KIPRONO, Elijah Koech

(kiprono@aims.ac.za)

October 20, 2020

1. Experiments

1.1 Dataset Description

The dataset used in this project consist of natural outdoor images of farm trees and the aim is to detect all the fruits in every image. This dataset was obtained from 3 sources:

i Aerobotics dataset- The images are sourced by flying the drone $\sim 2m$ above the tree canopy which generated $2.7k$ video imagery. These short videos are then segmented to generate images.

ii Fuji dataset[1] (Gen'e-Mola et al. (2019)) - This dataset is composed of 967 images of apple trees captured using Microsoft Kinect v2.

iii ACFR dataset[2] (Bargoti and Underwood (2016a) and Bargoti and Underwood (2016b)) - This dataset was collected by a team at Australian Centre for Field Robotics (ACFR), The University of Sydney, Australia. The orchard images were gathered using Shrimp which was built and is being maintained by ACFR. The dataset comprises of apples, mangoes and almond trees.

The total number of images collected from these three sources are 2853 - 2081 images used for training the model and 772 for testing. On the train set, a total of 12206 fruits were labelled whereas on the test set, 3459 fruits were labelled. Annotation/labelling was done manually using VGG Image Annoatator (VIA) Dutta and Zisserman (2019), Dutta et al. (2016) with each fruit labelled with a cirlce - center, (x, y) and the radius defined.

1.2 Evaluation Metrics

Objected detection models are often evaluated using confusion matrix. Out of confusion matrix components other perfomance metrics can be generated. These metrics include: precision, recall, F1 score, Average Precison (AP) and mean Average Precision (mAP). At a low-level, measuring the perfomance of a object detector involves detemining if a detection is valid or not.

Definitions

i True Positive (TP) - A valid detection.

ii False Positive (FP) - An invalid detection.

iii False Negative (FN) - A ground-truth model missed by the model.

iv True Negative (TN) - It is a negative instance that should not be detected. This metric does not apply to object detection problems because there are infinitely many regions within an image that should not be mapped as an object by the detector.

In the context of determining the validity of a detection (predicted mask), a supporting metric called Intersection over Union (also called Jaccard Index) is needed.

[1] FUJI dataset: https://zenodo.org/record/3715991
[2] ACFR Orchard Fruit Dataset: http://data.acfr.usyd.edu.au/ag/treecrops/2016-multifruit/

1.2.1 Intersection over Union (Everingham et al., 2015).

IoU is a metric that evaluates the degree of overlap between ground-truth mask (M_{gt}) and the predicted mask (M_{pd}). It is calculated as the area of intersection between M_{gt} and M_{pd} divided by the area of union of the two, that is,

$$\text{IoU} = \frac{\text{area}(\mathrm{M}_{gt} \cap \mathrm{M}_{pd})}{\text{area}(\mathrm{M}_{gt} \cup \mathrm{M}_{pd})} \tag{1.2.1}$$

IoU metric ranges from 0 to 1 with 0 signifying no overlap and 1 implying perfect overlap between M_{gt} and M_{pd}.

With IoU metric, we need to define a threshold (α, say) that is used to distinguish a valid detection from the one which is not. We can, therefore, redefine TP (correct detection) as a detection for which IoU$\geq \alpha$ and FP (incorrect detection) with IoU$< \alpha$. FN is a ground-truth missed by the model.

Example

At IoU threshold, $\alpha = 0.5$ (or 50%), we can define TP, FP and FN as shown in the Figure 1.1 below

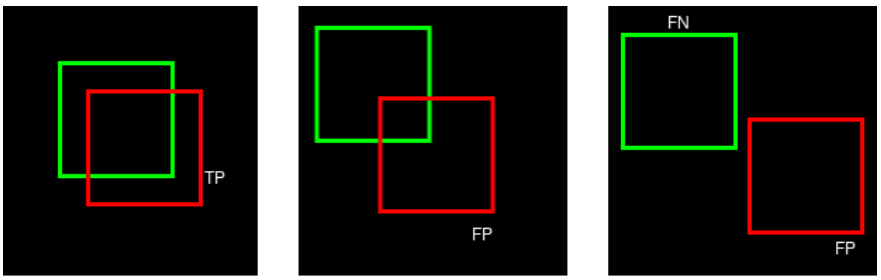


Figure 1.1: **[Best viewed in color]** Diagrammatic definition of TP, FP and FN. The red bounding boxes are the predictions while green boxes are the ground-truths.

Note: If we raise IoU threshold above 0.86 ,the instance on the first image will be a FP and if we lower IoU threshold below 0.24, the second instance becomes TP. FN is not determined through IoU thresholding.

As stated earlier, TN metric is not applicable to object detection problems and therefore one desists from using metrics that is based on this confusion matrix component such as True Negative Rate (TNR), False Positive Rate (FPR), Negative Predictive Value (NPC) and Receiver Operating Characteristic (ROC) curve. Instead, the performance evaluation of object detection models is based on Precision (P), Recall (R) and F_1 score which are defined as

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}} = \frac{TP}{\text{all detections}} \tag{1.2.2}$$

$$R = \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{TP}{\text{all ground-truths}} \tag{1.2.3}$$

$$F_1 = 2\left(\frac{P \times R}{P + R}\right) \tag{1.2.4}$$

Precision is the ability of a classifier to identify relevant objects only. It is the proportion of true positive detections. Recall, on the other hand, measures the ability of a model to find all relevant cases (that is, all ground-truths) - the proportion of true positives detected among all ground-truths. F_1 score is the harmonic mean of precision and recall. It is used to establish balance between the two.

A good model is a model that can identify most ground-truth objects (high recall) while only finding the relevant objects (high precision) often. A perfect model is the one with FN=0 (recall=1) and FP=0 (precision=1). The former is usually the objective, the latter often unattainable.

The precision-recall (PR) curve is a plot of precision as a function of recall. It shows trade-off between the two metrics for varying confidence values for the model detections.

If the number of FPs is low, precision will be high but the number of missed objects may also increase implying a high number of FNs and consequently low recall. On the contrary, lowering IoU threshold in order to accept more TPs, may yield less FNs - high recall but at the same time increase the number of FPs - low precision score. A good model is the one in which both precision, recall and consequently F_1 score remains high even if confidence threshold varies (Padilla et al., 2020).

Evaluation of Area Under the PR Curve (AUC-PR) yields another metric: Average Precision (AP). Mathematically, AP is defined as

$$AP\alpha = \int_0^1 p(r)\, dr \tag{1.2.5}$$

Notation: AP@α or APα means AP evaluated at α IoU threshold so that AP30 and AP50 are AP values at 30% and 50% IoU thresholds respectively.

A high AUC-PR implies high precision and high recall. Naturally, often, PR curve is a zigzag-like plot. For this reason, AP is approximated instead evaluating the exact value by solving the integral in Equation 1.2.5 analytically. There are two interpolation methods used to achieve this: (i) 11-point interpolation and (ii) All-point interpolation.

A 11-point PR curve is a plot of interpolated precision scores for a model results at 11 equally spaced standard recall levels, namely, $0.0, 0.1, 0.2, \ldots 1.0$.

$$\text{AP}\alpha_{11} = \frac{1}{11} \sum_{r \in R} p_{interp}(r), \tag{1.2.6}$$

where,

$R = \{0.0, 0.1, 0.2, \ldots 1.0\}$ and $p_{interp}(r) = \max_{r':r' \geq r} p(r')$, that is, interpolated precision at recall value, r, is defined as the highest precision for any recall value $r' \geq r$ Padilla et al. (2020).

Unlike 11-point, all-point interpolation interpolates through all the positions, that is,

$$\text{AP}\alpha = \sum_i (r_{i+1} - r_i) p_{interp}(r_{i+1}), \tag{1.2.7}$$

where, $p_{interp}(r_{i+1}) = \max_{r':r' \geq r_{i+1}} p(r')$ Padilla et al. (2020).

Remark. ***[Remark (AP and the number of classes)]*** *AP is calculated for individual class. In this context, there are as many AP values as the number of classes. These AP values can be averaged to obtain another metric: mean Average Precision (mAP)*

$$\text{mAP}\alpha = \frac{1}{n}\sum_{i=1}^{n}\text{AP}_i \quad \text{for n classes.} \tag{1.2.8}$$

Remark. ***[Remark (AP and IoU)]*** *AP is calculated at a given IoU threshold. With this reasoning, AP can be calculated over a range of thresholds. Therefore, in this context, AP can be reported as the average of AP values obtained at different thresholds. Microsoft COCO, Lin et al. (2014), calculates AP of a given object at 10 different IoU ranging from 50% to 95% at 5% step-size, usually denoted, AP@50:5:95.*

Example

Consider Figure 1.2 containing 3 images with 12 detections (red bounding boxes) and 9 ground-truths (green boxes). Each detection is labelled with a letter and confidence of the prediction. In this example we are considering that all the ground-truths are of the same class and use IoU threshold, $\alpha = 50\%$. The IoU for each detection-truth pair is indicated in Table 1.1. The columns cumTP and cumFP are cumulative values for TP and FP columns respectively. It accumulate TP and FP values above the corresponding confidence level.

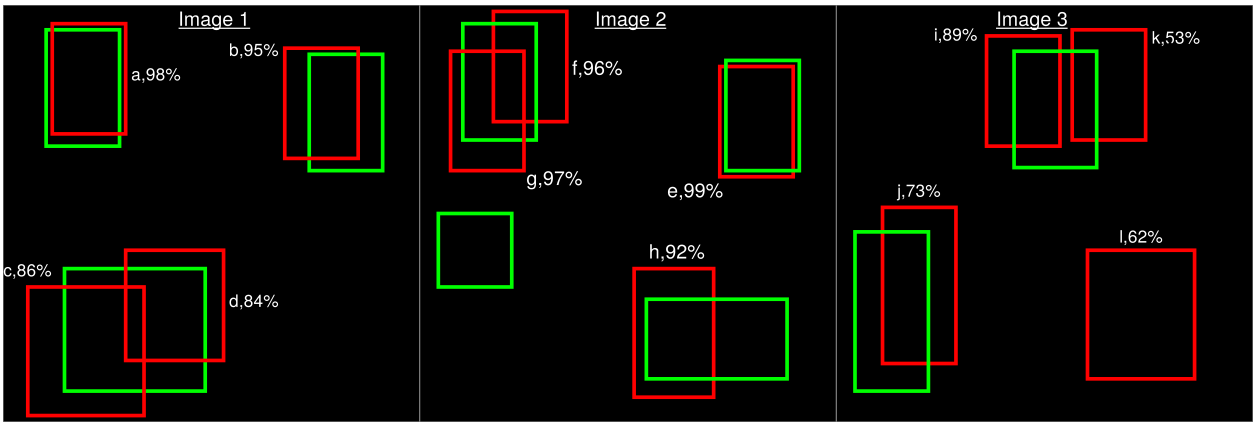


Figure 1.2: A model detecting objects of the same class. There are 12 detections and 9 ground-truths.

Remark (Multiple detections): In some cases, there are multiple detections overlapping one ground-truth, e.g. c,d in image 1, g,f in image 2 and i,k in image 3. For such cases, a detection with the highest confidence is considered as TP and the rest of the detections as FPs. This is, however, conditioned on the IoU threshold. The detection with the highest confidence should have IoU $>$ threshold otherwise all detections are FPs Everingham et al. (2015).

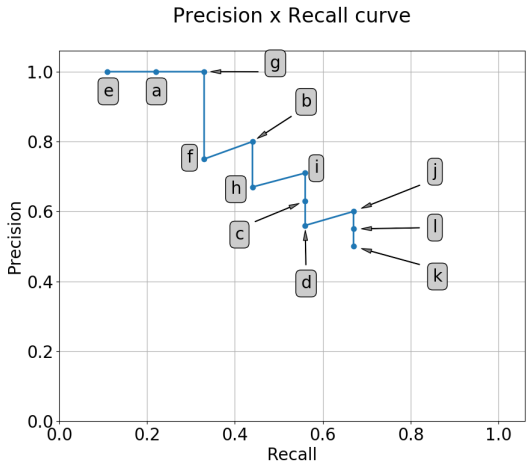
detection	confidence	TP	FP	cumTP	cumFP	all_detections	precision	recall	IoU
e	99	1	0	1	0	1	1	0.11	97
a	98	1	0	2	0	2	1	0.22	92
g	97	1	0	3	0	3	1	0.33	87
f	96	0	1	3	1	4	0.75	0.33	82
b	95	1	0	4	1	5	0.8	0.44	73
h	92	0	1	4	2	6	0.67	0.4	48
i	89	1	0	5	2	7	0.71	0.56	66
c	86	0	1	5	3	8	0.63	0.56	47
d	84	0	1	5	4	9	0.56	0.56	42
j	73	1	0	6	4	10	0.6	0.67	54
l	62	0	1	6	5	11	0.55	0.67	45
k	53	0	1	6	6	12	0.5	0.67	38

Table 1.1: Computation of precision and recall with IoU threshold, $\alpha = 50\%$. IoU is to determine whether a detection is TP or FP. cumTP and cumFP are cumulative values for TP and FP columns respectively.

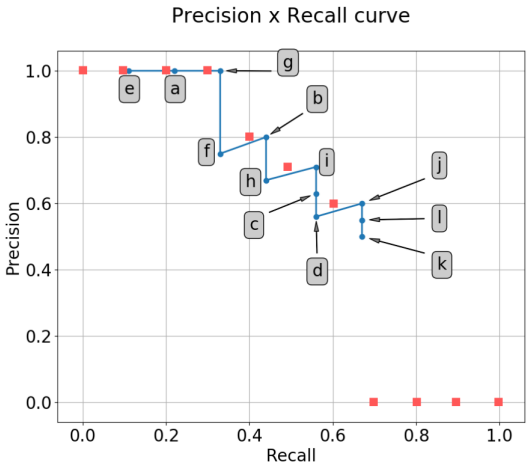
11-point interpolation

To calculate the approximation of AP50 using 11-point interpolation, we need to note the precision values for recall values in R (see Equation 1.2.6) as shown in Figure 1.3b

$$AP50_{11} = \frac{1}{11}(1 + 1 + 1 + 1 + 0.8 + 0.71 + 0.6 + 0 + 0 + 0 + 0) = 55.55\% \tag{1.2.9}$$



(a) PR curve with precision and recall values plotted for all detections.



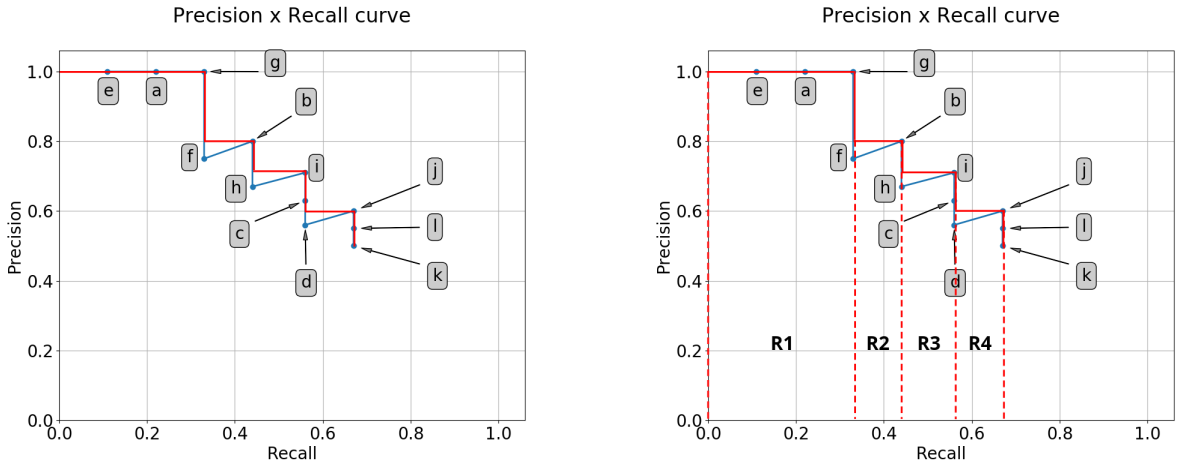
(b) PR curve for 11-point interpolation approach.

Figure 1.3: Precision-Recall curves

All-point interpolation

From the definition in Equation 1.2.7, we can calculate AP50 using all-point interpolation as follows

$$AP@50 = 1 * (0.33 - 0) + 0.8 * (0.44 - 0.33) + 0.71 * (0.56 - 0.44) + 0.6 * (0.67 - 0.56) \quad (1.2.10)$$
$$= 56.92\% \quad (1.2.11)$$



(a) All-point interpolation approach

(b) Regions of all-point interpolation approach.

Figure 1.4: Precision-Recall curves

Put simply, all-point interpolation involves calculating and summing area of the 4 regions (R1, R2, R3 and R4) in Figure 1.4b, that is,

$$AP@50 = \text{Area}(R1) + \text{Area}(R2) + \text{Area}(R3) + \text{Area}(R4), \quad (1.2.12)$$

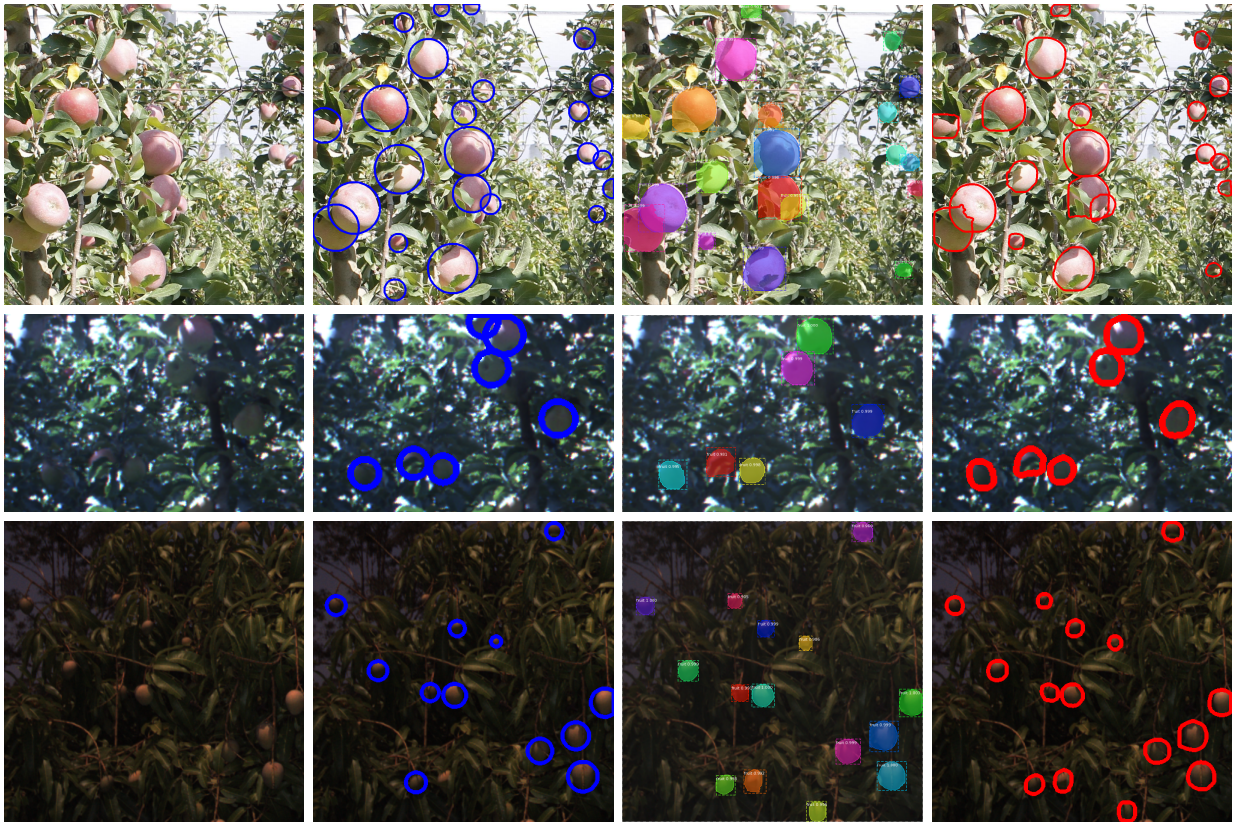
where,

$$\text{Area}(R1) = 1 * (0.33 - 0) = 0.33$$
$$\text{Area}(R2) = 0.8 * (0.44 - 0.33) = 0.088$$
$$\text{Area}(R3) = 0.71 * (0.56 - 0.44) = 0.0852$$
$$\text{Area}(R4) = 0.6 * (0.67 - 0.56) = 0.066$$

and thus $AP@50 = 56.92\%$.

1.3 Results of Mask R-CNN

The output of Mask R-CNN includes segmentation masks, mask confidence and bounding-boxes around the objects (as shown on the Figure to the right and Figure 1.5c). From the segmentation masks, we generated the mask contours (see Figure 1.5d). These masks and the ground-truth labels are then used to determined if a detection is valid or not. For this reason, IoU refers to mask IoU and not bounding box IoU unless otherwise stated.



(a) RGB Images (b) Truth labels (c) Mask R-CNN output (d) Segmentation masks

Figure 1.5: **[Best viewed in color]** Sample results of Mask R-CNN.(a) RGB images - input of Mask R-CNN. (b) Annotations created manually using VGG annotator. (c) Mask R-CNN output - bounding box, segmentation mask and mask confidence. (d) Segmentation masks exracted from Mask R-CNN output.

Remark. ***[IoU thresholds]*** *We will consider four IoU thresholds throughout the study, that is,* 20%, 30%, 40% *and* 50%.

Mask R-CNN model perfomed very well in detection of fruits in both training and testing sets. We evaluated the detector using confusion matrix, AP30 and AP50. PR-curve is also plotted for the two tresholds.

Despite the good perfomance, Mask R-CNN model has its shortfalls. Mask R-CNN, in some cases, yielded multiple detections for a single fruit. To merge these detections, just like in PASCAL VOC (Everingham et al., 2015), we consider the detection with highest confidence as TP (provided that its IoU is greater than threshold) and regard all the other detections as FPs.

The output of Mask R-CNN suggests that the model could not differentiate between fruits and background due to too much occlusion and visual complexity of the tree canopy. Some leaves have the same features (e.g. colour and shape) as the fruits, and therefore, the model marked them as fruits - false positive cases. Similarly, some fruit surfaces are highly similar to the colour of the leaves. Due to these similarities, the model marked these fruits as non-fruits - missed cases. By inspection, green fruits matched the green leaves, causing the misclassification.

Another problem affecting the performance of the model is the illumination challenge. The images used in this project are captured under natural outdoor conditions making it prone to visual complexities. These complexities negatively affected the performance of Mask R-CNN.

1.3.1 Confusion Matrix.

Mask R-CNN model was evaluated by calculating the proportion of TP, FP and FN at a given IoU threshold. In this kind of evaluation, we only considered the masks with confidence above 90%.

Threshold (α)	Set	TP(%)	FP(%)	FN(%)
0.2	Train	91.48	8.51	7.82
	Test	88.25	11.75	15.40
0.3	Train	91.35	8.65	7.96
	Test	87.86	12.14	15.79
0.4	Train	91.16	8.84	8.14
	Test	87.17	12.83	16.48
0.5	Train	90.61	9.39	8.70
	Test	85.65	14.35	18.01

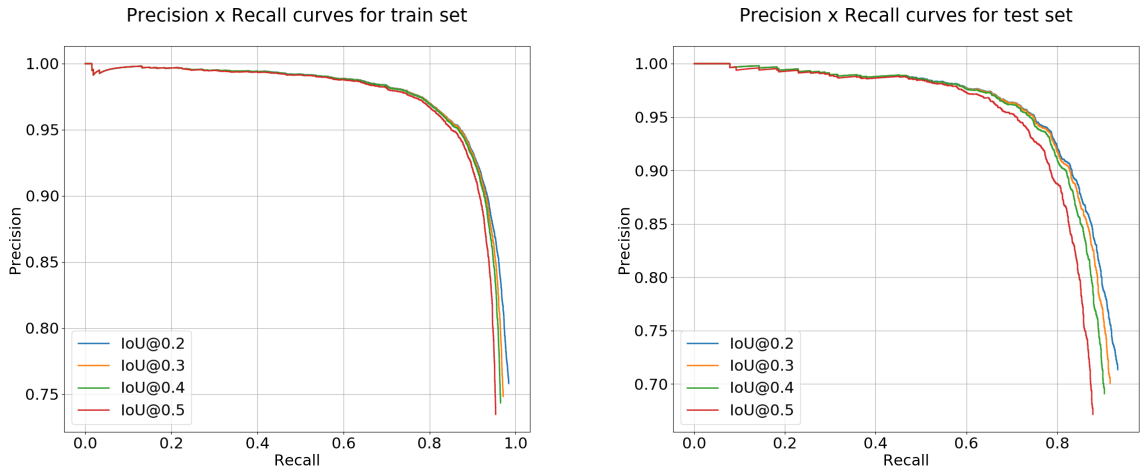
Table 1.2: Results of Mask R-CNN on the whole dataset for 4 different IoU thresholds. There are 12291 detections in the train and 3337 in the test set. During annotation, 12206 fruits were labelled in train set and 3459 in the test set. Each fruit was labelled with a circular annotation. For this reason, the evaluation of model was based on the segmentation mask and not the bounding box.

Remark. *$TP_\alpha(\%)$ and $FP_\alpha(\%)$ are calculated over the number of detections ($TP_\alpha + FP_\alpha$) while $FN_\alpha(\%)$ is calculated over the fruit count.*

For training set and IoU threshold of 0.5, 90.61% of the fruits were successfully detected, 9.39% of the detections were FPs. 8.70% fruits were missed by the model. The model achieved 85.65% TPs, 14.35% FPs and 18.01% FNs on the test set.

Remark. ***[Missed ground-truths]***. *These are ground-truths that are missed by the model (IoU=0) together with the cases of a single prediction overlapping multiple ground-truths. For example, if one prediction overlap 3 ground-truths, then, two of the truths with least overlap are considered FNs.*

1.3.2 Precision-Recall Curve.



(a) Train set.

(b) Test set.

Figure 1.6: **[Best viewed in color]** PR curves for IoU thresholds, $0.2, 0.3, 0.4$ and 0.5 for train and test set.

1.3.3 Average Precision.

As stated earlier, AP is a metric estimated by the AUC of the PR curves in Figure 1.6. In this project, AP is calculated with all-point interpolation method (discussed in page v) considering confidence between 50% and 100%.

Set	AP@20	AP@30	AP@40	AP@50
Train	0.9625	0.9514	0.9457	0.9344
Test	0.8997	0.8854	0.8732	0.8470

Table 1.3: Average Precision for Mask R-CNN output for 20%, 30%, 40% and 50% IoU thresholds.

References

Bargoti, S. and Underwood, J. Deep Fruit Detection in Orchards. *arXiv preprint arXiv:1610.03677*, 2016a.

Bargoti, S. and Underwood, J. Image segmentation for fruit detection and yield estimation in apple orchards. *To Appear in Journal of Field Robotics*, 2016b.

Dutta, A. and Zisserman, A. The VIA annotation software for images, audio and video. In *Proceedings of the 27th ACM International Conference on Multimedia*, MM '19, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-6889-6/19/10. doi: 10.1145/3343031.3350535. URL https://doi.org/10.1145/3343031.3350535.

Dutta, A., Gupta, A., and Zissermann, A. VGG image annotator (VIA). http://www.robots.ox.ac.uk/~vgg/software/via/, 2016. Version: 2.0.10, Accessed: Aug 20, 2020.

Everingham, M., Eslami, S. A., Van Gool, L., Williams, C. K., Winn, J., and Zisserman, A. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1): 98–136, 2015.

Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

Gen'e-Mola, J., Vilaplana, V., Rosell-Polo, J. R., Morros, J.-R., Ruiz-Hidalgo, J., and Gregorio, E. Kfuji rgb-ds database: Fuji apple multi-modal images for fruit detection with color, depth and range-corrected ir data. *Data in brief*, 25:104289, 2019.

He, K., Gkioxari, G., Dollar, P., and Girshick, R. Mask R-CNN. *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. doi: 10.1109/iccv.2017.322. URL http://dx.doi.org/10.1109/ICCV.2017.322.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft COCO: Common objects in context. pages 740–755. Springer, 2014.

Liu, L. and Özsu, M. T. *Encyclopedia of Database Systems*. Springer New York, NY, USA:, 2009. ISBN 978-0-387-35544-3. doi: 10.1007/978-0-387-39940-9_2002. URL https://doi.org/10.1007/978-0-387-39940-9.

Padilla, R., Netto, S. L., and da Silva, E. A. A survey on performance metrics for object-detection algorithms. In *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*, pages 237–242. IEEE, 2020.