural networks to improve the object detection over customised datasets. The Mask R-CNN used by us detects images with a minimum 98 percent confidence.**

Keywords– Object Detection, Transfer Learning, Faster R-CNN, Mask R-CNN

I. Introduction

The technique of identifying and locating objects within an image or video is known as Object Detection in Computer vision [1]. In general, object detection draws a square box around the object that is detected along with it's label near it. This helps in recognizing the objects easily. Object detection is generally confused with Image recognition. To clear out more on this, Image recognition is the task of assigning a label to image as whole. If an image contains many cats, the algorithm will label the whole image as a cat. Whereas, in the case of Object Detection, it draws a box around each cat present in the image and labels each one of them as a cat individually [1]. Hence, the model predicts its location and what the object is. Hence, we can say that Object Detection works more precisely. Object Detection is inextricably linked to other similar techniques such as Image Segmentation and Image Recognition. What separates Object Detection from all these techniques is the fact that it can detect the exact location of an object in the image or video. This object detection technique can be used in various other fields such as: Video Surveillance Face Detection Crowd Counting Anomaly Detection Self-Driving Cars The state-of-the-art methods of object detection can be categorized into two main types: one-stage methods and two stage-methods. One-stage methods prioritize inference speed, and example models include YOLO, SSD and RetinaNet [1]. Two-stage methods prioritize detection accuracy, and example models include Faster R-CNN, Mask R-CNN and Cascade R-CNN. The paper is divided into four section here on wards. The next section gives a brief literature review of the existing work done in the domain. Section 3 details Transfer Learning explaining the approach and the architecture of models used for transfer learning. Section 4 presents the results of the experimentation done and we conclude with Section 5.

II. Literature Review

Use of transfer learning for improving the performance of object detection through synthetic images and pre trained convolutional neural networks improves the performance of the deep neural networks [1]. Networks such as Faster RCNN, R-FCN, SSD are made used as the pre trained convolutional networks for the training purpose. In [1], the experiment conducted showed a mAP value 70.67 depicting an increase in the performance of object detection using Faster RCNN. In [2], there is an improvement achieved over the accuracy of object detection performed over CUB 200-2011 dataset making use of convolutional neural networks for transfer learning. The study presented in [3] suggests addressing the underwater object detection challenge. For the optical underwater images, the authors have worked over Yolov3-tiny network as the backbone and pre

[1]Srishti Sharma, AU2049002, PhD in Engineering, School of Engineering and Applied Sciences

[2]Bhumiti Gohel,AU1841051, Department of Information and Communication Technology

trained it on Pascal VOC datasets. The results of experimentation prove that making use of transfer learning improves the performance of the region based object detectors such as Mask RCNN and Faster RCNN as compared to SD algorithm and Yolov3. The experimentation here was conducted on CHINAMM 2019 dataset. In [4], the authors have demonstrated how region based object detection models such as Mask RCNN and Faster RCNN give improved object detection on live video streams and images. In [5], the authors have shown an interesting experiment where they borrow and train examples from other classes so as to train their classes to make the object detector more efficient. They created an object detector based on this concept and showed results on the SUN09 object detection dataset. In [6], the authors have made use of the capabilities of the convolutional neural networks such as data driven approach, high representation, extracting hierarchical image features from a good amount of training data. These experimentation was conducted in the domain of medical imaging and they proved how transfer learning using region based deep neural networks working on the region proposal approach improved the mAP value achieved for object detection.

III. Transfer Learning

A. Model Architecture

The models used in our experiment are:

- Faster R-CNN
- Mask RCNN

Faster R-CNN:

The algorithm works in 3 phases namely

- Convolution Layers
- Region Proposal Network
- Classes and Bounding Boxes Prediction

Convolutional Layers:

In this layers First we train filters to extract the appropriate features the image. We know that in CNN,Convolution networks are generally consist of Convolution layers, pooling layers and a fully connected or another extended thing that will be used for an appropriate task like recognition,object detection or classification [7]. after that we are performing convolution using sliding filter along with our image and as a result we will get two dimension matrix. This is known as feature map.

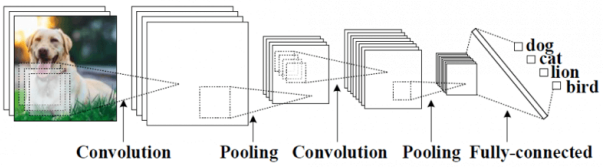

Fig. 1. Convolution Layer

In Pooling, we decrease quantity of features in the features map by removing pixels of low values.And the last thing is using the fully connected layer to classify those features which not our case in the Faster RCNN [7].

Region Proposal Network (RPN):

Region Proposal Network (RPN) is small neural network sliding on the last feature map of the convolution layers and predict whether there is an object or not. It also predicts the bounding box of those objects [7].

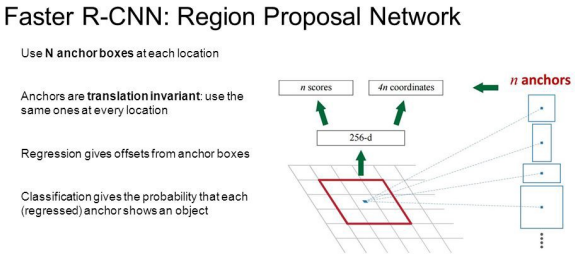

Fig. 2. Region Proposal Network

Classes and Bounding Boxes prediction:

In this layer we use another fully connected neural networks which takes the regions that are proposed by the RPN as an input and predicts class of the object i.e. Classification and the Bounding boxes i.e. Regression. After this, we train this architecture by using SGD to optimize filters in convolution layers, RPN weights and last fully connected layer weights [7].

Mask R-CNN:

Mask R-CNN is an extension of the Faster R-CNN approach for pixel level segmentation. This is done by adding a branch to Faster R-CNN that presents as an output a binary mask that conveys whether or not the pixel given is a part of an object. This branch added to the Faster R-CNN is in real terms a fully convolutional network. This convolutional

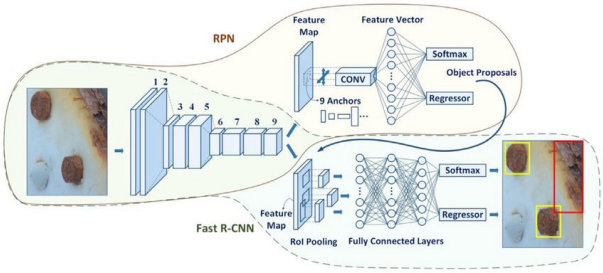

Fig. 3. Faster RCNN Architecture

network is built over the top of the CNN based feature map generated by Faster R-CNN [8].

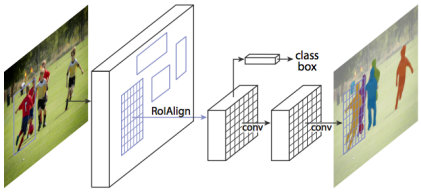

Fig. 4. Mask R-CNN Architecture

Inputs to Mask R-CNN was the feature Map generated by CNN and Outputs were matrices that contain 1s in the location of pixels that belong to the object and 0s in position of the pixel that doesn't belong to the object. While using the same faster R-CNN architecture, the region proposals of feature map selected by the ROIPool obtained by Mask R-CNN were a little misaligned.As the image segmentation task required a higher quality of pixel unlike that obtained with bounding boxes, using original faster R-CNN architecture led to inaccuracies [8].

To overcome this problem, RoIAlign was used instead of RoIPool that gave a precise alignment. With RoIAlign we made use of bilinear interpolation technique to have a clear and accurate idea about a pixel. Hence the issue of misalignment was overcome. Once the masks were generated this way, Mask R-CNN combined the masks with the classification and the bounding boxes that were generated by the Faster R-CNN and thereby generate precise segmentation [8].

B. Approach

Faster R-CNN:

Dataset:

The dataset collected from Google's OpenImages dataset [9]. It conatins which millions of images grouped into thousands of labels with bounding boxes.we downloaded only 5 objects such as person,scissors,mobile phone,calculators,pen etc and each object has 200 images.this is how we prepared dataset.These images were 1024 x 704 in dimension.

Approach:

After collecting the dataset we download the dataset along with it's annotations in xml format.after that we are spliting dataset into test images and train images.Here the 10 percent of the images are used for the testing other for training.after that we are creating a label for the dataset.after that we are converting files into TFRecords.TFRecord file stores your data as a sequence of binary strings. This can improve the performance of your process drastically. So here we are going to convert the data in our XML files to tfrecord format for faster execution.after that download the pretrained model(Faster R-CNN ResNet50 V1 640x640) from the Tensorflow Model Zoo.and trained the model in our dataset which we preprocessed.after doing sucessful training we exported the model.and doing prediction

Mask R-CNN:

Dataset:

The dataset was prepared by us by clicking pictures using a mobile phone. There were 50 pictures clicked containing daily utility objects such as Plate, Facewash, Snack, Keys, Book, Box, Case, Mask, Bottle, Glue, Pen, Remote, Mobile Phone etc. These images were 1280*960 in dimension and the size of each image was around 360 KBs.38 images were used for the purpose of training while 14 images were used for the purpose of testing.

Approach:

These images clicked were then annotated using the Labelme graphical tool for annotations available in python. Each of the annotation for images were stored in JSON format. These images were then converted to the COCO format using the COCO API. A list of classes was then prepared containing the names of objects

in the images. Detectron DatasetCatalog was registered and MetadataCatalog was initiated with class labels for both training and testing data. The trained image were fetched as a dictionary and visualized with annotations. The dectron

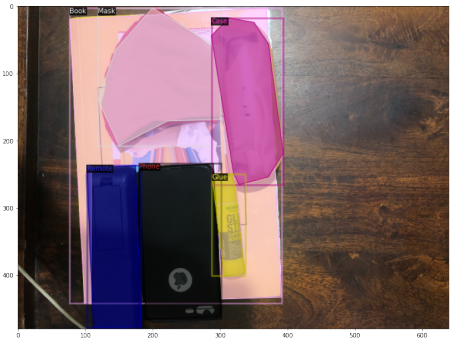

Fig. 5. Visualizing Train Images

Fig. 6. Visualizing Train Images

engine trained Mask R-CNN named COCO Image Segmentation Mask R-CNN X 101 32*8d FPN 3x is made use of for the purpose of training. This R-CNN uses FPN in the backbone 4 conv layers. The kernel sizes used are 1*1, 3*3 and 7*7 in different layers with max pooling. This a bottleneck block in the architecture and ResNet is used for the task of weight assignment.This Mask R-CNN runs 1197 iterations in a training time 25 minutes. The test images are then fed to the training model to generate object detection masks using dectron visualizer and colormode.

IV. Results

Faster R-CNN:

The test images are generated with objection detected label with bounding box. In the image shown below, the objects present in the image were a mobile phone,a person,a pen. The model identified all four objects correctly with a 80 percent confidence and returned their bounding box along with the label.

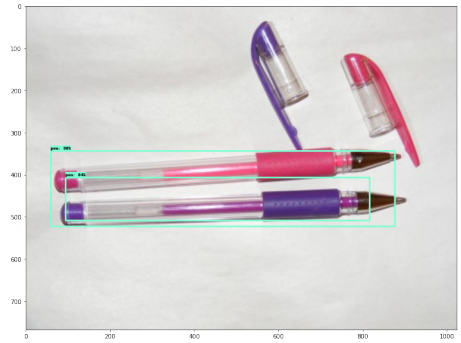

Fig. 7. Faster R-CNN:Object Detected Results of Test Image

Fig. 8. Faster R-CNN:Object Detected Results of Test Image

Mask R-CNN:

The test images are generated with objection detected label and their mask. In the image shown below, the objects present in the image were a box, a mask, a remote and a mobile phone. The model identified all four objects correctly with a 99 percent confidence and returned their masks.

In the image shown below, the objects present in the image were a box, biscuit, a remote, mobile phone, snack, glue, keys and facewash. The model identified all objects correctly with a 98 to 99 percent confidence and returned their masks.

The results generated by the Mask R-CNN detected objects with a minimum confidence of 98 percent.

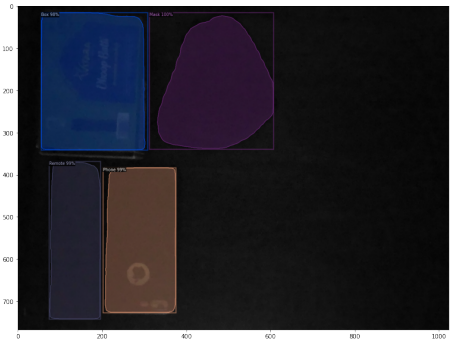

Fig. 9. Mask R-CNN:Object Detected Results of Test Image

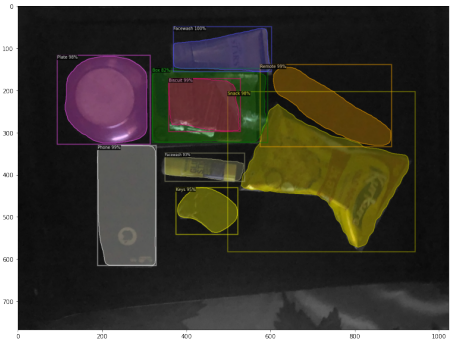

Fig. 10. Mask R-CNN:Object Detected Results of Test Image

V. Conclusion

While Faster R-CNN is a good model that can be used for object detection, Mask R-CNN combines the Faster R-CNN and a Fully Convolutional Networks into one mega architecture giving improved results along with the identified object masks. The Mask R-CNN implemented on our custom dataset detects objects with a minimum 98 percent confidence. Making use of the transfer learning approach for training the dataset using these pre trained networks not only reduces the computationally complexity and saves time, but also improves the results of object detection.

VI. References

1. Talukdar, Jonti, et al. ”Transfer learning for object detection using state-of-the-art deep neural networks.” 2018 5th International Conference on Signal Processing and Integrated Networks (SPIN). IEEE, 2018.

2. Bamne, Bulbul, et al. ”Transfer learning-based Object Detection by using Convolutional Neural Networks.” 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE, 2020.

3. Kaiyan, Zhu, Li Xiang, and Song Weibo. ”Underwater Object Detection using Transfer Learning with Deep Learning.” Proceedings of the 2020 International Conference on Computers, Information Processing and Advanced Education. 2020.

4. Shetty, Jyothi, and Pawan S. Jogi. ”Study on different region-based object detection models applied to live video stream and images using deep learning.” International Conference on ISMAC in Computational Vision and Bio-Engineering. Springer, Cham, 2018.

5. Lim, Joseph Jaewhan. Transfer learning by borrowing examples for multiclass object detection. Diss. Massachusetts Institute of Technology, 2012.

6. Shin, Hoo-Chang, et al. ”Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning.” IEEE transactions on medical imaging 35.5 (2016): 1285-1298.

7. Achraf Khazri, ”Faster RCNN Object Detection”, Towards Data Science, 2019.

8. Umer Farooq, ”From R-CNN to Mask R-CNN”, Towards Data Science, 2018.

9. https://storage.googleapis.com/openimages