Date out : July 20, 2020 Due on: July 27, 2020

ENIN 880CA – Spring/Summer 2020

Instance Segmentation Using Mask R-CNN

The subject of this assignment is using Mask R-CNN for following two applications:

- Segmentation and masking of skin lesions
- Segmentation and masking of round pills

1. Mask R-CNN

1.1. From Object Detection to Semantic and Instance Segmentation [1]

Before reviewing Mask R-CNN, we would review object detection and segmentation.

Object detection

Objects in an image are detected and the output are corresponding bounding box coordinates and class labels for each object (Figure 1, top-right).

Traditional segmentation

Image is partitioned into objects and background. The outputs are pixel-wise masks for objects. Semantic and instance are two variations of segmentation and defined as follows:

Semantic segmentation

The image is partitioned into meaningful parts. While each class-object has a pixel-wise masks, multiple objects of the same class are treated as a single entity (Figure 1, bottom-left).

Instance segmentation

The image is partitioned into meaningful parts. Not only multiple objects of the same class, but also multiple objects of the same class are treated as distinct individual objects (or instances) and have their individual masks (Figure 1, bottom-right).

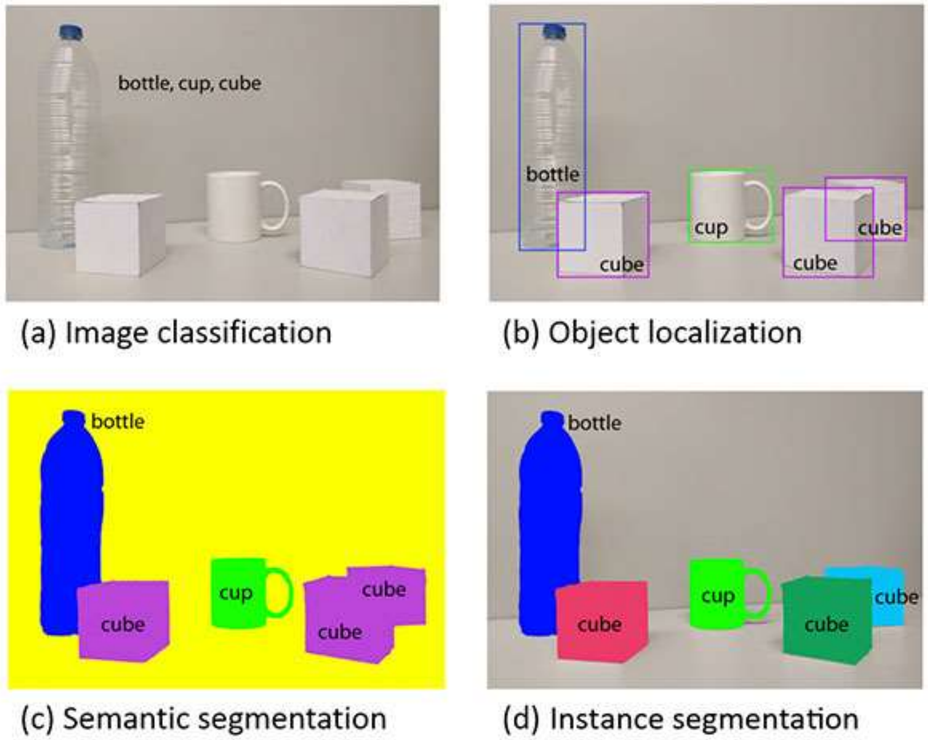


Figure 1: Object Detection vs. Instance Segmentation [1]

1.2. Mask R-CNN

Mask R-CNN is a state-of-the-art model for instance segmentation, developed on top of Faster R-CNN. To understand Mask R-CNN, let's first review the architecture of Faster R-CNN:

Faster R-CNN [2]

Faster R-CNN is a region-based convolutional neural networks that returns bounding boxes for each object and its class label with a confidence score. Faster R-CNN works in two stages (figure 2):

Stage1:

The first stage consists of two networks, backbone (ResNet, VGG, Inception, etc..) and Region Proposal Network (RPN). These networks run once per image to give a set of region proposals. Region proposals are regions in the feature map which potentially contain the object.

Stage2:

In the second stage, the network predicts bounding boxes and object class for each of the proposed region obtained in stage1. Each proposed region can be of different size whereas fully connected layers in the networks always require fixed size vector to make predictions. Size of these proposed regions is fixed by using either RoI pool (which is very similar to MaxPooling) or RoIAlign method. The features used by both stages can be shared for faster inference.

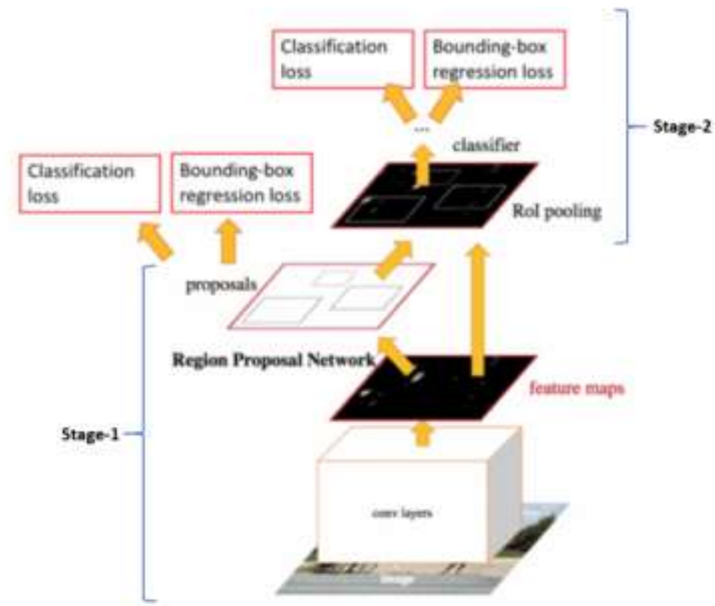


Figure 2: Faster R-CNN

Mask R-CNN [1, 3]

Mask R-CNN is conceptually simple: Faster R-CNN has two outputs for each candidate object, a class label and a bounding-box offset; to this we add a third branch that outputs the object mask — which is a binary mask that indicates the pixels where the object is in the bounding box. But the additional mask output is distinct from the class and box outputs, requiring extraction of much finer spatial layout of an object. To do this Mask R-CNN uses the Fully Convolution Network (FCN) (figure 3).

So, in short, we can say that Mask R-CNN combines the two networks — Faster R-CNN and FCN in one mega architecture (figure 4). The loss function for the model is the total loss in doing classification, generating bounding box and generating the mask. Mask R-CNN results in following outputs:

- Label prediction
- Bounding box prediction
- Mask prediction

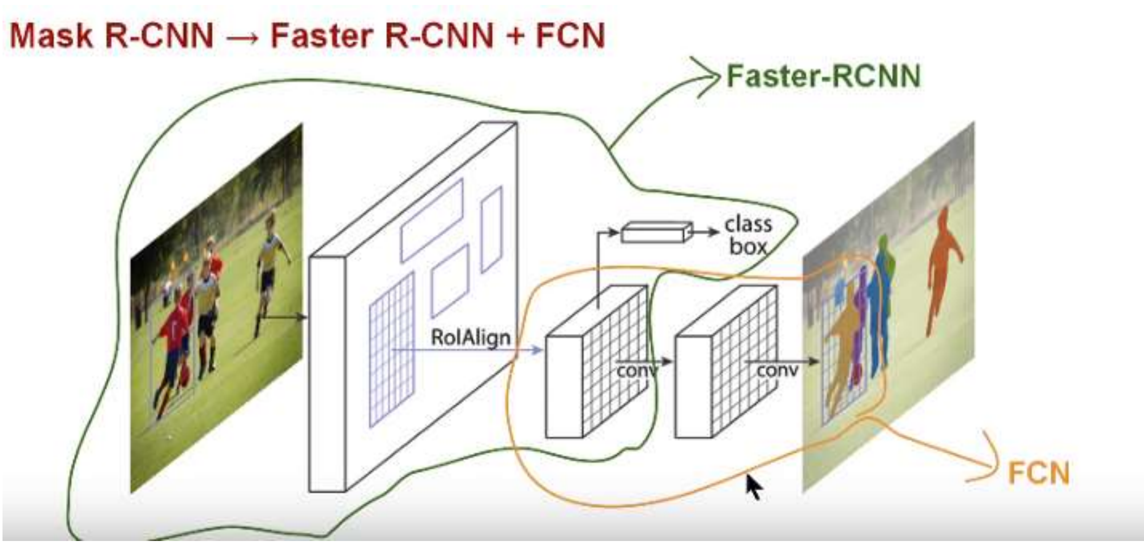


Figure 3: Mask R-CNN=Faster R-CNN + FCN [2]

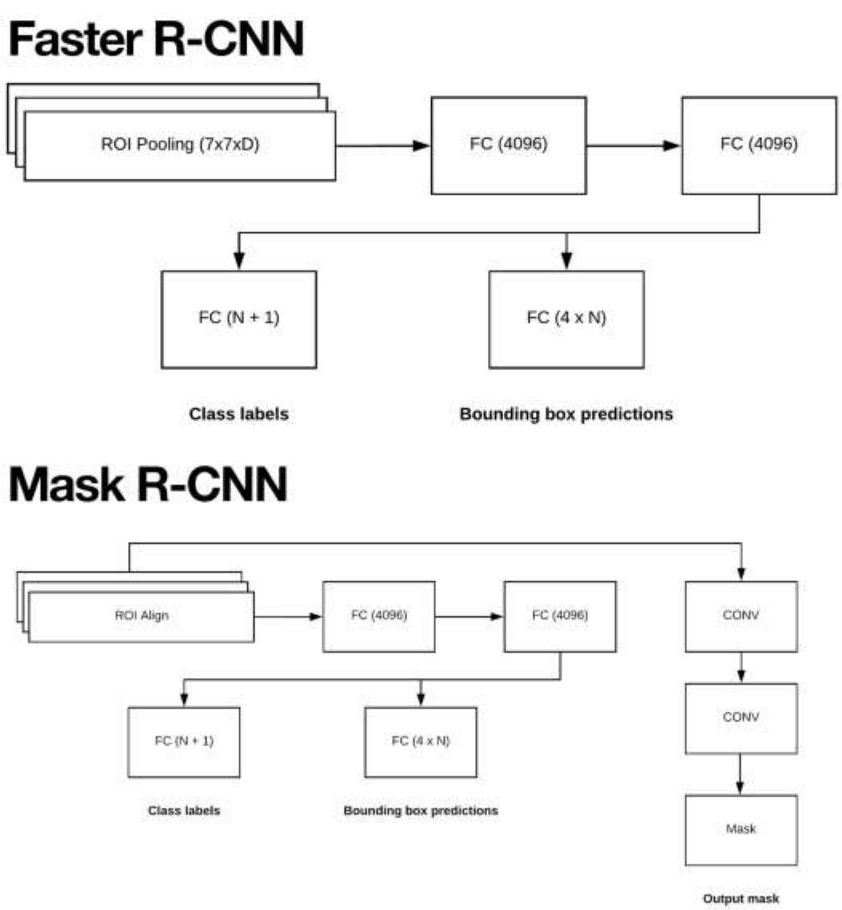


Figure 4: Faster R-CNN & Mask R-CNN [1]

2. Mask R-CNN for Segmenting Skin Lesions

2.1. Motivation and Dataset [1]

For this application we have used the dataset provided by the International Skin Imaging Collaboration (ISIC) as 2018 Skin Lesions dataset [4]. This particular dataset has been provided with the hope of facilitating and encouraging researches toward early detection of cancer, in particular melanoma which is the deadliest type of skin cancer.

Similar to any other cancer, early detection plays a vital role in melanoma treatment. When detected early enough, survival rates would be up to 95%. Therefore, early detection of the cancer is very important.

The dataset released by ISIC in 2018 is consisted of following parts (figure 5):

- Training_Input: 2594 images of skin lesions
- Training_GroundTruth: 2594 corresponding masks

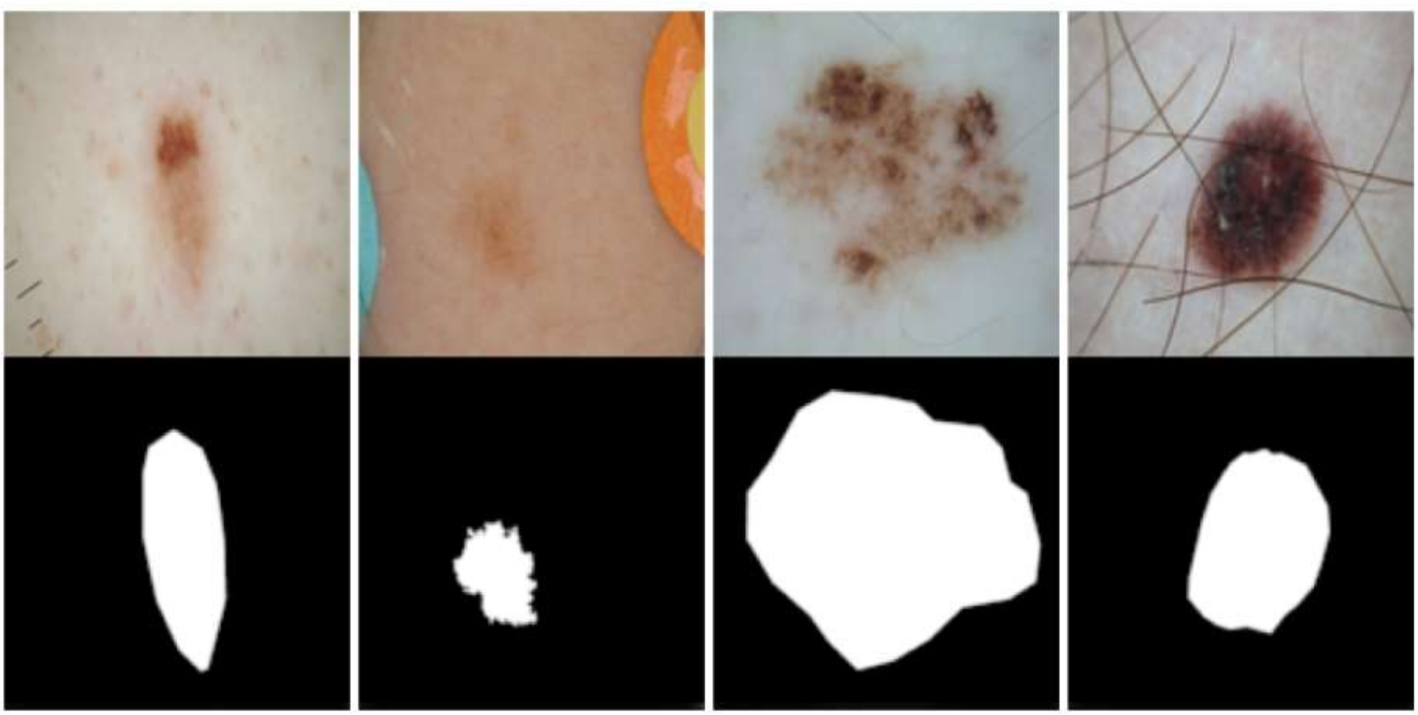


Figure 5: Sample of Dataset: the International Skin Imaging Collaboration (ISIC) 2018 Skin Lesions challenge [1]

Table 1 summarises the number of images used in training, evaluation and prediction.

Table 1: Skin Lesions Dataset

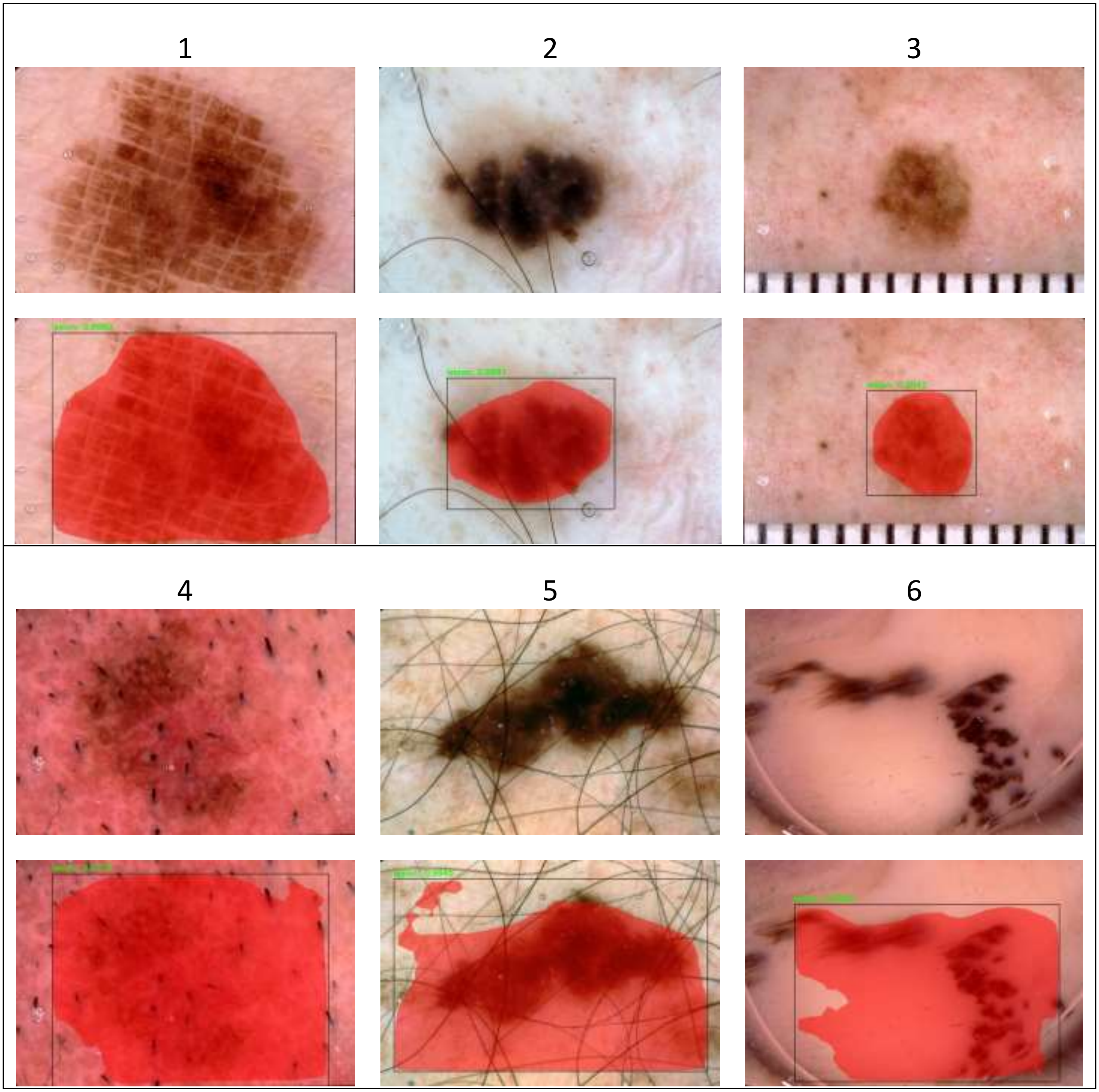
Usage	Number of Images	Source Dataset
Training	80% x 2000 = 1600	ISIC Skin Lesion Dataset (2018)
Evaluation	20% x 2000 = 400	ISIC Skin Lesion Dataset (2018)
Prediction	594	ISIC Skin Lesion Dataset (2018)

2.2. Training Mask R-CNN for Segmenting Skin Lesions

After building a base model using Mask R-CNN library [5], initial weights were loaded from mask_rcnn_coco.h5 file (trained on COCO dataset). The head model was trained for 20 epochs. Next the whole model would be trained for another 20 epochs. This last training could not be done due to memory shortage.

2.3. Results of Mask R-CNN for Segmenting Skin Lesions

2.3.1. Predicting Images from ISIC Skin Lesion Dataset



Remarks from results:

- These 6 images of skin lesions and their corresponding masks have been selected form prediction dataset in a way to represent the best and worst performances of Mask R-CNN
- Results 1 to 4 show the significant ability of Mask R-CNN network for segmenting skin lesions;
- On the other hand, results 5, and 6 show that the mask precision would suffer for challenging skin lesions (e.g. when highly scattered, covered by hair or etc.)

3. Mask R-CNN for Segmenting Round Pills

3.1. Motivation and Dataset [1]

Each year, over 3.3 million people are hurt or even dead as a result of taking incorrect prescription pill being. In order to facilitate researches in this area, the National Library of Medicine's (NLM) has proposed a pill identification challenge and provided a dataset for this challenge consisting of following parts [6] (figure 6):

- Consumer-quality Images
- Reference Images
- Ground Truth Table

Figure 6: a subset of the National Library of Medicine's (NLM) 2016 pill

Table 2 summarises the number of images and datasets used in training, evaluation and prediction.

Table 2: Datasets used for Segmenting Round Pills

Usage	Number of Images	Source Dataset	Sample Image
Training	75% x 60 = 45	NLM (2016) pill dataset-Reference Images	
Evaluation	25% x 60 = 15	NLM (2016) pill dataset-Reference Images	
Prediction	100	NLM (2016) pill dataset-Consumer-quality Images	
Prediction	20	Google Images Search Results	

3.2. Annotating images in VGG Image Annotator (VIA) [1]

Skin lesions dataset provided both input images and their corresponding masks. However, the pills dataset does not include corresponding masks so we need to creat masks for them. There are several software for this purpose. One of the most appropriate ones is VGG Image Annotator (VIA).

VIA is a installation free program that can be ran using any web browser. It also has tools for drwing circular boundries (just perfect for our application). And finally, it can export the created region boundries in text formats (json) which are compatible with python. Figure 7 illustrate an instance of annotating a pill image in this program (left) and the output attribution of annotation (right).

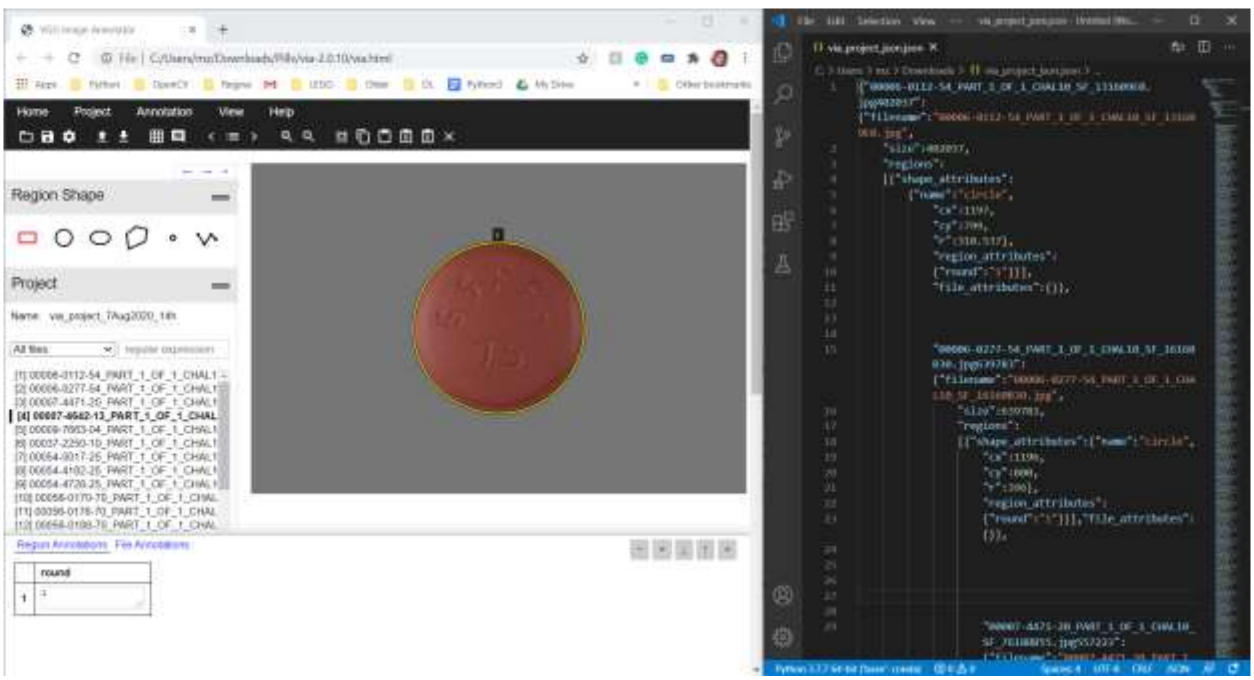


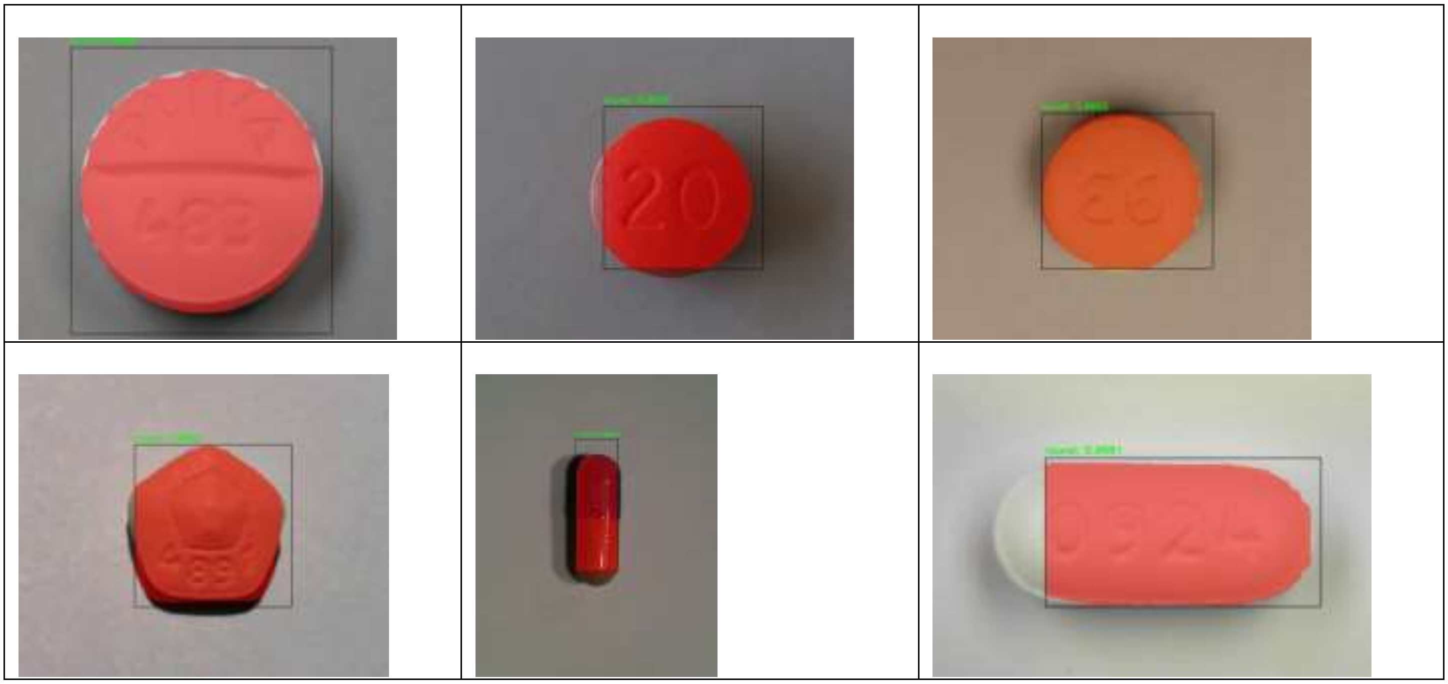
Figure 7: Left: Annotating images in VIA. Right: Annotation text in json file

3.3. Training Mask R-CNN for Segmenting Round Pills

After building a base model using Mask R-CNN library, initial weights were loaded from mask_rcnn_coco.h5 file (trained on COCO dataset) [5]. The head model was trained for 10 epochs. Next the whole model would be trained for another 10 epochs. This last training could not be done due to memory shortage.

3.4. Results of Mask R-CNN for Segmenting Round Pills

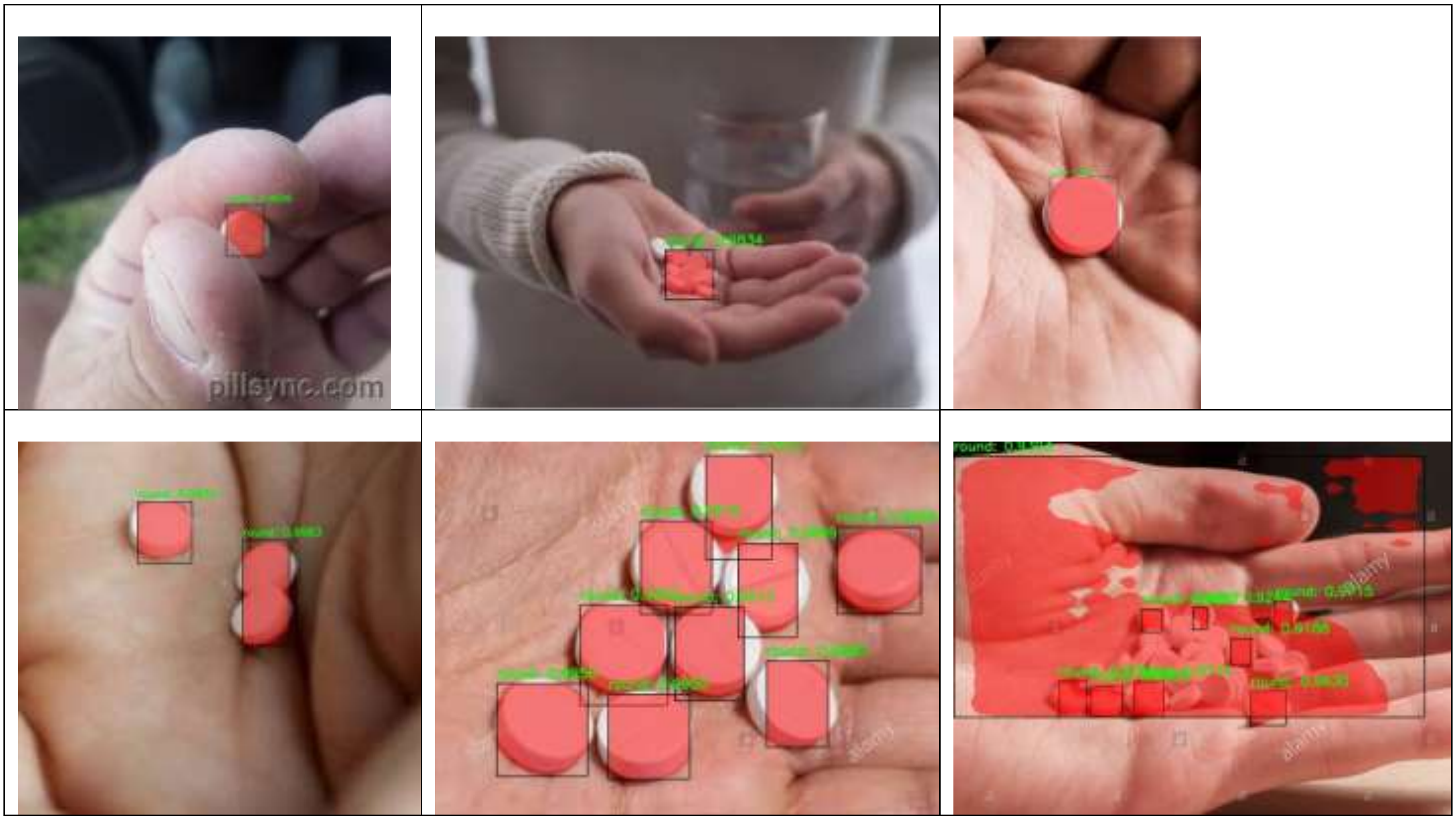
3.4.1. Predicting Consumer-quality Images



Remarks from results:

- Mask R-CNN is successfully segmenting and masking pills from Consumer-quality Images;
- However, the problem is that it mistakenly detects non-round pills as round.

3.4.2. Predicting Images from Internet



Remarks from results:

- Mask R-CNN precision is significantly lower when applied to images of pills in hands; I
- For 30% of images, Mask R-CNN is able to segment all round pills; However, for the other 70% (such as the result 6) segmentation is not satisfactory.

References

[1] Deep Learning for Computer Vision with Python (Bonus Bundle) by Dr. Adrian Rosebrock

[2] https://developers.arcgis.com/python/guide/how-maskrcnn-works/

[3] https://towardsdatascience.com/semantic-segmentation-popular-architectures-dff0a75f39d0

[4] https://challenge.isic-archive.com/data#2018

[5] https://github.com/matterport/Mask_RCNN

[6] https://pir.nlm.nih.gov/challenge/submission.html

Appendix A

Required software and libraries for implementation of Mask RCNN 2.1

Many issues arising during installation and ruuning are due to incompatible versions of CUDA, cuDNN, tensorflow-gpu, and Keras. Thanks to Niloufar explanations about how to find the compatible versions for each of these libraries, I could successfully run Mask RCNN using following configuration:

Table 3 Required software and libraries for implementation of Mask RCNN 2.1

Hardware/Software/Library	Version	Useful Link
Operating System	Windows 10	
GPU	GEFORCE GTX 1660 Ti	
Visual Studio	2017 (RTW and all updates)	https://docs.nvidia.com/cuda/archive/10.0/cuda-installation-guide-microsoft-windows/index.html
NVIDIA CUDA	10.0	https://www.tensorflow.org/install/source_windows#gpu
NVIDIA cuDNN	7.4	https://www.tensorflow.org/install/source_windows#gpu
Python	3.6	
tensorflow_gpu	1.13.1	
Keras	2.2.2	
opencv-python		
opencv-contrib-python		
imgaug		
ipython		
imutils		
Mask RCNN	2.1	https://github.com/matterport/Mask_RCNN