Object Detection Using MASK R-CNN

K. MARUTHI PAVAN SURYA[1], M. PADMASRI BALASUBRAHMANYAM[2], N. KUMAR VENKATA SIVA[3], M. VENKATA SIVANJANEYULU[4]

[1, 2, 3, 4] *Department of Electronics and Communication Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur, Andhra Pradesh, India*

Abstract- In today's world, Automation takes the front seat of growth, and one of the important things for that machine vision. We need good models for better results for computer vision. Here we present a conceptually clear, versatile, and general structure for segmentation of object instances. Our model detects objects in an image more efficiently and at the same time it generates a high quality segmentation mask for each object in the image. The model is called Mask R-CNN it actually extends Faster R-CNN model by adding a branch to predict an object mask in parallel with the bounding box recognition branch. Mask R-CNN is quick to train and adds a slight overhead to Faster R-CNN which runs at 5 fps. In addition, Mask R-CNN is easy to generalize for other functions, e.g. allowing us to estimate human poses within the same framework.

Indexed Terms- CNN (Convolution Neural Network), ROI (Region of Interest), Instance Segmentation, Region Proposal Nertwork (RPN)

I. INTRODUCTION

Performing Instance Segmentation involves the precise identification of each and every objects in an image, while simultaneously segmenting each instance precisely. It therefore incorporates elements from the classical computer vision tasks of object detection, with the objective of classifying individual objects and locating each object using a bounding box, and semantic segmentation.

Here the goal of the process is to classify each and every pixel into a fixed set of categories without differentiating instances of the objects. Given this, in order to achieve good results, one would assume a complex method is required. We demonstrate, however, that a surprisingly simple, versatile, and fast framework can surpass the results of prior best instance segmentation results.

The model is called Mask R-CNN, which extends already an existing model, Faster R-CNN by adding a branch to predict an object mask on Region of Interest (ROI), in parallel with the classification, bounding box regression branch.

The mask branch is a small Fully Convolutional Network added to each ROI which predicts a pixel-to-pixel segmentation mask. Given the Faster R-CNN architecture, which encourages a wide range of versatile architectural designs, Mask R-CNN is easy to implement and train. Moreover, the mask division only adds a minimal overhead computational, allowing for a simple system and simple trial.

Mask R-CNN is, in theory, an intuitive extension of Faster R-CNN, but properly designing the mask branch is crucial for good performance. Most notably, the Faster RCNN was not designed to match pixel to pixel between inputs and outputs of the network.

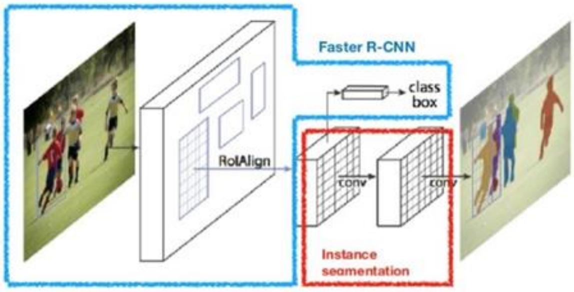

Figure-1: Mask R-CNN architecture

This is most apparent in the way RoIPool conducts coarse spatial quantization for extraction of information, the de facto core process for attending instances. We propose a plain, quantization-free layer, called RoIAlign, to correct the misalignment, which faithfully preserves exact spatial locations.

II. INSTANCE SEGMENTATION

Using the effectiveness of Region based CNNs, many approaches to instance segmentation are based mostly on proposals for segments. Older techniques resorted to divisions from the bottom up. DeepMask and subsequent works learn to suggest segment candidates which are then identified by Fast R-CNN. In these approaches, segmentation precedes slow and less reliable identification.

Li et al, proposed a model combining segment proposal system and object detection system for "fully convolutional instance segmentation" FCIS. The general idea here is to completely convolutionally predict a series of position sensitive output channels. These channels address object classes, boxes, and masks simultaneously, thus making the device fast. Yet FCIS reveals systemic mistakes regarding simultaneous instances and produces false edges.

The effectiveness of instance segmentation is also driven by another family of solutions to semantic segmentation. Beginning with results of per-pixel classification (e.g., FCN outputs), these approaches aim to split the pixels in the same group into various instances. Mask R-CNN is based on an instance-first approach, contrary to the segmentation-first approach of those methods.

2.1 Advantage of Instance segmentation

Classification tells us the object belongs to a given class. It is not considering the image's complex pixel level structure. It consists of having a prediction of a whole data.

Semantic segmentation allows complex predictions stating labels for each and every pixel so that each pixel in the image is labelled with the respective class of its enclosing object.

Object detection not only offers the class names for the objects but also gives the spatial position of those classes. It takes overlapping of objects into account.

Instance segmentation involves recognition of the boundaries for every object in the image at the level of the pixels.

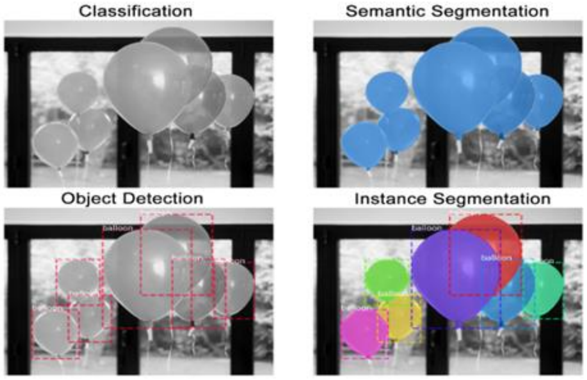

Figure-2: Difference between results of other methods and Instance segmentation

III. EXISTING METHODS

3.1 R-CNN (Region based CNN)

The Region-based CNN (R-CNN) approach towards bounding-box object detection is to attend to a realistic number of candidate object regions and independently test the convolutional networks of each RoI. R-CNN has been expanded to allow the use of RoIPool to attend to RoIs on feature maps, resulting in faster speed and greater precision.

3.2 Faster R-CNN

We start by looking briefly at the Faster R-CNN detector. Faster R-CNN is composed of two stages. The first step, called a Region Proposal Network (RPN), proposes bounding boxes of candidate objects. The second step, which

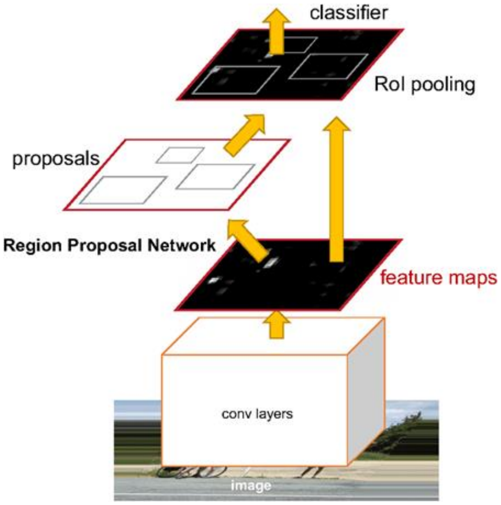

Figure-3: Faster R-CNN Architecture

is essentially Faster R-CNN, extracts from each candidate box features using RoIPool, and performs classification and bounding- regression. With a Region Proposal Network (RPN), Faster R-CNN advanced the stream in R-CNN by learning the attention mechanism. Faster R-CNN is scalable and resilient to several follow-up enhancements, and in many benchmarks, it is the current leading platform.

IV. PROPOSED METHOD

4.1 MASK R-CNN

The Mask R-CNN is a much more conceptually clear method. Where with each candidate object, Faster R-CNN has two outputs, a class tag and an offset bounding-box; to this we add a third branch that outputs the object mask; The R-CNN mask is therefore a natural and intuitive concept. Yet the additional mask output is distinct from the outputs of class and box, which involves the extraction of an object's much finer spatial structure. Next, we implement the core elements of Mask R-CNN including pixel-to-pixel alignment, which is the Fast / Faster R-CNN's main missing component.

Mask R-CNN also works in the same two-stage process, the first stage being similar to Faster R-CNN (which is RPN). In the second stage, Mask R-CNN also outputs a binary mask for each RoI, in parallel with the prediction of the class and box offset. This is contrary to most recent systems, where classification is based on predictions of masks. Our technique follows the Fast R-CNN spirit, which applies parallel bounding-box classification and regression (which turned out to simplify the original R-CNN multi-stage pipeline to a large extent).

4.2 Mask representation

A mask is that which encodes an input object's spatial structure. Therefore, unlike class labels or box offsets that are eventually condensed into short output vectors by fully-connected (fc) layers, extracting the spatial layout of masks can be handled simply by the pixel-to-pixel correspondence given by convolutions.

Specifically, using an FCN, we predict a mxm mask from each RoI. This helps each layer in the mask branch to preserve the spatial structure of the specific m x m object without collapse into a vector representation lacking spatial dimensions. Their completely convolutionary representation requires less parameters and is more accurate than previous approaches that use fc layers for mask prediction.

4.3 Network Architecture

To show the generalization of our methodology, Mask R-CNN is instantiated using multiple architectures. For clarification, we distinguish between: (i) the convolutionary backbone architecture used to extract features from a whole image, and (ii) the bounding-box recognition network head (classification and regression) and the mask prediction applied separately to each RoI.

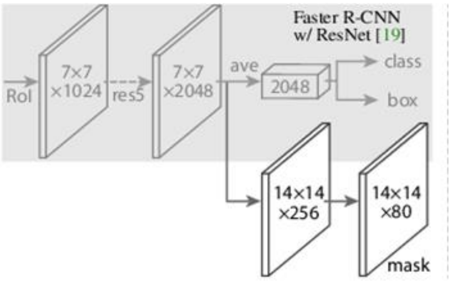

Figure-4: Head Architecture

We use the nomenclature network-depth-functions to denote the backbone architecture.

We analyze using ResNet networks of depth 101 layers.

We closely follow architectures outlined in previous research for the network head, to which we add a branch of fully convolutionary mask prediction. Specifically, we extend the ResNet and FPN papers to the Faster R-CNN box headers. See Figure 4 for details.

V. TRAINING AND RESULTS

5.1 Training

As with Fast R-CNN, if it has IoU with a ground-truth box of at least 0.5 and negative otherwise, a RoI is considered positive. The Lmask's mask loss is only established on positive RoIs. The mask target is the intersection between a RoI and the ground-truth mask it associates.

We follow the training oriented on image. Frames are resized so that they are 800 pixels in length (shorter edge). Each mini-batch is done with 2 images per GPU, and each image will be comprised of 512(N) sampled RoIs, with 1:3 positive to negative ratio. We train 1 image per GPU with the same number of iterations, with a learning rate of 0.01.

5.2 Testing

At the time of testing, the proposal number (N) will be 1000. On these proposals, we run the box prediction branch, followed by a non-maximum suppression. The mask branch is then added to 100 detection boxes with the highest scores. Although this differs from the parallel computation used in training, it accelerates testing and increases precision (due to fewer, more accurate RoIs being used). The mask branch can predict K masks for every RoI, but we only use the k-th mask, where k is a classification branch predicted class label. The output of the floating numbers is then resized to the size of the RoI and binarized at a threshold of 0.5.

Note that because we only measure masks on the top 100 detection boxes, Mask R-CNN brings a slight overhead to its Faster R-CNN equivalent (e.g., standard models at 0.20 percent).

5.3 Results

We used MS COCO Dataset 2015 to train and test the model and compared the results with the best instance segmentation results available and the results are satisfactory and the range of accuracy is about 88 percent-92 percent with real images, i.e. camera images.

If it comes to animated images or cartoon images the prediction accuracy decreases significantly.

Also the image resolution consistency plays a key role in prediction accuracy. Though the frames are resized to a length of 800 pixels, this aspect tends to influence the results.

The result images are given below in figures 5, 6. The figures 5a, 6a are inputs and 5b, 6b are outputs as we can see the objects the images are identified properly with RoI (Region of interest) i.e., a bounding around the objects and the class label is given at the top left corner of every bounding box along with it the accuracy percentage given in terms of score.

The masks for each objects are generated at the pixel-to-pixel level as per the feature maps obtained through the RPN (Region Proposal Network). These masks for the objects are generated even for the items overlapped on each other with maximum precision. So, that the computer vision can differentiate between same objects even when they are overlapped over each other as in figure 5 where two zebras are overlapped on each other but they differentiated in result with maximum precision.

Even when the objects are crowded tighter in an image the masking and the class label prediction are to the par with normal images this can be observed with figure 6.

Figure-5a: Input image

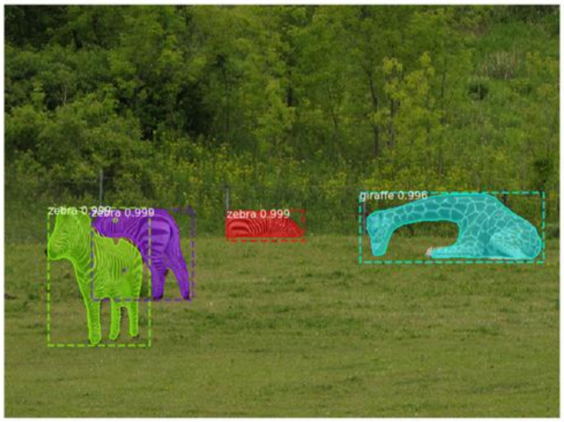

Figure-5b: Result image

Figure-6a: Input image

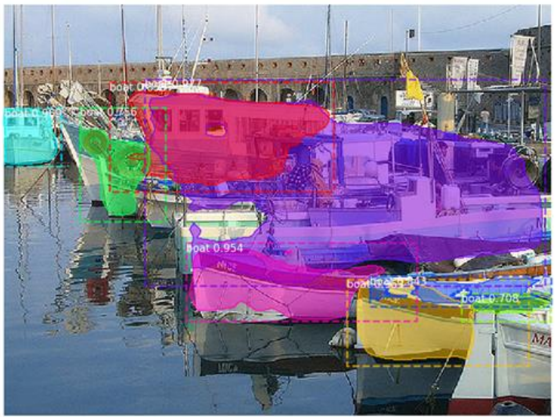

Figure-6b: Result image

CONCLUSION

Mask R-CNN is an instance segmentation technique that locates every pixel of each object in the image rather than the bounding boxes. It has two stages: regional proposals and then the proposals are classified and bounding boxes and masks created. It does this by using an additional fully convolutional network on top of a CNN-based feature map with input as feature map and gives matrix with 1 at all locations where the pixel belongs to the target and 0 elsewhere as output.

It uses the RPN, which scans all of the Feature Pyramid Network through top-bottom approach and then it proposes the regions that might consist artifacts. It uses anchors that are a series of boxes with predefined positions, and scales themselves according to the images input. Individual anchors are allocated to the bounding boxes and ground-truth classes. For each anchor, RPN produces two outputs-anchor class and bounding box specifications. The class of anchors is either a foreground class or a background class

Thus, this proves as one of the best for computer vision models. This model can be used when the most precision is required with the object identification. That is the applications such as vision for self-driving cars, for human pose estimations, Cityscapes Analysis and many more computer vision applications

REFERENCES

[1] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[2] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*. 2014.

[3] M. Andriluka, L. Pishchulin, P. Gehler, and B. Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[4] P. Arbelaez, J. Pont-Tuset, J. T. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.

[5] A. Arnab and P. H. Torr. Pixelwise instance segmentation with a dynamically instantiated network. In *CVPR*, 2017.

[6] M. Bai and R. Urtasun. Deep watershed transform for instance segmentation. In *CVPR*, 2017.

[7] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside outside net: Detecting objects in context with skip pooling and recurrent neural networks. In *CVPR*, 2016.

[8] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh. Realtime multiperson 2d pose estimation using part affifinity fifields. In *CVPR*, 2017.

[9] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The Cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016.

[10] J. Dai, K. He, Y. Li, S. Ren, and J. Sun. Instance-sensitive fully convolutional networks. In *ECCV*, 2016

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016