1. Maciej JUREWICZ[1], 2. Bartosz Świderski[1], 3. Jarosław KUREK[1]

Department of Artificial Intelligence, Institute of Information Technology, Warsaw University of Life Sciences (1),
ORCID: 1. 0000-0003-4206-3723; 2. 0000-0002-7490-8930, 3. 0000-0002-2789-4732

doi:10.15199/48.2024.05.55

Application of Mask R-CNN Algorithm for Apple Detection and Semantic Segmentation

Streszczenie. *Artykuł ten przedstawia zastosowanie algorytmu Mask R-CNN do wykrywania i semantycznej segmentacji jabłek, mając na celu zwiększenie automatyzacji w sektorze rolniczym. Pomimo rosnącego wykorzystania technik uczenia głębokiego w zadaniach detekcji obiektów, ich stosowanie w kontekstach rolniczych, szczególnie w wykrywaniu i semantycznej segmentacji owoców, pozostaje stosunkowo niezbadane. Niniejsze badanie ocenia wydajność algorytmu Mask R-CNN poprzez serię eksperymentów numerycznych, wykorzystując metryki takie mIoU, wynik F1, dokładność oraz analizę macierzy pomyłek. Nasze wyniki wykazały, że model Mask R-CNN był skuteczny w wykrywaniu i segmentacji jabłek z dużą dokładnością, osiągając mIoU wynoszące 0.551, wynik F1 równy 0.704 oraz dokładność 0.957. Jednakże zidentyfikowano również obszary potencjalnych ulepszeń, takie jak zmniejszenie fałszywie negatywnego wskaźnika modelu. To badanie dostarcza wglądów w zastosowanie algorytmów uczenia głębokiego w sektorze rolniczym, torując drogę do bardziej wydajnych i zautomatyzowanych systemów zbierania owoców.* (***Zastosowanie algorytmu Mask R-CNN do wykrywania jabłek i segmentacji semantycznej***)

Abstract. *This research presents an application of the Mask R-CNN algorithm for apple detection and semantic segmentation, aiming to enhance automation in the agricultural sector. Despite the growing use of deep learning techniques in object detection tasks, their application in agricultural contexts, specifically for fruit detection and semantic segmentation, remains relatively unexplored. This study evaluates the performance of the Mask R-CNN algorithm through a series of numerical experiments, with metrics including mean intersection over union (mIoU), F1 score, accuracy, and a confusion matrix analysis. Our results demonstrated that the Mask R-CNN model was effective in detecting and segmenting apples with a high degree of precision, achieving an mIoU of 0.551, an F1 score of 0.704, and an accuracy of 0.957. However, areas for potential improvement were also identified, such as reducing the model's false negative rate. This study provides insights into the application of deep learning algorithms in the agricultural sector, paving the way for more efficient and automated fruit harvesting systems.*

Słowa kluczowe: jabłka, detekcja obiektów, segmentacja semantyczna, MASK R-CNN
Keywords: apple, object detection, semantic segmentation, MASK R-CNN

Introduction

The application of machine learning, specifically deep learning methods in agriculture has gained significant attention in recent years. These methods provide substantial potential in automating and enhancing various processes in agricultural practices such as detection, identification, and segmentation of fruits within orchards. This work focuses on the application of the Mask R-CNN (Region-based Convolutional Neural Networks) algorithm for apple detection and semantic segmentation [9].

The convolutional neural network (CNN) has been a widely adopted technique for image analysis tasks due to its high accuracy and reliable performance. It has brought about a significant breakthrough in areas such as image recognition, object detection, and semantic segmentation. One such influential architecture is the Region Convolutional Neural Network (R-CNN) and its variations, including Fast R-CNN, Faster R-CNN, and Mask R-CNN [1]-[5]. The Mask R-CNN, introduced by He, K., Gkioxari, G., Dollár, P., and Girshick, R., et al. in their seminal 2017 paper titled "Mask R-CNN". In this paper [1], the authors present a simple, flexible, and generally effective approach for object segmentation. The methodology effectively identifies objects within an image and subsequently generates high-quality segmentation for each object and stands out for its unique feature of generating bounding boxes and segmentation masks for each instance of an object in the image. It extends Faster R-CNN by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition.

The algorithm has seen wide adoption in various domains, including agriculture where it has been used for fruit detection and semantic segmentation. The challenge in such applications lies in accurately detecting and segmenting fruits in complex natural environments, amidst foliage, varying lighting conditions, and often with fruits occluded or at different stages of maturity.

Iqbal et al. [13] applied Mask R-CNN for fruit count and diameter estimation of coconut and oil palm trees. They found that Mask R-CNN outperformed traditional machine learning methods, providing a robust solution to the problem of fruit detection in complex environments.

Similarly, Bargoti and Underwood [15] used the Faster R-CNN architecture to detect and count apples in orchard environments. Despite the complexity of the task, their method achieved high precision and recall rates, thereby demonstrating the feasibility of deep learning methods for fruit detection.

Choi exploited the depth information in 3D images to pinpoint fruit areas on a map, leveraging a Convolutional Neural Network (CNN) for accurate identification. Through comparison, it was concluded that Near InfraRed (NIR) images yielded the most promising results with a true positive rate of 96% for detecting and classifying circular objects (Choi, 2017 [16]).

Chen, on the other hand, developed a blob detector applying a fully-connected CNN to pull out potential regions in images, segment them, and then use a subsequent CNN algorithm to estimate the quantity of fruits (Chen et al., 2017 [17]).

Meanwhile, Tian modified the yolo-v3 network to boost feature propagation and improve feature reusability, enabling detection of apples at varying stages of growth. Despite the positive results, the algorithm was still affected by obscured fruits (Tian et al., 2019 [14]).

Yu, Zhang et at. [8] applies Mask R-CNN to strawberry detection for harvesting robots in non-structured environments. The Resnet50 backbone, combined with Feature Pyramid Network (FPN), enhanced the model's feature extraction ability, proving significantly effective for recognizing overlapping and hidden fruits under varying illumination.

In [6], an improved Mask Scoring R-CNN model (MS-ADS) is proposed for apple detection and instance segmentation in natural environments. By incorporating ResNeSt and Dual Attention Network, this approach successfully manages challenges introduced by various apple conditions, achieving high accuracy and real-time performance.

The 'Deep Orange' study [10] presents a unique approach of using both RGB and HSV images as inputs for Mask R-CNN in order to detect and segment oranges under natural lighting conditions. The findings emphasize the role of HSV data in boosting precision.

Forest inventory management is another practical application as seen in [7]. Mask R-CNN is used to detect tree-crowns and estimate their height simultaneously. This study shows the efficiency of the algorithm when using the Normalized Difference Vegetation Index (NDVI) and Canopy Height Model (CHM) as input images.

In [11], the use of Mask R-CNN for road crack detection is explored and compared with Faster R-CNN. Although both methods demonstrate good performance, the joint training strategy results in a degradation in the effectiveness of the bounding box detected by Mask R-CNN.

Finally, [12] proposes an improved Mask R-CNN model for detecting disease spots on various fruits. Through enhancement of the FPN structure of Mask R-CNN, the model exhibits high detection accuracy and speed, outperforming other methods like Fast R-CNN and SSD algorithms.

Sa utilized the Faster RCNN on a multi-vision sensor for the detection of peppers, rock-melons, and apples via transfer learning (Sa et al., 2016 [18]). Similarly, Bargoti implemented the Faster-RCNN model for the detection of apples and mangos in orchards, yielding high detection accuracy but struggling to detect clustered fruits (Bargoti and Underwood, 2017 [15]).

However, the application of Mask R-CNN for apple detection and semantic segmentation is a relatively under-explored area of research. Therefore, this work aims to contribute to this field by providing a detailed investigation of the Mask R-CNN algorithm's application for apple detection and segmentation.

Several scientific articles have provided insights into the practical applications of the Mask R-CNN algorithm, shedding light on its versatility and effectiveness. Notable studies in the field of computer vision, autonomous driving, medical imaging, and agricultural technology demonstrate the algorithm's potential.

The main goal of this article is to explore the application of Mask R-CNN for apple detection and semantic segmentation. In the broader scheme of agricultural technology, this application can potentially streamline the process of apple harvesting by automating the identification and localization of apples in orchards. While some studies have hinted at the potential of deep learning in fruit detection, there is a clear gap in research focusing on the specific application of Mask R-CNN for apple detection and segmentation. By bridging this gap, this article aims to contribute to the optimization of agricultural practices, potentially leading to increased yield and profitability.

Data Set

The critical cornerstone of any machine learning algorithm's performance is the dataset used for both training and evaluation. In this study, we applied a unique dataset to ensure the robustness and applicability of the Mask R-CNN algorithm for apple detection and semantic segmentation.

The dataset comprised a total of 115 images. The selection was based on diverse factors to build an encompassing and comprehensive dataset. The images used spanned various apple species, which are significantly different in shape, size, and color. Both red and green apples were included, promoting diversity and improving the model's ability to generalize.

Out of the 115 images, 99 were used for the training phase. This selection of images was carefully curated to contain different apple species under varying lighting conditions and from multiple viewpoints. This large number of training images ensured the algorithm was exposed to ample data, leading to better learning and ultimately improved performance.

The remaining 16 images were reserved for the testing phase. It is vital to separate a subset of data for testing to evaluate the model's performance on unseen data, which directly mirrors real-world performance. The test images were chosen randomly from the initial dataset to ensure impartiality and a representative sample.

Mask R-CNN Algorithm

Mask R-CNN (Mask Region with Convolutional Neural Networks) is a powerful model that offers both object detection and semantic segmentation capabilities. Introduced by He et al. [1], it is an extension of Faster R-CNN, a model known for its efficiency in object detection.

A. Architecture

The Mask R-CNN model consists of two primary stages. The first, known as the backbone, is a deep convolutional network that serves as a feature extractor. The second, known as the head, is used for predicting class labels, bounding box adjustments, and pixel-wise masks for each object [1].

Backbone:

The backbone of Mask R-CNN is a convolutional neural network (CNN) that acts as a feature extractor. Typically, a pre-trained ResNet is used for this purpose. The ResNet backbone is further extended by a Feature Pyramid Network (FPN) which creates a rich, multi-scale feature representation beneficial for detecting objects at various scales [20].

Region Proposal Network (RPN):

RPN is used to generate a set of proposed regions where an object might be located (objectness). RPN shares its convolutional layers with the object detection part of the model, leading to efficiency and speed improvements. The RPN generates multiple region proposals, and with the help of RoI (Region of Interest) pooling, it aligns regions to a fixed size to enable the use of fully connected layers.

RoIAlign**:**

In Mask R-CNN, a method called RoIAlign is applied which preserves the exact spatial locations of the given RoIs, avoiding the harsh quantization in RoI Pooling, essentially aligning the regions perfectly and providing more accurate region proposals.

Heads**:**

The head of Mask R-CNN consists of two parts: the detection head and the mask head. The detection head is responsible for classifying the object and providing refined bounding box coordinates. The mask head generates a binary mask for each RoI at pixel level. It is important to note that the mask generation is decoupled from class prediction, allowing masks to be generated without competition among classes.

B. Performance

Mask R-CNN demonstrates strong performance on the COCO benchmark, outperforming previous state-of-the-art models in object instance segmentation. In terms of efficiency, Mask R-CNN's system is capable of running at 5 fps (frames per second) on a GPU [1].

C. Improvements over Previous Models

One of the key improvements of Mask R-CNN over its predecessors is its ability to handle occlusions, a frequent scenario in image analysis where parts of an object are covered by another. Furthermore, Mask R-CNN can generate more precise segmentation results due to its fine-grained segmentation capability, allowing better detection of objects' boundaries.

Numerical Experiments and Results

In this section, we delve into the statistical analysis and interpretation of the results of our application of the Mask R-CNN algorithm for apple detection and semantic segmentation.

Figure 1 provides a visual representation of the input to the Mask R-CNN network—a photograph captured within an apple orchard, showcasing the raw data as perceived by the algorithm. The vibrant red tones of the apples stand out against the green foliage, setting the stage for the network's segmentation task.

Figure 2 illustrates the outcome post-processing by the Mask R-CNN algorithm, highlighting the successful semantic segmentation of apples. Each apple is encircled, denoting the algorithm's detection accuracy. The varying shades of the bounding boxes represent the algorithm's confidence levels, showcasing the precise differentiation between the fruit and the surrounding foliage.

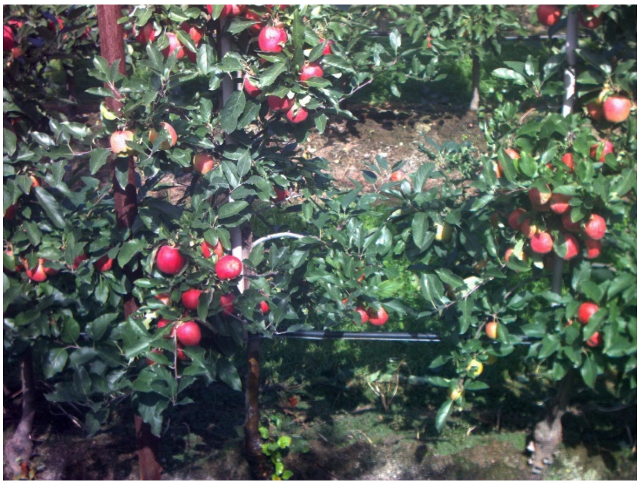

Figure.1. Sample input to the Mask R-CNN network in the form of an apple orchard photo

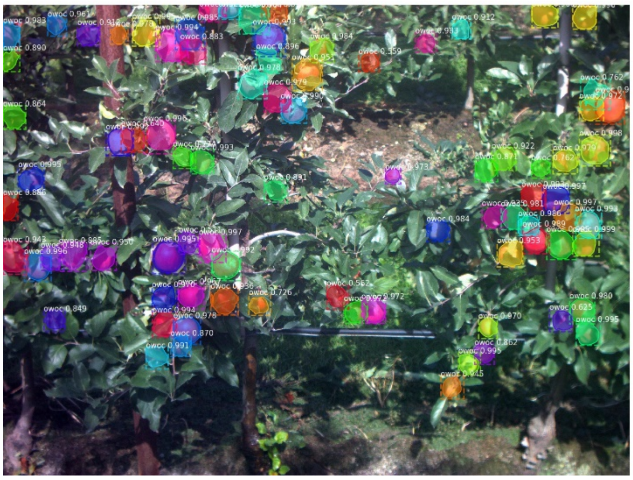

Figure.2. Result of semantic segmentation (Mask R-CNN algorithm) for the above input photo

The performance of our model was evaluated using several key metrics. The mean intersection over union (mIoU), F1 score, and accuracy were 0.551, 0.704, and 0.957 respectively. The mIoU, which measures the prediction accuracy of our model by comparing the area of overlap with the total area, shows that more than half of the predictions were correct. The F1 score, a measure of the model's balance between precision and recall, and the accuracy both exceed 70%, indicating a high degree of correctness in the model's predictions.

The confusion matrix, which summarizes the performance of the classification model by presenting the true negatives (TN), false positives (FP), false negatives (FN), and true positives (TP), was analyzed. These metrics provide a comprehensive view of the performance of the model and can further help identify specific areas of improvement.

The results of these metrics are tabulated as shown below:

Table 2. Performance metrics of the Mask R-CNN model

Metric	Score
mIoU	0.551
F1	0.704
Accuracy	0.957

Table 3. Confusion Matrix of the Mask R-CNN model

	Predicted: No	Predicted: Yes
Actual: No	0.899 (TN)	0.016 (FP)
Actual: Yes	0.027 (FN)	0.058 (TP)

As reflected by the TN value, the model correctly predicted the absence of apples 89.9% of the time. The FP value suggests that the model incorrectly predicted the presence of apples 1.6% of the time. The model incorrectly predicted the absence of apples 2.7% of the time, as indicated by the FN value. Finally, the model correctly predicted the presence of apples 5.8% of the time, as reflected by the TP value.

While the model demonstrates a high degree of accuracy in its predictions, there are still improvements that can be made, especially in reducing the number of false negatives to improve recall.

Conclusions

In this study, we successfully applied the Mask R-CNN algorithm for apple detection and semantic segmentation. The analysis of our numerical experiments provided important insights into the strengths and weaknesses of the implemented model.

The performance metrics, as revealed by the mIoU, F1 score, and accuracy, confirmed the model's efficacy in detecting and segmenting apples with high precision. These metrics were all well above 50%, with the F1 score and accuracy both exceeding 70%, indicating a significant level of success in the model's predictions.

The confusion matrix further confirmed the model's ability to accurately predict both the presence and absence of apples. The model showed particularly strong performance in predicting the absence of apples, with a true negative rate of 89.9%.

However, while the model's performance was generally high, there were some areas identified for potential improvement. Specifically, reducing the model's false negative rate could improve its recall performance, ensuring that fewer actual instances of apples are overlooked.

The findings from this research contribute to the growing body of evidence supporting the use of advanced deep learning techniques, such as Mask R-CNN, for object detection and semantic segmentation tasks in agricultural contexts. The application of these techniques could significantly advance efforts towards automated fruit detection and harvesting, ultimately contributing to increased efficiency and productivity in the agricultural sector.

Future work could aim to further refine the model and expand its capabilities, such as improving the detection of apples at different stages of ripeness or under varying light conditions. Additionally, research into the incorporation of other contextual features may improve the model's overall performance and applicability in real-world scenarios.

REFERENCES

[1] He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. (arXiv,2017)

[2] Suh, S., Park, Y., Ko, K., Yang, S., Ahn, J., Shin, J. & Kim, S. Weighted Mask R-CNN for improving adjacent boundary segmentation. J. Sens..2021 pp. 1-8 (2021,1)

[3] He, K., Gkioxari, G., Dollar, P. & Girshick, R. Mask R-CNN. IEEE Trans. Pattern Anal. Mach. Intell.. 42, 386-397 (2020,2)

[4] Ren, S., He, K., Girshick, R. & Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. (2016)

[5] Girshick, R. Fast R-CNN. (2015)

[6] Wang, D. & He, D. Apple detection and instance segmentation in natural environments using an improved Mask Scoring R-CNN Model. Front. Plant Sci.. 13 pp. 1016470 (2022,12)

[7] Hao, Z., Lin, L., Post, C., Mikhailova, E., Li, M., Chen, Y., Yu, K. & Liu, J. Automated tree-crown and height detection in a young forest plantation using mask region-based convolutional neural network (Mask R-CNN). ISPRS J. Photogramm. Remote Sens.. 178 pp. 112-123 (2021,8)

[8] Yu, Y., Zhang, K., Yang, L. & Zhang, D. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN. Comput. Electron. Agric.. 163, 104846 (2019,8)

[9] Chu, P., Li, Z., Lammers, K., Lu, R. & Liu, X. Deep learning-based apple detection using a suppression mask R-CNN. Pattern Recognit. Lett.. 147 pp. 206-211 (2021,7)

[10] Ganesh, P., Volle, K., Burks, T. & Mehta, S. Deep orange: Mask R-CNN based orange detection and segmentation. IFAC-PapersOnLine. 52, 70-75 (2019)

[11] Xu, X., Zhao, M., Shi, P., Ren, R., He, X., Wei, X. & Yang, H. Crack detection and comparison study based on Faster R-CNN and Mask R-CNN. Sensors (Basel). 22, 1215 (2022,2)

[12] Wang, H., Mou, Q., Yue, Y. & Zhao, H. Research on detection technology of various fruit disease spots based on mask R-CNN. 2020 IEEE International Conference On Mechatronics And Automation (ICMA). (2020,10)

[13] Iqbal, M., Ali, H., Tran, S. & Iqbal, T. Coconut trees detection and segmentation in aerial imagery using mask region-based convolution neural network. IET Comput. Vis.. 15, 428-439 (2021,9)

[14] Tian, Y., Yang, G., Wang, Z., Wang, H., Li, E. & Liang, Z. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. Comput. Electron. Agric.. 157 pp. 417-426 (2019,2)

[15] Bargoti, S. & Underwood, J. Deep Fruit Detection in Orchards. (2017)

[16] Choi, D., Lee, W., Schueller, J., Ehsani, R., Roka, F. & Diamond performance comparison of RGB, NIR, and depth images in immature citrus detection using deep learning algorithms for yield prediction, 2017 Spokane, Washington July 16 - July 19, 2017. (2017)

[17] Chen, S., Shivakumar, S., Dcunha, S., Das, J., Okon, E., Qu, C., Taylor, C. & Kumar, V. Counting Apples and Oranges With Deep Learning:A Data-Driven Approach. IEEE Robotics And Automation Letters. 2, 781-788 (2017)

[18] Sa, I., Ge, Z., Dayoub, F., Upcroft, B., Perez, T. & McCool, C. DeepFruits: A Fruit Detection System Using Deep Neural Networks. Sensors. 16 (2016), https://www.mdpi.com/1424-8220/16/8/1222

[19] He, K., Zhang, X., Ren, S. & Sun, J. Identity Mappings in Deep Residual Networks. Computer Vision – ECCV 2016. pp. 630-645 (2016)

[20] Lin, T., Dollar, P., Girshick, R., He, K., Hariharan, B. & Belongie, S. Feature Pyramid Networks for Object Detection. (2017)