Using Mask-RCNN to Identify Defective Parts of Fruits and Vegetables

Sai Raghunandan Suddapalli and Perugu Shyam(✉)

National Institute of Technology, Warangal, India
shyamperugu@nitw.ac.in
https://nitw.ac.in/

Abstract. Fruits and vegetables are a major source of food for humans after cereals. Since the evolution of civilizations, they have been gathered, cultivated and modified according to our needs. During the process of modification and harvesting, there might be diseased variants that are unfit for consumption. Manual removal/segregation of the diseased fruits and vegetables is a time consuming process on a large scale, which could be automated in the near future with the help of artificial intelligence. This could be done by training a machine, using machine learning algorithms, to recognize which fruits and vegetables are fit for consumption and which ones are not, with the help of an annotated dataset. The goal of this study is to introduce a dataset that contains 11 classes of fruits and vegetables that are annotated for instance segmentation tasks and the effectiveness of the dataset in simplifying quality testing and analysis. This paper begins by explaining the usage of Mask-RCNN [5] algorithm, and then explains the properties of the dataset and further discusses the areas of application where the dataset can be used.

Keywords: Fruits · Vegetables · Instance segmentation · RCNN · Mask-RCNN

1 Introduction

At various stages of cultivation, fruits and vegetables are prone to a variety of diseases such as rust, blight, scab and wilt. These diseases can occur due to a wide variety of reasons, but can be broadly categorized into three types: non-parasitic, parasitic, and viral diseases [11]. During the time of separation/removal, it becomes cumbersome to segregate the fruits and vegetables manually, especially at a large scale. For this purpose, we can employ automation techniques, particularly machine learning algorithms, to automate this process, thereby reducing the human factor. This study proposes an annotated dataset for instance segmentation of popular fruits and vegetables. This dataset is named FV-11, as it contains 11 classes of annotated fruits and vegetables. Currently, there are 355 images with annotations, which are encouraged to be extended further by the users. Classification tasks usually require high quality images that are often fed

J.-H. Kim et al. (Eds.): IHCI 2021, LNCS 13184, pp. 637–646, 2022.
https://doi.org/10.1007/978-3-030-98404-5_58

to the networks as one instance of a class per image. In instance segmentation tasks, it is required to have the position of the class in the image as well, so that it can be identified by the neural network. There exists a dataset by the name Fruits-360 [9], which contains 131 fruits and vegetables that are categorized for classification purposes. Now, we attempt to propose an annotated dataset that can be used for instance segmentation as well, where the localization, and the classification task is done at the same time. This dataset could prove to have important applications in the field of agriculture, and quality testing for the food industry, where identification of defects could be made easier and faster. In this study, we have used Mask-RCNN architecture for training our model. It is a region convolutional neural network that localizes the object and also generates a bitmap for the mask that is overlaid on the object. Considering the rapid growth in the food and nutrition industries where the need for rapid categorization of raw material is of high significance, our dataset can provide a way for automation of this process. This not only saves time, but also aids in accurate categorization of usable fruits and vegetables, thereby reducing human effort for such processes. Instance segmentation [4] algorithms provide a way of identifying objects that are scattered throughout the canvas of the given/captured image at a given time. This quality of instance segmentation is very useful for quality control methods in such factories that utilize fruits and vegetables as their raw material.

2 Data Preparation

To prepare the dataset, we have chosen 200 high quality images from Unsplash [12], which is a website with images with Creative commons license. We have gathered 155 images of fruits and vegetables from a local market, where there were multiple vendors who had stalls of fruits and vegetables. The images gathered from the local market were captured using a mobile device. Since most images were in varying dimensions, we have decided to upscale the images onto a 1024×1024 canvas before sending the input to the model for training purposes. So, the input layer of the Mask-RCNN model we initialized was of the shape (1024, 1024, 3). The third dimension handles the RGB color format. All images were then annotated with an image labelling and annotating tool called Sense sixgill [1]. Here the labelling was done in a way such that we can generate masks for each instance of every class. For 11 classes on an average, we could get 60 instances that the model could be trained on. The classes are listed as follows: potato, tomato, carrot, brinjal, rotten_apple, rotten_banana, rotten_orange, apple, banana, orange, guava.

Sense sixgill is an industry standard annotation tool that can be used to generate datasets for various object detection problems. The annotations were later exported in Sense's annotation format. The data gathered should be split into training, test, and validation sets. This is done so as to prevent overfitting of the model. Overfitting is a scenario in machine learning problems where the model predicts inferences with very high accuracy on the training set data, but fails to accurately draw inferences on the test data, which the model hasn't been

trained on. This leads to an erratic model, which performs well only on the data it was trained on, which is not feasible for practical reasons. So, the dataset was split into a ratio of 8:1:1 of training: validation: test data respectively.

3 Model Training

All experiments were conducted in our local workstation using tensorflow 2.0 on a Lenovo®computer; Intel®Core^{TM}i5-5200U CPU @ 2.20 GHz, 2201 MHz, 8 Core(s), 4 Logical Processor(s), with 16 Gb RAM and 1.95 Gb Cache and 8 GB NVIDIA®GeForce 840M graphics card. When we discuss about artificial intelligence and its usage in computer vision applications, one of the most common networks we come across are artificial neural networks or ANNs [9]. To put in simple terms, ANN is an imitation of the human brain, where we structure the network in the following format:

1. Input layer
2. Hidden layer, where the computation occurs. It may consist of multiple hidden layers, depending on the complexity
3. Output layer, which consists of values ranging between 0 and 1(greater than or equal to 0 and less than or equal to 1).

Computation in ANNs occurs by the virtue of the inter-connection strengths, which are known as weights. These weights are usually represented as "W", which can take positive or negative values. Negative values cause inhibition of the input signals, and positive values cause addition. Output neurons have an activation function, which yields a result in the form of 0s or 1s. For multiple instance classification problems, ANNs use an activation function called softmax, which is defined as

$$\sigma(z) = \frac{e^{zi}}{\sum_{j=1}^{K} e^{zj}} \tag{1}$$

Where σ is the softmax function, K is the number of clusters or classes, z is the input vector and e^zi is the standard exponential function for input vector and e^{zj} is the standard exponential function for output vector. For computer vision applications, an extension of the ANNs, called Convolutional Neural Networks, or CNNs [2] have proved to be of great importance, due to the ability of the network to extract the features of a given object. It takes this name from mathematical linear operation between matrixes called convolution. CNNs have proved to be of excellent use in natural language processing tasks, and image classification tasks. Mask-RCNN model runs on a backbone of a much complex variant of CNNs, The basic model consists of a ResNet [6] backbone, which has extended output layers that perform regression for the object's position, and classification to detect the instances of the classes the model was trained on. To create an instance segmentation model, we chose Mask-RCNN for our model's architecture. There is a choice between choosing a resnet-101 backbone and a resnet-50 backbone, which depends on the user's choice. The backbone of Mask-RCNN is usually a Feature Pyramid Network, which is usually a ResNet

backbone or a VGG backbone. Formally, a Resnet building block is defined as follows

$$y = F(x, W_i) + x \tag{2}$$

As it can be observed, x and y correspond to the input and output vectors of the layers and Wi is the vector of weights. It is observed that the corresponding input vector is further connected to the final output layer in a 'shortcut'. This layer performs identity mapping on the final output layer, leading to a feed forward like behavior of the network. This leads to the layer having a less number of connected layers, and produces a competitive accuracy with much complex networks like VGG-16 [10] and Alex net [8]. The image below summarizes the working of a Mask RCNN model's working and structure [14] in a manner that is highly approachable for beginners and experts alike where we see 2 types of machine learning algorithms working. First, the classification is done to identify the type of fruit/vegetable and linear regression is performed to identify the co-ordinates of the classified objects. Figure 1 summarizes the working of a Mask RCNN model's working and structure in a manner that is highly approachable for beginners and experts alike where we see 2 types of machine learning algorithms working. First, the classification is done to identify the type of fruit/vegetable and linear regression is performed to identify the co-ordinates of the classified objects. RPN here stands for a region proposal network, which upon receiving the input image, suggests areas of interest where the objects are likely to be present. In stage 2, the 3 boxes "class", "bbox" and "mask" denote the final

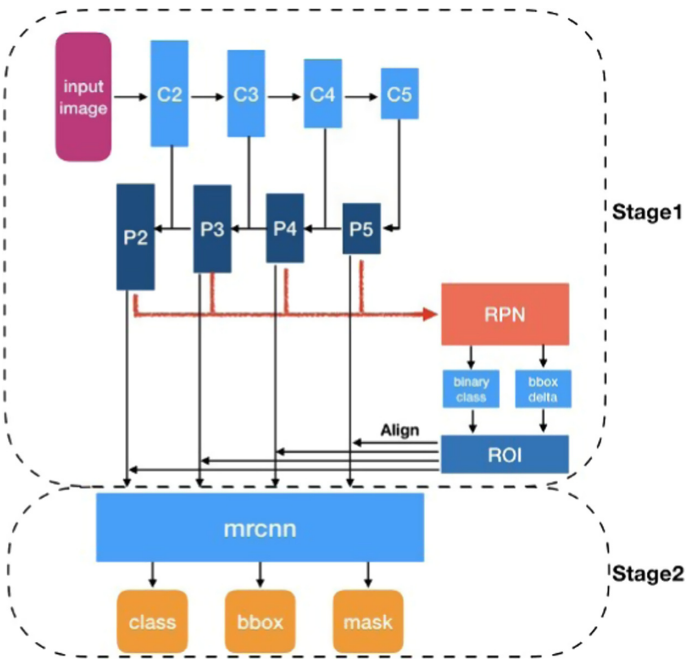

Fig. 1. Block diagram representing the network of Mask-RCNN

outputs that are presented by the model: the predicted class, the bounding box around the object of interest, and the bitmap mask generated over the object of interest. For real-time detection, other architectures like YOLACT [3] also can be used, which we will discuss in the conclusions section.

4 Results

Post training of the model, the overall accuracy was obtained to be 89% on the test dataset and 91.75% on the training dataset. Out of a 100 images for each instance that were sent for prediction, there were 10 wrongly predicted images. This gives us a practical accuracy of 90%, which can be improved by training the model further on various instances of the same classes, using more examples. This can be further improved by increasing the training time, and fine tuning the hyper parameters of the model, such as changing the learning rate, steps per epoch, no. of epochs, batch size and momentum of the optimizer. We have taken normal as well as diseased apples and were analyzed using Mask-RCNN network. The model generated a bitmap over each of the identified instances of apples. Thus localization is generated and can be seen in Fig. 2. In Fig. 4, it is shown that the model would detect guavas even when they're cut in half. We made sure the dataset contains cut fruits and vegetables as well, since some defects cannot be identified on the outer appearance. In Fig. 7, we can see the model identifies multiple instances of potatoes and carrot in the same picture. This is a key feature that can be observed in instance segmentation models, where multiple instances of different classes can be identified accurately. In Fig. 9, we see that the rotten bananas are identified here. These bananas are treated as an entirely different class, which separates them from healthy bananas. Further complexity can be added regarding the localization of the rotten part of the fruit, which is currently out of the scope of the project.

5 Discussion

Current specifications of the dataset meet a small yet significant needs of any life-scale problem. Post expansion of the dataset, we expect a significant rise in

Fig. 2. Instance segmentation of apples by the model

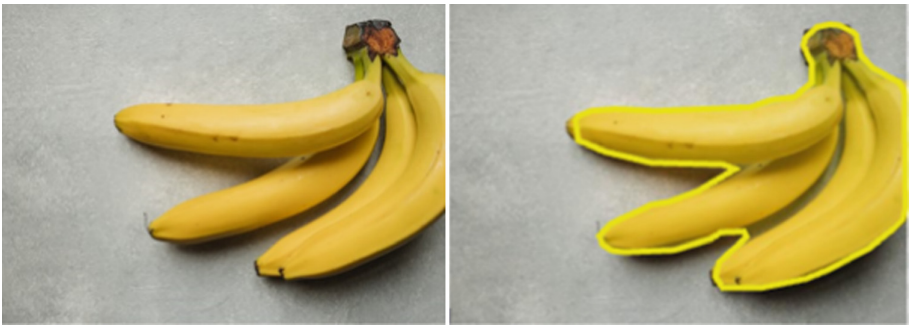

Fig. 3. Unlike Fig. 2, here the hand of bananas are recognized as a single unit, instead of being identified individually

Fig. 4. Instance segmentation of guavas when cut in half

Fig. 5. Here, oranges are detected by the model, both full oranges and cut oranges

the accuracy, and the usability of the dataset. As of now, the dataset contains 11 classes of fruits and vegetables that were used to train our current model. We didn't encounter any cases of overfitting in our model training phase, which indicates sufficient splitting of data. This data can be used to train automated

Fig. 6. Picture showing identification of Aubergines. Aubergines or Eggplants or Brinjals as they are called, are also termed as the king of vegetables in the Indian subcontinent due to its versatility in preparation of recipes

Fig. 7. Instance segmentation of multiple vegetables in a single image

systems to segregate fruits and vegetables for further processing in food and beverage industry, where fruits and vegetables are processed from their raw form into desired products such as juices, ready-made salads, etc. (Figs. 3, 5, 6, 8, 10 and 11).

We are going to train various object detection models, apart from Mask-RCNN, such as MobileNet [7], and Detectron2 [13], to obtain benchmarks of the dataset.

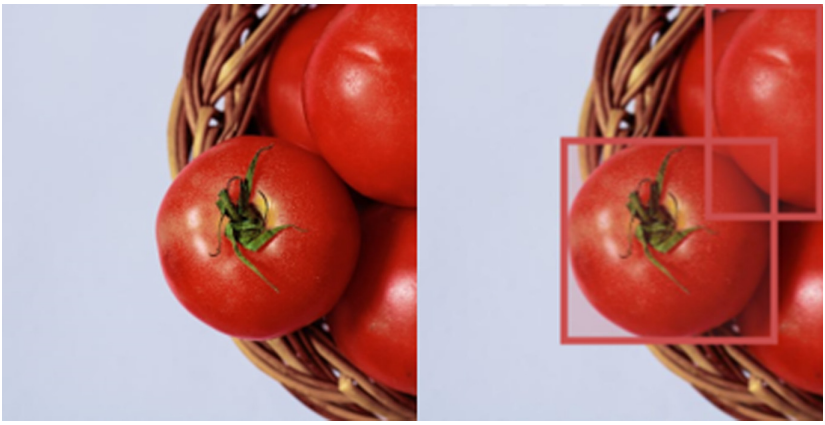

Fig. 8. Tomatoes being identified by the model

Fig. 9. Instance segmentation performed over diseased/rotten bananas

Fig. 10. Illustration showing the identification of a rotten apple by the model

Fig. 11. Illustration showing the identification of a rotten orange by the model

6 Conclusions and Further Work

We have proposed a novel dataset for instance segmentation of diseased/rotten fruits and vegetables, and we have utilised Mask-RCNN algorithm to perform instance segmentation on our dataset.

We aim to design a standalone image sensor that can capture images and can use the model trained for instance segmentation. The entirety of this device is expected to be mobile and strong enough, so that we can carry it to places where detection of diseased fruits is key to the quality checks and analyses.

We also hope to train the model using a single shot detector model such as YOLACT which can be used in real-time video stream to identify the defects in fruits and vegetables.

We also aim to extend the dataset to identify specific diseases of fruits and vegetables, wherein timely identification of the diseases would help in implementing remedies for the crop, thereby helping boost the yield. The diseased fruits are classified purely on the basis of the presence/absence of degradation in the fruit or vegetable. This is one of our future prospects of this project where we extend the dataset to specific diseases, and not separation of healthy and diseased fruits in general. This could directly help the horticulture industry in identifying the disease on time, and implementing the remedial measures to boost the yield of the crop.

References

1. https://sense.sixgill.com/
2. Albawi, S., Mohammed, T.A., Al-Zawi, S.: Understanding of a convolutional neural network. In: 2017 International Conference on Engineering and Technology (ICET), pp. 1–6 (2017). https://doi.org/10.1109/ICEngTechnol.2017.8308186
3. Bolya, D., Zhou, C., Xiao, F., Lee, Y.J.: YOLACT: real-time instance segmentation. CoRR abs/1904.02689 (2019). http://arxiv.org/abs/1904.02689

4. Hafiz, A.M., Bhat, G.M.: A survey on instance segmentation: state of the art. CoRR abs/2007.00047 (2020). https://arxiv.org/abs/2007.00047
5. He, K., Gkioxari, G., Dollár, P., Girshick, R.B.: Mask R-CNN. CoRR abs/1703.06870 (2017). http://arxiv.org/abs/1703.06870
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015). http://arxiv.org/abs/1512.03385
7. Howard, A.G., et al.: MobileNets: efficient convolutional neural networks for mobile vision applications. CoRR abs/1704.04861 (2017). http://arxiv.org/abs/1704.04861
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, NIPS 2012, vol. 1, pp. 1097–1105. Curran Associates Inc., Red Hook (2012)
9. Murean, H., Oltean, M.: Fruit recognition from images using deep learning. Acta Univ. Sapientiae Inform. **10**, 26–42 (2018). https://doi.org/10.2478/ausi-2018-0002
10. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556 (2015)
11. Thind, S.K.: Principles of disease management in fruit crops, May 2017. https://medcraveonline.com/ICPJL/principles-of-disease-management-in-fruit-crops.html
12. Unsplash: Download free pictures & images [HD]. https://unsplash.com/images
13. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2 (2019)
14. Zhang, X.: Simple understanding of mask RCNN, April 2018. https://alittlepain833.medium.com/simple-understanding-of-mask-rcnn-134b5b330e95