Detecting strawberries using different Convolutional neural networks

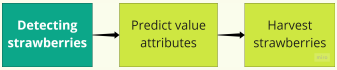
By Jeroen Bechtold, Supervisors: Thomas Abeel, Junhan Wen TU Delft

Research Question

How effective can a convolutional neural network be used to efficiently detect strawberries in the by TU Delft provided dataset?

1. Introduction

The goal of this project is to combat food waste by implementing part of the pipeline to more efficiently harvest strawberries.



Current research does tackle strawberry detection, however these papers do not compare multiple CNNs against each other. Therefore this research consists of the following sub questions:

- which CNN model does the best detection for the dataset provided by the supervisors?

2. Methodology

CNNs chosen and implemented with detectron2 [1]:

- Faster R-CNN [2] (figure 1),
- Mask R-CNN [3],
- RetinaNet [4]

Training data:

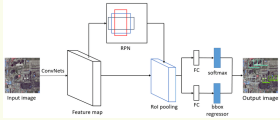
- 3 camera
  - The same one used as testing camera
- 4.370 usable images with 130.665 strawberries
- Ground truth provided (figure 2)

Test Metric:

- Precision (figure 3)
- mean Average Precision (AP)
  - From 50% to 95% confidence with steps of 5%

$$Precision = \frac{tp}{tp + fp}$$

3. How precision is calculated



1. General layout of faster r-cnn

source: https://www.researchgate.net/figure/The-framework-of-Faster-R-CNN-RPN-region-proposal-network-RoI-region-of-interest-FC_fig1_333048961

(a) The orignal image (B) The Ground truth plotted on image 2.a

2. The images provided.

3. Results & Discussion

- 100 epoch single camera
  - overfitted (figure 4)
- Doubling the data was tested.
  - On average double camera dataset 8.42% better.
  - Still overfitting

In total the following was tested:

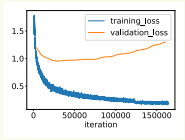
- 1 camera trainingset:
  - 10 & 100 epoch
  - learning rate from 0.002 to 0.2
- 2 camera trainingset:
  - 5, 10, 25, 50 epoch
  - learning rate from 0.002 to 0.2

Model	mAP BBox	LR	Epoch	no. cams
Faster r-cnn	51.00	0.002	5	2
Mask r-cnn	**51.63**	**0.002**	**5**	**2**
RetinaNet	49.36	0.002	10	2

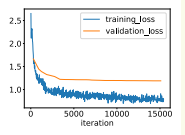
8. The best BBox mAP results for each CNN

Model	mAP Seg	LR	Epoch	No. cams
Mask r-cnn	**73.20**	**0.02**	**10**	**2**
Mask r-cnn	71.56	0.2	10	1
Mask r-cnn	71.27	0.02	5	1

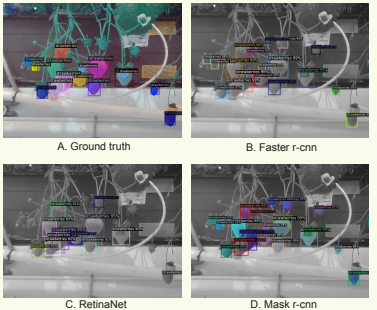
9. The top 3 segm mAP



4. Overfitting on faster r-cnn 100 epochs Loss consists of all losses summed. The base losses are calculated by smoorth l1 loss and cross-entropy loss.



5 No overfitting on the best BBox mAP model



3. Incorrect ground truth

4. Conclusion

- Best results:
  - BBox mAP: **Mask r-cnn**, lr:**0.002** epochs:**5**, no. cams: **2**
  - Segm mAP: **Mask r-cnn**, lr:**0.02** epochs:**10**, no. cams: **2**
- Training favors lower learning rates.
- Higher learning rate sometimes made the training loss diverge.
- Adding more data improves the accuracy.

References

[1] . Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2." https://github.com/facebookresearch/detectron2, 2019
[2] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, pp. 1440–1448, 2015
[3] K. He, G. Gkioxari, P. Doll´ar, and R. Girshick, "Mask r-cnn," in Proceedings of the IEEE international
conference on computer vision, pp. 2961–2969, 2017.
[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Doll´ar, "Focal loss for dense object detection," in Proceedings
of the IEEE international conference on computervision, pp. 2980–2988, 2017.