The Effects of Model Certainty, Test-Time Augmentation, and Their Trade-Offs on Leaf Segmentation and Counting

Douglas P. S. Gomes and Lihong Zheng
Charles Sturt University
Wagga Wagga, Australia
douglas.uf@gmail.com and lzheng@csu.edu.au

1. Introduction

The tasks of leaf segmentation and counting have been receiving increasing attention with some distinguishable work focusing solely on either or both of these tasks ([4, 6, 8]). However, it is important to note that localization (segmentation) and detection have an intrinsic trade-off, which is usually regulated by a detection threshold that influences how many instances are counted as positive detections and how well-localized they are.

The following experiments presents techniques adopted in training and analyses that led to a model ranking first in the LSC (when first proposed [1]) without using synthetic data generation methods while accessible computer-vision frameworks.

2. Data and methods

The analysis proposed here is based primarily on the CVPPP data set while the validation of the presented hypotheses is performed on the Leaf Segmentation Challenge server and on an independent similar data set of a different plant species (Komatsuna).

The CVPPP images have been collected from several sites from growth chamber experiments, and it is divided into four groups named A1 to A4. There are 810 images with ground-truth masks available, from which 783 are from Arabidopsis and 27 from Tobacco, showing that the later represents only a small part of that data. These images are all taken from the top. The organizers of this data set made the critical decision of creating a separate data set, A5, and uploading it to an evaluation server in CodaLAB labelled as the Leaf Segmentation Challenge (LSC).

In a similar manner to the CVPPP data set, the Komatsuna ([5]) comprises images of plants taken from the top but now on a different species, which is hoped to help attest generalization. Such a test set is composed of 271 images, and the evaluation is performed in the same metrics as used in the CVPPP testing.

2.1. Methods

Previous works proposing algorithms for segmenting leaves on the CVPPP have been using a specific method for computer vision called Mask R-CNN ([2]). The detection trade-off comes into play in the mask proposal head, which is the addition to the classic Faster R-CNN architecture. Such an effect is relevant to the approached tasks since recent methods for leaf segmentation often does not discuss the effect that different thresholds can have on the final output. Therefore, experiments with different thresholds were performed here and analysed in the Mask R-CNN framework.

To the subject of model depth and cardinality, two different depths were compared with a variation with different cardinality on the 101-layer backbone was added. This addition resulted in two models having a ResNet backbone [3], one with 50 and the other with 101 layers, and one model having a ResXNet backbone [7]. The difference between these two types of backbone is in the size of transformations, which the authors call cardinality [7], but they do have the same dimensions regarding depth and width.

A solution proposed here to extract more from the used models is to apply a technique called Test-Time Augmentation (TTA) to mitigate the effects that a different detection threshold causes in the leaf counting task while preserving or even boosting the leaf segmentation performance. In short, test-time augmentation comprises the process of performing inference on the original image and different augmented versions of it. The predictions are then averaged through a pixel-wise voting system and the final output is composed of averaged masks of each instance. The application of TTA here is composed of definite parts: (i) augmentation of the original image, (ii) inference, (iii) reversing the augmentation on the predictions, (iv) aligning instances of leaves, and (v) averaging the instances masks. The augmentation of leaves is composed of four, simple ones: horizontal and vertical flipping, and 90 degrees rotation clockwise and anticlockwise.

Threshold	SBD	abs. DiC
0.5	0.8117	1.48
0.7	0.874	1.59
0.9	0.9137	2.88

Table 1: Comparison of the effect of different detection thresholds on the CVPPP test set in the leaf segmentation metric (SBD) and counting (absolute DiC).

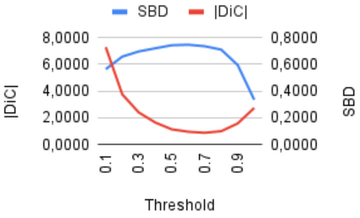

Figure 1: The effect of different thresholds on the Komatsuna test set in the SBD and abs. DiC metrics.

3. Results

For the CVPPP, Table. 1 is set to illustrate the comparison between the performance of three key thresholds in the metrics of leaf segmentation (SBD) and counting (DiC). What is most interesting about this comparison is the gains in performance that one makes from a 0.5 to 0.7 threshold.

The testing different thresholds on the Komatsuna data set showed that the performance gains might be somewhat overly expressed in the CVPPP data set. Fig. 1 illustrates the difference in the SBD and DiC in detail. The results show that although the detection threshold can have a great influence on such an external data set, the leaf segmentation metric does not monotonically increase to from 0.5 to 0.9, as seen in the CVPPP. The increase in performance with very hard thresholds points to the fact that the CVPPP might represent a dataset with narrow distribution and that the performance mainly gains comes from an overfitting artifact.

A consequent question is how models with different depth will perform in these tasks while trained by the same algorithm. The results from such an experiment were also revealing; Fig. 2 is set to illustrate some comparisons. In this case, the figure compares the 50- and the two 101-layer models (ResNet and ResXNet), all for the same threshold of 0.7. The Komatsuna data set again clarifies that such gains don't generalize well, highlighting the importance of testing on external datasets (which is not common practice in the related papers encountered). Cardinality did not allow the deeper model to overfit as it happened with the ResNet of the same depth. It points to the fact that having higher cardinality — the only difference between these backbones — may work better in problems of objects with high occlusion, allowing the model to generalize significantly better to other data sets.

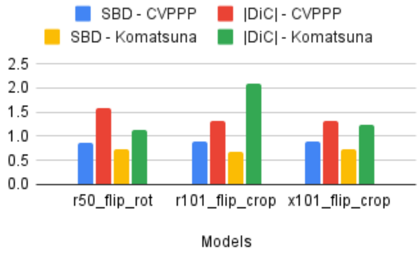

Figure 2: Comparison of performance of models of different depth and cardinality on the tests sets CVPPP and Komatsuna. Detection threshold = 0.7.

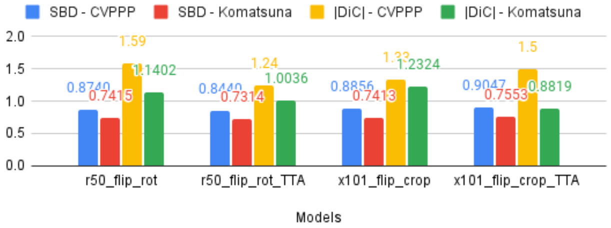

Figure 3: TTA performance comparison in the 50-layer ResNet backbone and the 101-layer ResXNet backbone on the CVPPP and Komatsuna data set.

Regarding test-time augmentation, the experiment showed to be consistent for both data sets but it showed to be more useful for ResXNet backbones as illustrated in Fig. 3. This result becomes more important when the Komatsuna is considered. By applying TTA, the model not only makes a better compromise between segmentation and counting on the CVPPP data set but also generalises much better.

The results illustrated that adjustments like increasing the detection threshold might significantly boost performance on a isolated dataset like the CVPPP, but that will not generalize to a similar external dataset. The experiments also reinforced the trade-off effect between segmentation and counting tasks, which is regulated by the detection threshold. The use of test-time augmentation was then proposed to mitigate such a trade-off effect while using metrics of segmentation and counting adopted in the field. The methodology proposed showed its effectiveness by achieving competitive results - best ranking segmentation in the LSC when first proposed - in both the Leaf Segmentation Challenge and the external dataset. It is worth noting that the past best-performing models used methods to generate synthetic data for achieving their performance while this work only used the simple and less onerous adjustments in training and testing such as test-time augmentation.

References

[1] Douglas Pinto Sampaio Gomes and Lihong Zheng. Leaf segmentation and counting with deep learning: on model certainty, test-time augmentation, trade-offs. *arXiv preprint arXiv:2012.11486*, 2020.

[2] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Dmitry Kuznichov, Alon Zvirin, Yaron Honen, and Ron Kimmel. Data augmentation for leaf segmentation and counting tasks in rosette plants. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Hideaki Uchiyama, Shunsuke Sakurai, Masashi Mishima, Daisaku Arita, Takashi Okayasu, Atsushi Shimada, and Rinichiro Taniguchi. An easy-to-setup 3d phenotyping platform for komatsuna dataset. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2038–2045, 2017.

[6] Daniel Ward and Peyman Moghadam. Scalable learning for bridging the species gap in image-based plant phenotyping. *Computer Vision and Image Understanding*, page 103009, 2020.

[7] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017.

[8] Yezi Zhu, Marc Aoun, Marcel Krijn, Joaquin Vanschoren, and High Tech Campus. Data augmentation using conditional generative adversarial networks for leaf counting in arabidopsis plants. In *BMVC*, page 324, 2018.