УДК 004.932.2

ANALYSIS OF THE BASIC NEURAL NETWORK APPROACHES TO THE PROBLEM OF THE INSTANCE SEGMENTATION

БОБРОВА НАТАЛЬЯ ЛЕОНИДОВНА
к.т.н., доцент
ХАРКЕВИЧ АНТОН ПАВЛОВИЧ,
СТЕЦКО ВАДИМ ЮРЬЕВИЧ
магистранты
УО «Белорусский государственный университет информатики и радиоэлектроники»

Научный руководитель: Шевалдышева Елена Зигфридовна
к.ф.н., доцент
УО «Белорусский государственный университет информатики и радиоэлектроники»

Аннотация: в данной работе рассмотрены две популярные нейросетевые модели для осуществления сегментации экземпляров на изображении, изучены ключевые отличия данных моделей от прошлых моделей, а также произведено их сравнение на при их обучении практике.
Ключевые слова: Python, mask r-cnn, yolo, нейросеть, сегментация экземпляров.

АНАЛИЗ ОСНОВНЫХ НЕЙРОСЕТЕВЫХ ПОДХОДОВ К ЗАДАЧЕ СЕГМЕНТАЦИИ ЭКЗЕМПЛЯРОВ НА ИЗОБРАЖЕНИИ

Bobrova Natalya Leonidovna,
Kharkevich Anton Pavlovich,
Stetsko Vadim Yurievich

Scientific adviser: Shevaldysheva Elena Zigfrididovna

Abstract: This paper reviews two popular neural network models for performing instance segmentation in an image, examines the key differences between these models and past models, and compares them during training practices.
Keywords: Python, mask r-cnn, yolo, neural network, instance segmentation.

Currently, neural networks are widely used for image analysis.

The main neural network tasks are image classification, object detection and image segmentation.

Image classification answers the questions:

1) is there an object of a given class in the image?
2) does the image belong to a given class?

Object detection allows you to find an object that belongs to a certain class (people, animals, cars, etc.) in an image.

Semantic segmentation works with many objects of the same class as a single whole.

Instance segmentation is similar to semantic segmentation, but it divides similar objects into many instances of a class.

You can see the examples of basic neural network tasks below in Figures 1 and 2.

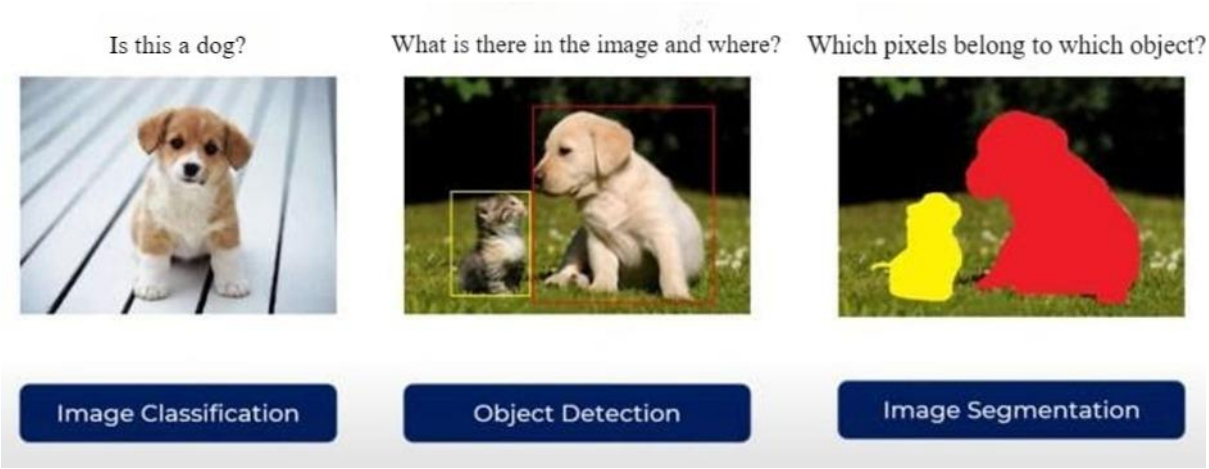

Fig. 1. Image classification, object detection, image segmentation

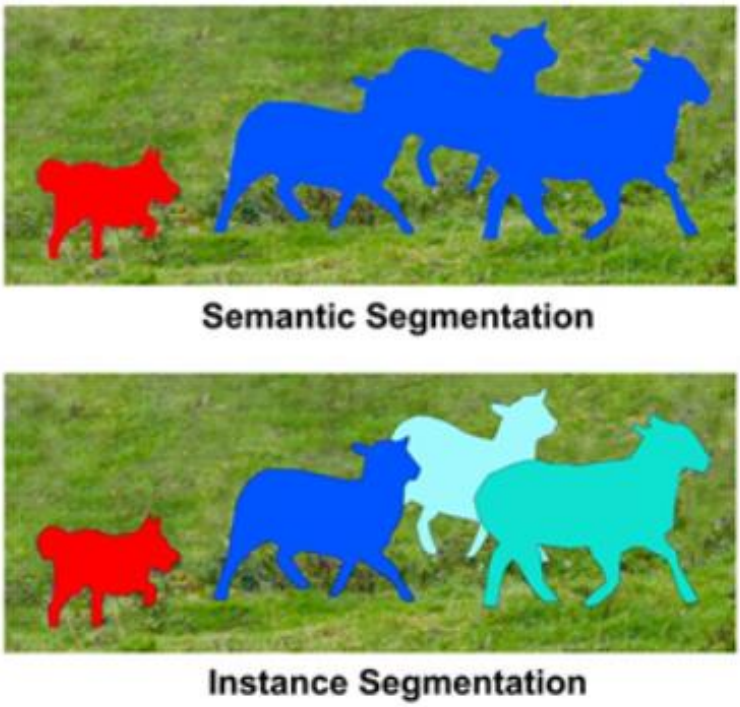

Fig. 2. Semantic segmentation, instance segmentation

It is important to understand the key advantage of neural networks that perform instance segmentation. While earlier solving the problem of object detection and semantic segmentation required two models working in parallel, in our case of instance segmentation two models are no longer required. The model that performs instance segmentation performs these two tasks simultaneously.

Let's consider two of the most popular neural network models applicable to the problem of instance segmentation.

One of the most popular neural networks for performing instance segmentation is MASK R-CNN. It is based on Faster R-CNN, but it allows you to display an object mask.

The architecture of MASK R-CNN is shown in figure 3 below:

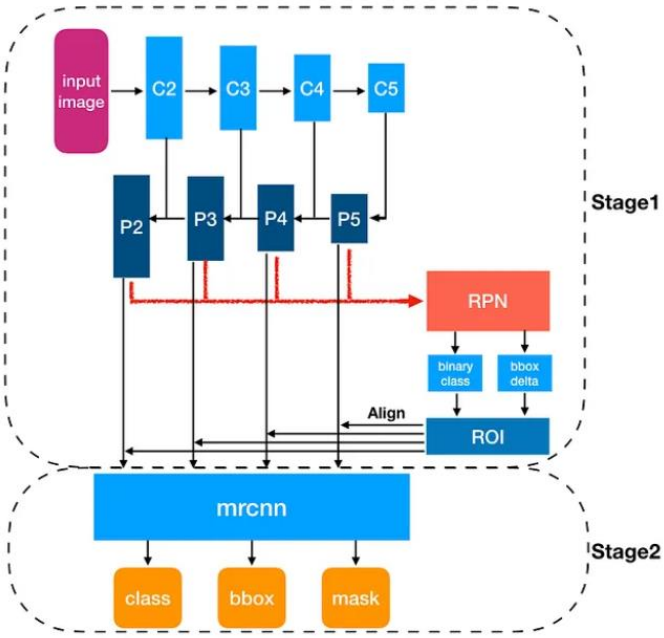

Fig. 3. The architecture of MASK R-CNN

The main differences between MASK R-CNN and Faster R-CNN [1, p.4]:

1) In addition to outputting the predicted class of the object and the predicted region of the object, MASK R-CNN has an additional output, from which a binary mask of size m×m is determined for each of the K predicted classes [2, p.7].

The outputs of Faster R-CNN and MASK R-CNN can be seen below in Figures 4 and 5.

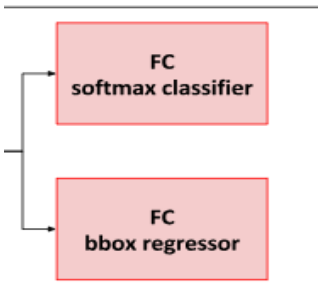

Fig. 4. Faster R-CNN output

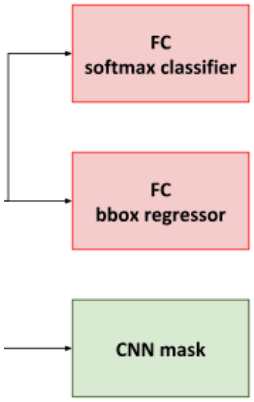

Fig. 5. MASK R-CNN output

2) The ROI Pooling layer was replaced by the ROI Allign layer in MASK R-CNN [3, p.7].

Max Pooling is an operation that divides an area of interest into non-overlapping cells of the same shape and generates an output by taking the maximum value found in each cell.

When we retrieve values for a region using the RoIPool procedure, the region's coordinates are rounded and the cells into which the region is divided are aligned to feature boundaries. If we only want to get the class of an object and its location in the image, this not very accurate approach suits us quite well, since these characteristics of the object are resistant to small shifts. But this simplification does not provide sufficient accuracy to obtain an object mask.

When we perform the ROI Allign operation, coordinates are not rounded, and cells are not shifted. Instead of this we use bilinear interpolation of 4 points in each cell to display the output information.

For better understanding, let's give an example in figures 6-8:

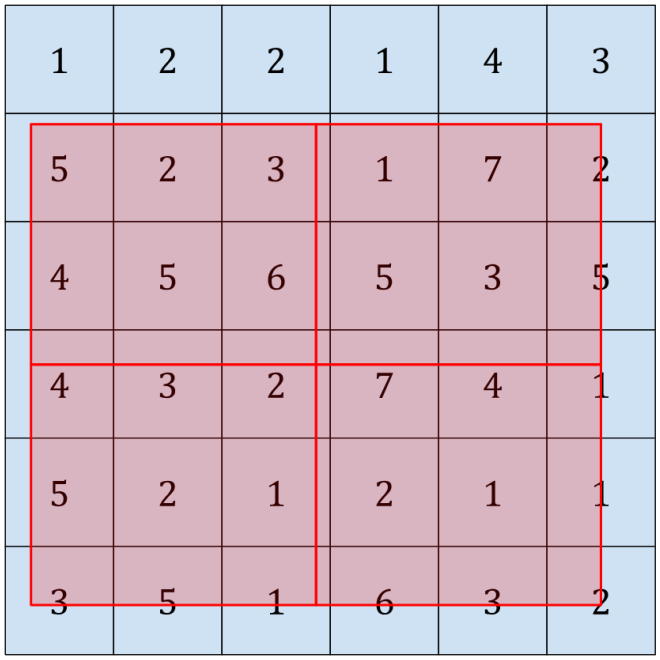

Fig. 6. Splitting the source area into 4 cells

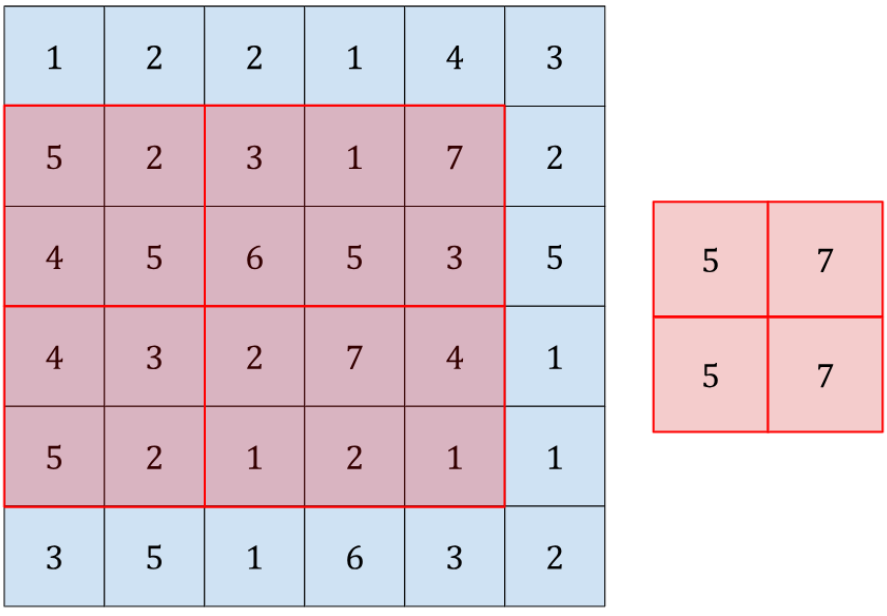

Fig. 7. Performing the ROI Pool operation

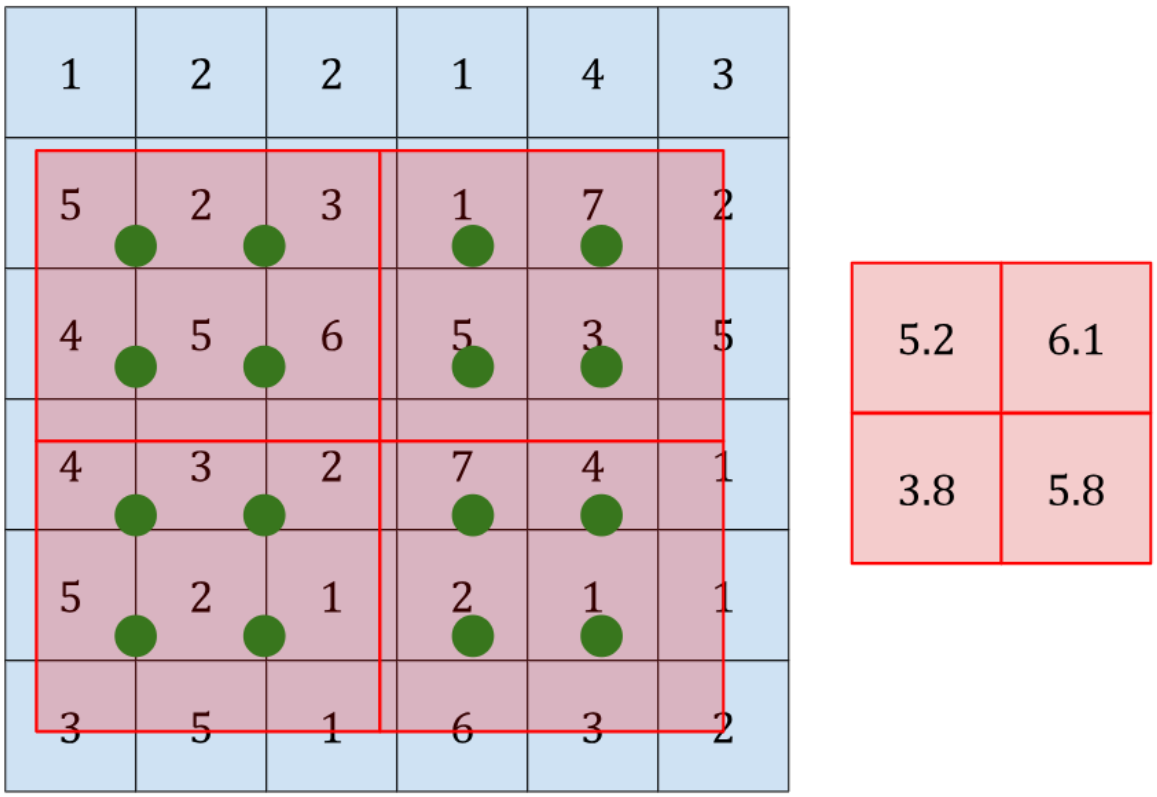

Fig. 8. Performing the ROI Allign operation

The example shows that the layer replacement was done to improve the accuracy of the output.

3) Changing the loss function.

For Faster R-CNN, the loss function looked as follows in figure 9:

$$L = L_{cls} + L_{box}$$

Fig. 9. Loss function for Faster R-CNN

Lcls is responsible for class selection and Lbox is responsible for coordinate region error. For MASK R-CNN, the loss function looked as follows in figure 10:

$$L = L_{cls} + L_{box} + L_{mask}$$

Fig. 10. Loss function for MASK R-CNN

So, a new Lmask summand has been added to the loss function for the Faster R-CNN model. Lmask is defined as the average of the cross-entropy error over the generated mask.

Let's consider a second popular network that performs instance segmentation. This network is YOLO.

This network has several versions. Starting from YOLOv7 this network can be used for various tasks, including the task of instance segmentation.

This is achieved by integration with the BlendMask algorithm.

The architecture of YOLOv7 designed for instance segmentation is shown in figure 11 below.

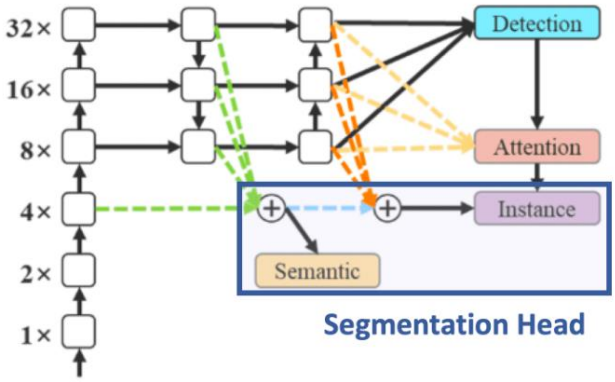

Fig. 11. YOLO architecture for instance segmentation

The architecture of the BlendMask algorithm consists of three parts [4, p.3].
It is shown in figure 12 below.

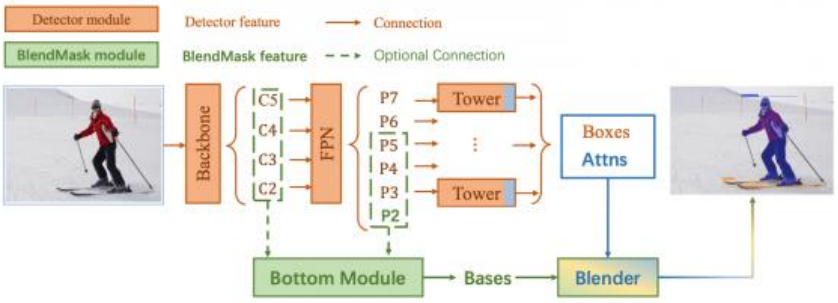

Fig. 12. BlendMask algorithm architecture

The first part of the algorithm is a lower module that predicts score maps.
The second part of the algorithm is an upper module that predicts attention maps for objects.
The third part of the algorithm is a Blender module that combines score maps and attention maps.
An example of how the algorithm works is shown in figure 13 below:

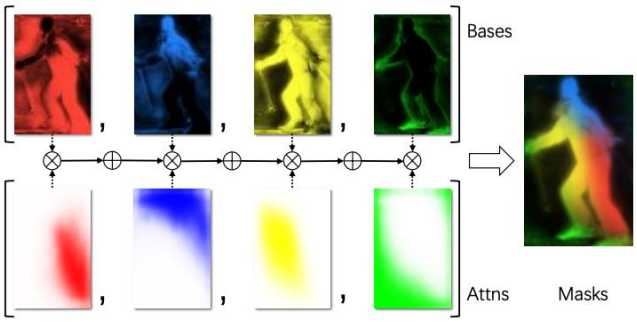

Fig. 13. Example of BlendMask algorithm

The first row in the figure above contains the score maps and the second row contains the attention maps. Here ⊗ represents the item product and ⊕ represents the item sum. Each score map is multiplied by its attention map and then summed to output the final mask maps according to the predicted object boundaries.

Let's check the models' performance in practice.

For this purpose, let's take a small dataset (about 300 images for training) with images of butterflies.

An example of an image from the dataset is shown in figure 14 below:

Fig. 14. Example of an image in a dataset

Let's bring the data in the dataset to the desired form, and then let's start training the models.

Finally, we have the following results:

The YOLOv8 model was trained in a google colab environment for about half an hour, the training went in the section of 100 epochs, the final accuracy of the model was 98 percent, and the model showed high accuracy already on the first iterations of training.

The model training metrics are summarized in figure 15 below:

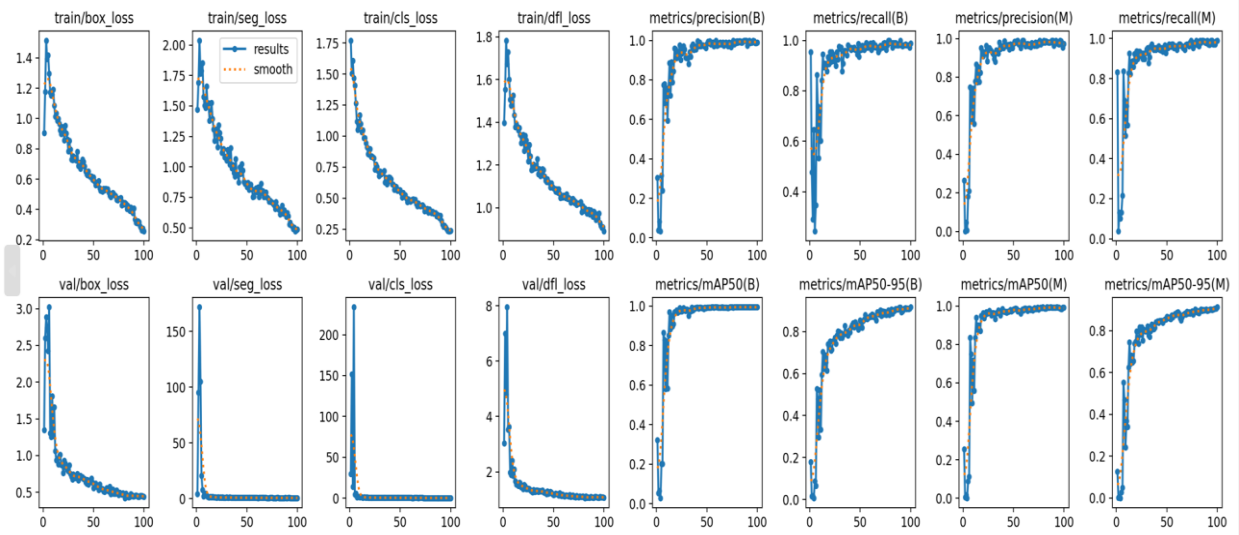

Fig. 15. YOLOv8 model learning metrics

An example of a segmented image is shown in figure 16 below:

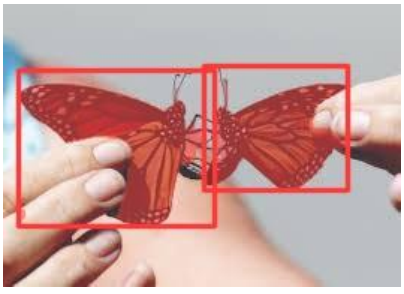

Fig. 16. An example of a segmented image

During training the MASK R-CNN model, a large number of problems arose due to the fact that MASK R-CNN was written on an outdated version of the tensorflow 1.x library, while modern environments such as

google colab have stopped supporting versions lower than version 2. The resulting bugs had to be fixed directly in the code in which the MASK R-CNN model was written.

The MASK R-CNN model took twice as long to train as YOLOv8 did, but no significant results in image segmentation were achieved.

Thus, based on the validation of the models in practice, we can conclude that YOLOv8 is a more modern and relevant model, it is faster to train and more accurate. It is more suitable for work on a small dataset than MASK R-CNN.

References

1. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick “Mask R-CNN”//Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2961-2969

2. S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz and D. Terzopoulos “Image Segmentation Using Deep Learning: A Survey”//IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 7, pp. 3523-3542

3. Yiqing Zhang,Jun Chu,Lu Leng andJun Miao “Mask-Refined R-CNN: A Network for Refining Object Details in Instance Segmentation”//Sensors, 2020, vol.20, no. 1010

4. Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, Youliang Yan “BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation”//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8573-8581