Assignment 4: Object Detection Models Survey

Moustafa Abd El Haliem Eltawy 19016662
Abdel Rahman Ahmed Bahaa 19015882
Louay Nasr Zahran 19016198

December 2023

1 Introduction

Object detection problem has become a famous classic problem which is solved using many ways .In this report three state of the art models are represented: SSD , Faster-RCNN and RetinaNet.Moreover, a comprehensive analysis is made by using inference on the validation sets of COCO and VOC datasets

2 Datasets

2.1 Common Objects in Context(COCO)

Common Objects in context dataset(COCO)[1]is a large-scale dataset designed for object detection, segmentation, and captioning tasks in computer vision. Developed by Microsoft Research, COCO is widely used as a benchmark for evaluating and advancing algorithms in these areas.COCO focuses on complex scenes with multiple objects, enabling researchers to address challenges related to object recognition and segmentation.It includes a diverse range of object categories, such as people, animals, vehicles, and household items.The 2017 version contains 5000 images for validation.

2.2 The PASCAL Visual Object Classes(VOC)

The PASCAL Visual Object Classes(VOC)[2] is a benchmark dataset in computer vision, primarily used for object recognition and detection tasks. The PASCAL VOC challenges were initiated to encourage research and evaluate algorithms in the field. 1000 images from the validation set were taken

[1]https://cocodataset.org/home
[2]http://host.robots.ox.ac.uk/pascal/VOC/

3 Model Architectures

3.1 Faster R-CNN

In the R-CNN family of papers,CPU based region proposal algorithms, like: Selective search algorithm which takes around 2 seconds per image and runs on CPU computation. The Faster R-CNN introduces a new region proposal network instead of region selection algorithms which reduces time to milliseconds per image and also allowed layer sharing with the following detection stages.
The region proposal network (RPN) starts with the input image being fed into the backbone convolutional neural network.The backbone is usually a pretrained model like: VGG, ResNet.For every point in the output feature map, the network has to learn whether an object is present in the input image at its corresponding location and estimate its size. This is done by placing a set of "Anchors" on the input image for each location on the output feature map from the backbone network. These anchors indicate possible objects in various sizes and aspect ratios at this location.Then network checks whether these anchors actually contain objects, and refine these anchors' coordinates to give bounding boxes as "Object proposals" or regions of interest.An anchor is considered to be a "positive" sample if The anchor has has an IoU greater than 0.7 with any groundtruth box.if its IoU with all groundtruth boxes is less than 0.3, it is considered negative. The remaining anchors (neither positive nor negative) are disregarded for RPN training.The output of the RPN is passed to the region of interest pooling layer (ROI) in Fast R-CNN and the weights of the Backbone network of RPN is shared with the backbone network of Fast RCNN(Layer sharing).The output from the ROI pooling layer has a size of (N, 7, 7, 512) where N is the number of proposals from the region proposal algorithm. After passing them through two fully connected layers, the features are fed into the sibling classification and regression branches.Fig1 shows the Faster-RCNN Network architecture.

3.1.1 Single Shot Detector (SSD)

Single shot detector is a single stage detector which is much faster than multi-stage detectors like Faster-RCNN. SSD uses a backbone VGG-16 model for feauture extraction.Then,image is divided into cells where each cell is responsible to detect objects in it's region to provide more localization. Additional CNN layers are used to detect 8736 bounding box for object. The bounding boxes are then filtered by using IOU and applying Non-max suppression.Fig2 shows the SSD network

3.2 Retina Net

Retina net is considered one of the state of the art single stage models.RetinaNet uses a Feature Pyramid Network (FPN) as its backbone architecture. FPN enhances the model's ability to detect objects at different scales by creating a

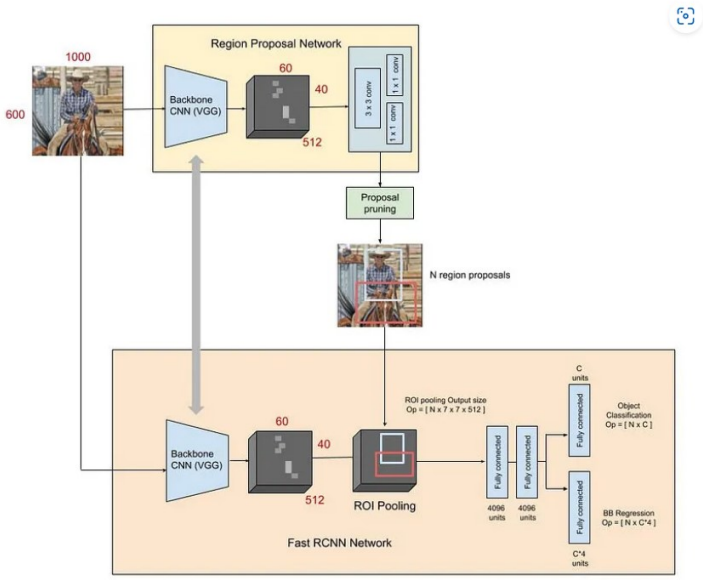

Figure 1: Faster RCNN network

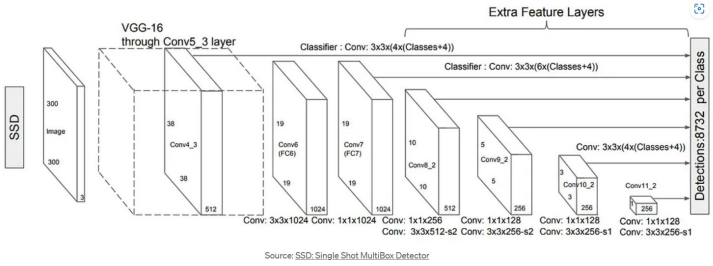

Figure 2: SSD network

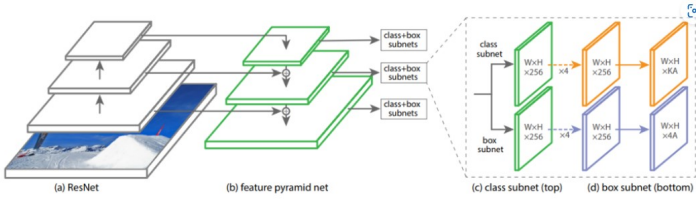

Figure 3: Retina net network

feature pyramid with high-level semantic information and detailed spatial information.RetinaNet utilizes anchor boxes to predict bounding boxes for potential object locations. However, unlike traditional anchor-based methods, RetinaNet introduces a set of anchors with a wide range of scales and aspect ratios at each pyramid level.Moreover,Retina net introduces Focal Loss, a novel loss function that addresses the issue of class imbalance in object detection. Focal Loss down-weights well-classified examples and focuses on hard, misclassified examples, thereby improving the model's ability to handle imbalanced datasets.Fig3 shows the Retina Net network

4 Evaluation metrics

4.1 Intersection over Union (IOU)

IOU is calculated as the ratio of the area of overlap between predicted and ground truth bounding boxes to the area of their union. A higher IOU signifies more proximity to the true box

4.2 Mean Average Precision (mAP)

mAP is a comprehensive metric that considers precision-recall curves across different object categories.It is the mean of the Average Precisions(AP) across all classes in the dataset.The mAP score ranges from 0 to 1, with a higher score indicating better overall performance.

5 Analysis

5.1 Evaluation results

table 1 shows the results of the models during inference of COCO dataset. table 2 shows the results of the models during inference of Pascal VOC dataset.

from the results it is observed that SSD is the fastest while Faster R-CNN has the highest MAP and IOU. Also note that these numbers are based on a

model name	IOU	MAP	samples/s
Faster R-CNN	0.36696651106773465	0.34474258303522526	14.0
SSD	0.07270479127449074	0.18150702503806054	28.0
Retina Net	0.07270479127449082	0.29717884699303193	18.4

Table 1: Inference Results on COCO dataset

model name	IOU	MAP	samples/s
Faster R-CNN	0.2550200854765254	0.35635901036310746	35.7
SSD	0.053130498461261055	0.2807325824312871	90.9
Retina Net	0.05313049846126114	0.35878097826866184	52.9

Table 2: Inference Results on VOC dataset

confidence level threshold of 0.5. More results are in the notebooks provided with the report.

5.2 Feature maps

This section is used to show Fast R-CNN feature maps for a sample image.More samples are provided with the report in /feature maps directory

5.3 Ablation Studies

5.3.1 Faster R-CNN

1. Most accurate of all three models
2. Multi-stage model
3. Can detect drawings as shown in Fig.7
4. weakness: A lot of output boxes refer to parts not all of the object, figures 6 & 8 shows this

5.3.2 SSD

1. Fastest of all three models
2. Single-stage model
3. can predict very localized boxes as shown in 9 ,10 & 11
4. weakness: Not accuracte,bad for images with many objects as shown in figures 12 & 13

Figure 4: Image

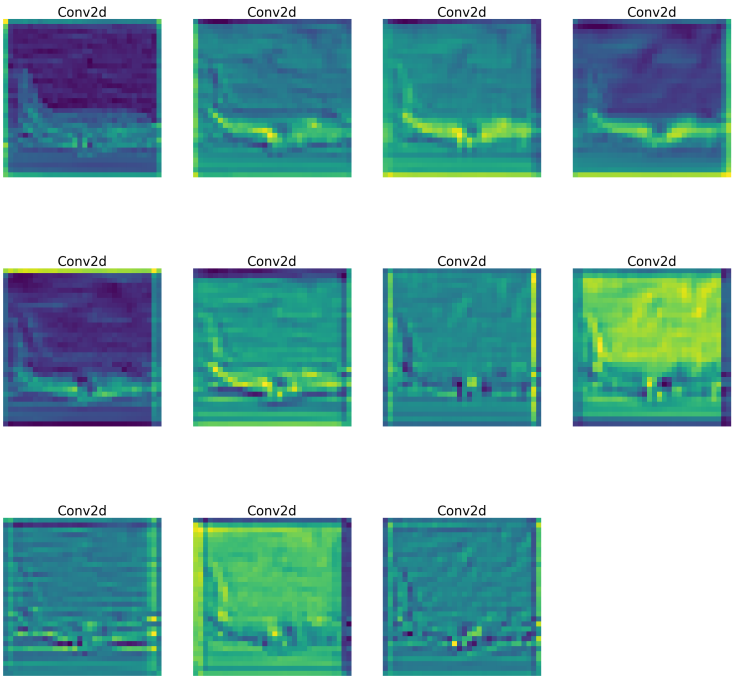

Figure 5: feature map

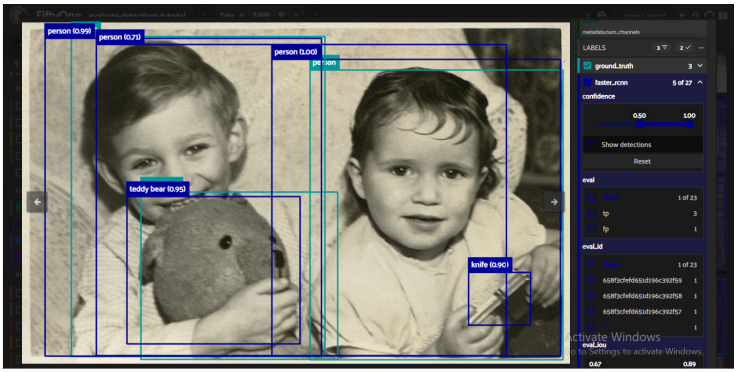

Figure 6: detected part of the person as a whole person thus detecting 3 persons

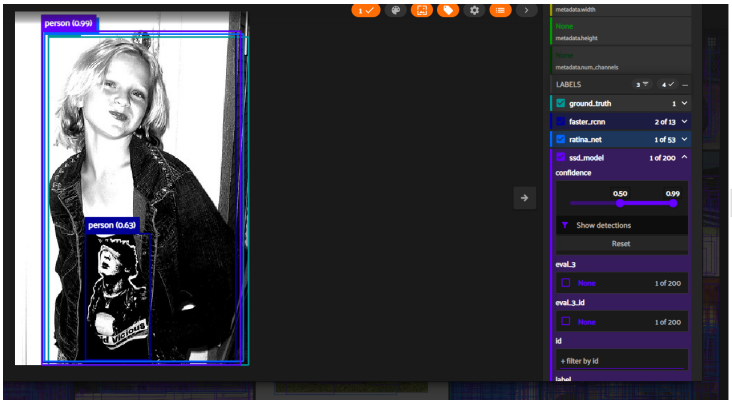

Figure 7: detected person drawing on the shirt

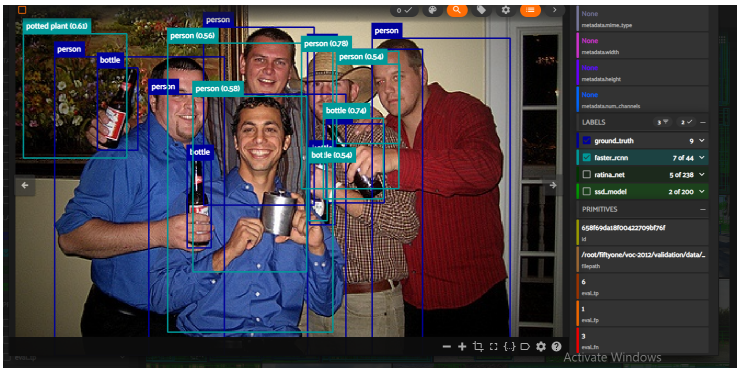

Figure 8: A lot of the boxes are parts of people not people themsolves

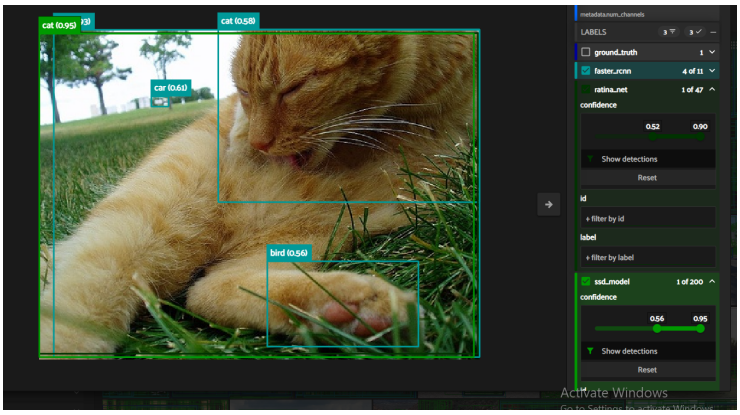

Figure 9: success case

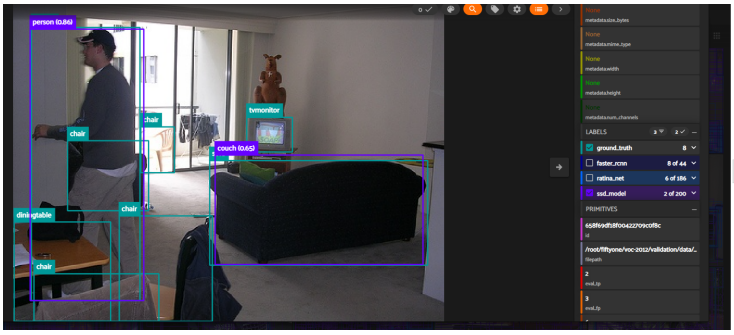

Figure 10: success case

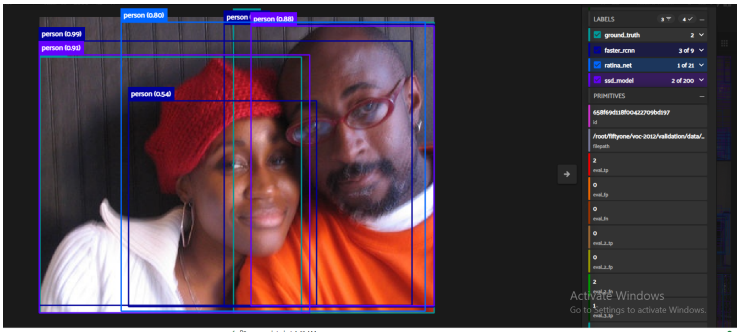

Figure 11: success case

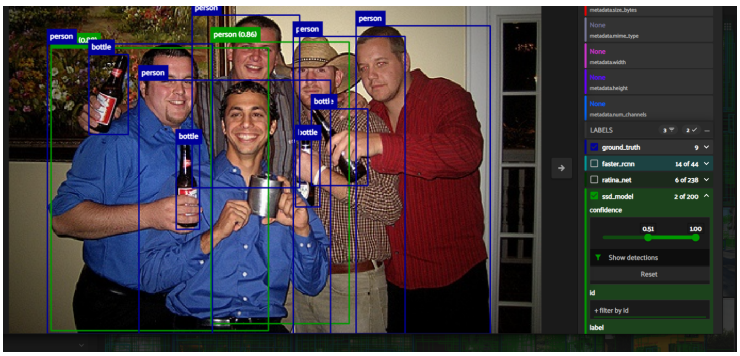

Figure 12: fail case

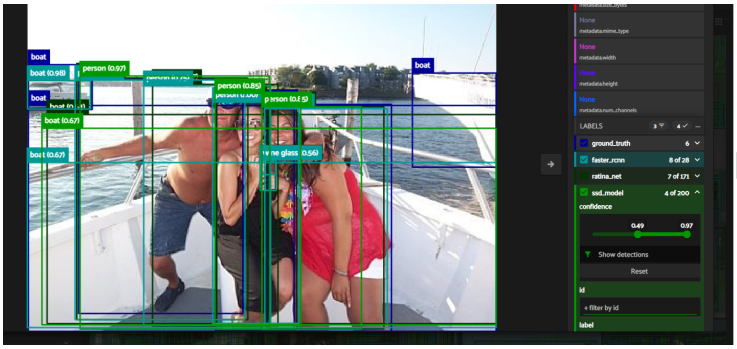

Figure 13: fail case

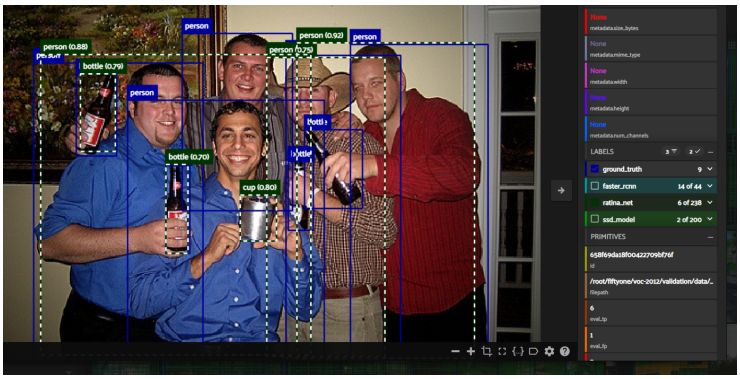

Figure 14: All people are inside 2 objects

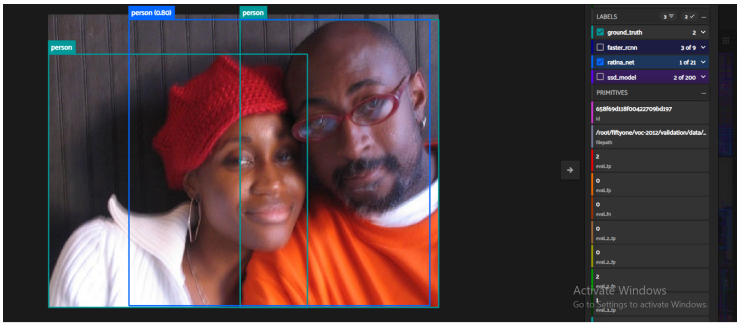

Figure 15: All people are inside 1 object

5.3.3 Retina net

1. State of the Art for Single-stage models
2. The use of Feature Pyramid Network(FPN) improved the accuracy
3. Introduced novel metric : Focal Loss
4. weakness: Large bounding Boxes(unlike SSD) which is shown in fig.14 and fig.15

6 Conclusion

To summarize ,Table 3 concludes all info gathered in this survey

Properties	**Faster R-CNN**	**SSD**	**Retina Net**
Accuracy	Highest	Low	High
Number of Stages	Multi	Single	Single
Localization	Medium	Good	Bad
FPN	NO	NO	Yes
Focal loss	NO	NO	Yes
Speed	Slow	Fastest	Fast

Table 3: Conclusion Table