Assignment 1 Report

group 131

Task 1

Task 1. Object Detection Metrics

For evaluating image classification methods, accuracy is an easy and good metric. For object detection, measuring performance is a more complicated task. First, we need to localize the object correctly with a good bounding box. Once we have localized the object, we want to ensure the localized object is classified correctly. When the PASCAL VOC dataset was proposed, they introduced a standard metric to measure the performance of object detection architectures. This metric was *mean average precision (mAP)*.

Take a look at the recommended resources, and **answer the following in your report:**

(a) [2*pts*] Explain the Intersection over Union and how we can find it for two bounding boxes. Illustrate it with a drawing.

(b) [2*pts*] Write down the equation of precision and recall, and shortly explain what a true positive and false positive is.

(c) [4*pts*] Given the following precision and recall curve for the two classes, what is the mean average precision?

Precision and recall curve for class 1:
$\text{Precision}_1 = [1.0, 1.0, 1.0, 0.5, 0.20]$
$\text{Recall}_1 = [0.05, 0.1, 0.4, 0.7, 1.0]$

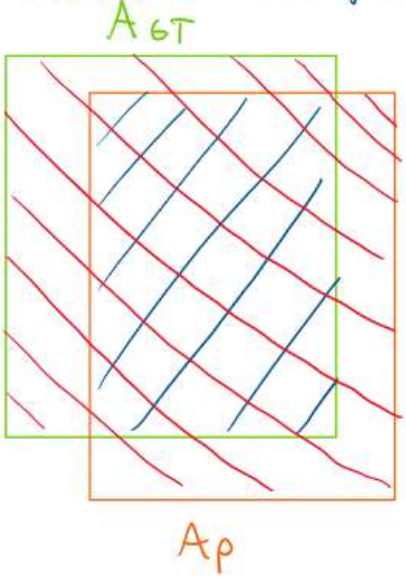
Precision and recall curve for class 2:
$\text{Precision}_2 = [1.0, 0.80, 0.60, 0.5, 0.20]$
$\text{Recall}_2 = [0.3, 0.4, 0.5, 0.7, 1.0]$

Hint: To calculate this, find the precision for the following recall levels: 0.0, 0.1, 0.2, ... 0.9, 1.0.

task 1a)

a) Intersection over union is a way to score how much a predicted bounding box over laps with the ground truth one. The IoU is calculated by deviding the area that is overlapping with the union of the two.

$$IoU = \frac{A_p \cap A_{GT}}{A_p \cup A_{GT}}$$



Blue is $A_p \cap A_{GT}$

Red is $A_p \cup A_{GT}$

There is often a threshold of IOU before a bounding box prediction is considered a correct guess.

task 1b)

N) Precission : $\frac{TP}{TP+FP}$

Recall : $\frac{TP}{TP+FN}$

TRue Positive (TP) : is when a predicting box is Correct in relation to the Ground truth.

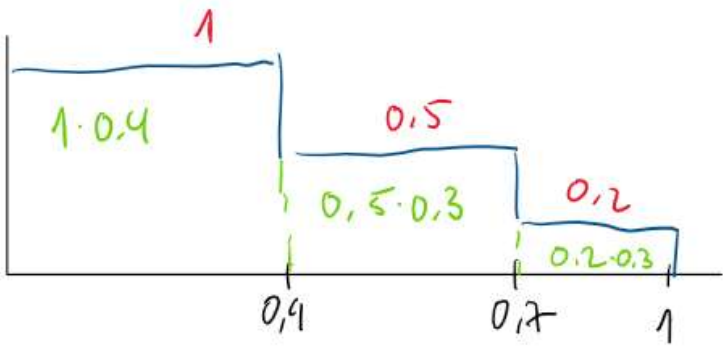
False Positive (FP): is when the network predicts a positive that does not match with the Ground truth

False Negative (FN):

a False negative is when there is no positive predicton on the positive ground truth.

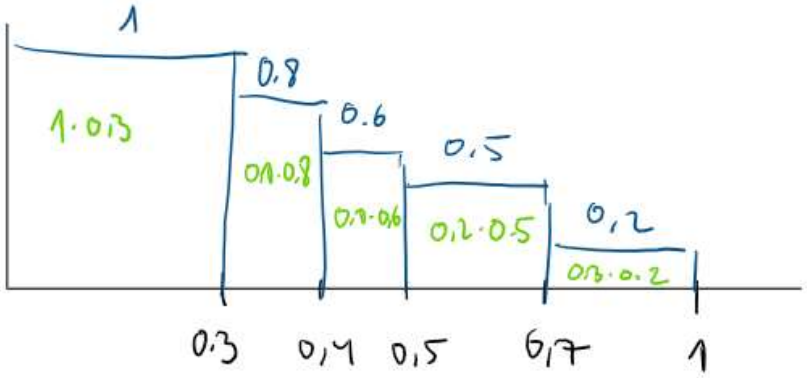
task 1c)

c) class 1:



0.4 + 0.15 + 0.06 = 0.61 ⇒ 61%

class 2:



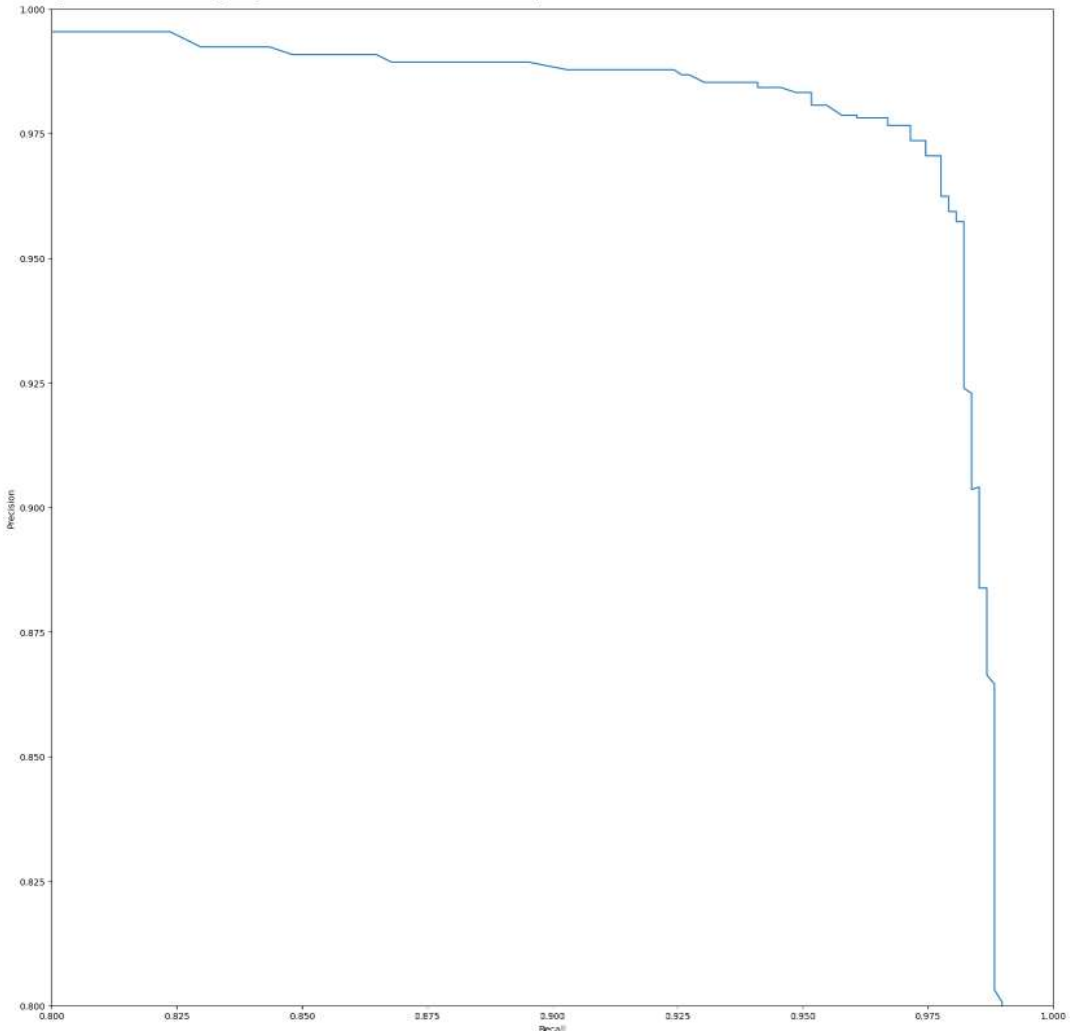
0.3 + 0.08 + 0.06 + 0.1 + 0.03 + 0.06 = 0.63 ⇒ 63%

$$mAP = \frac{1}{2}(0.61 + 0.63) = 0.61$$

Task 2

Task 2f)

```
length of precisions:  500
Mean average precision: 9.9771
```



Task 3

Task 3a)

a) This filtering is called non-Maximum supression. This is we only want to use the box with the best score.

Task 3b)

b) False: Deeper layers have smaller resolution and thus better to detect larger structures for smaller objects layers with higer resolution (shallow layers)

Task 3c)

c) This is so the network can detect classes with diffent aspect ratios so the network can catch matches in all diffent ratios, sizes and locations.

Task 3d)

d) SSD creates multiple Box proposal at all cells/positions to better catch all classes in all forms.
YOLO Predicts only one box per grid/position.

Task 3e)

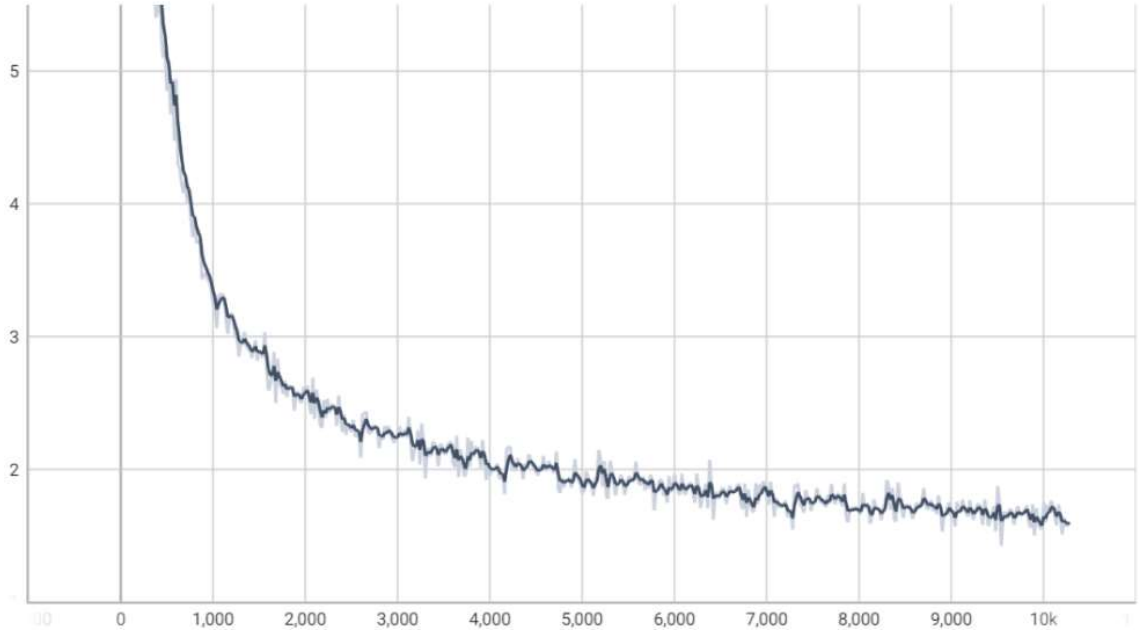
e) $38 \times 38 \times 6 = 8664$ Boxes.
in this feature map.
If we place an anchor in each
cell in this layer

Task 3f)

f) $38 \cdot 38 \cdot 6$
$+ 19 \cdot 19 \cdot 6$
$+ 10 \cdot 10 \cdot 6$
$+ 5 \cdot 5 \cdot 6$
$+ 3 \cdot 3 \cdot 6$
$+ 1 \cdot 1 \cdot 6$
$= 11640$

Task 4

Task 4b)



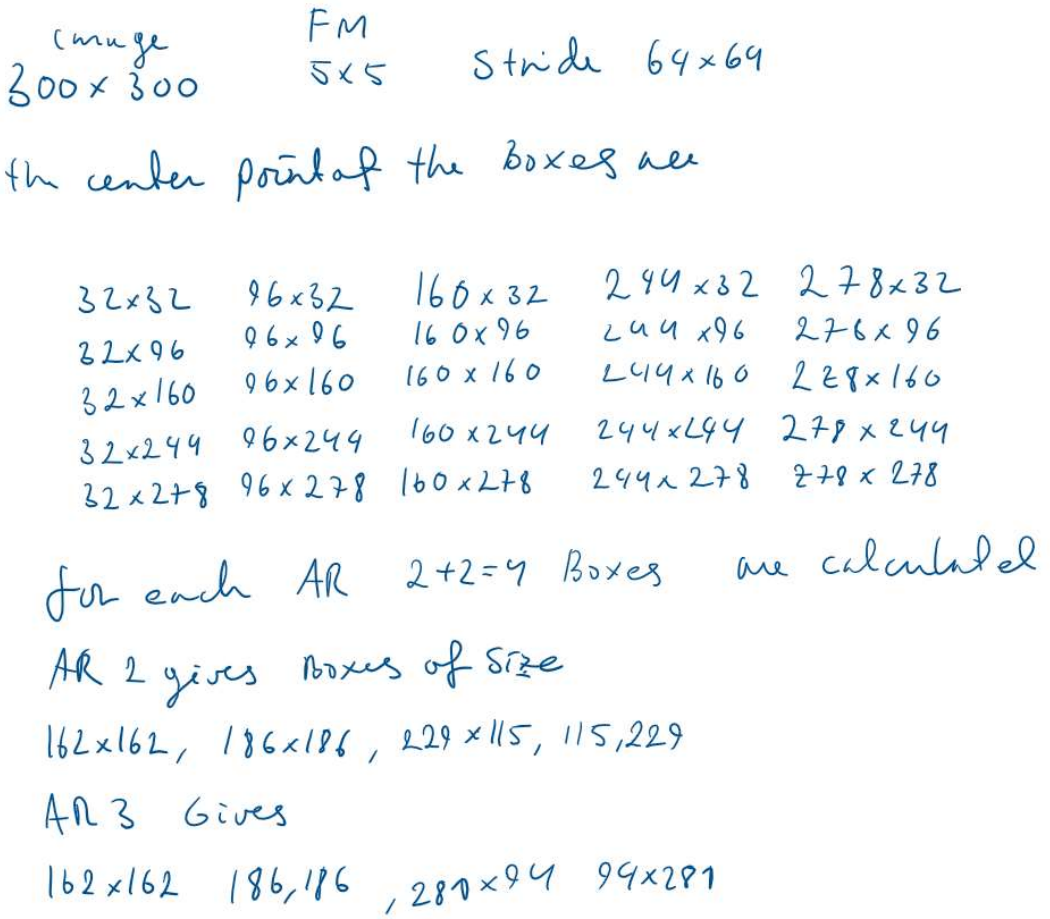
The Mean average precicion after 6000 iterations was 0.7394

Task 4c)

Added some image aumentation with image flips and rotations using the gpu

Making the network deeper with more convolutions, by either expanding the existing layers or adding more layers. for example you could add a new layer with 75x75 output before the 38x38 layers

Task 4d)



Task 4e)

Task 4f)