INSTANCE SEGMENTATION OF VIDEO USING MASK R-CNN

E. Annapoorna
Mtech, Computer Science and Engineering
Sri Venkateswara University,
Tirupati.

Dr.D.Vivekananda Reddy
Sr.Assistant Professor,CSE
Sri Venkateswara University,
Tirupati.

Abstract:

Object detection in the digital images, videos plays an important role in the real world. At present object detection is done in the images with the help of the models like Fast RCNN. In the object detection the objects in the image are identified with the help of bounding boxes around objects. The object detection is done by generating the bounding box around the object with (x,y) coordinates but with this we cannot identify which pixel belongs to the background and which pixel belongs to the foreground. Image classification only categorizes the objects in the input image. Object detection localizes every object in the input image. Whereas the semantic segmentation can classify all the similar objects as a single instance, the main drawback of this is we cannot identify how many similar objects are present in the single instance.

In semantic segmentation we can only segment objects by using bounding boxes and cannot identify individual instances of even same classes. For each instance of an object in an image the model generates bounding boxes and segmentation masks.

In the Instance segmentation objects of same class will assign a different instance. Here we can compute a mask (pixel level) for each and every object in the input image this can be achieved with the help of Mask R-CNN

In this paper the Mask R-CNN is proposed to add the mask feature to every object at pixel level along with the bounded box. The Mask R-CNN can perform the instance segmentation along with the object detection. It outperforms all the other previous models.

Keywords:Object detection, R-CNN, region proposal network, instance segmentation

1. Introduction:

Image segmentation is computer vision process which is designed to simplify the image by dividing the image into segments or group of pixels that indicates the objects or parts of objects based on some criteria. This segmentation is done based on deep learning techniques. Image segmentation is applied in many fields like self-driving, satellite imaging, Military, recognizing objects in an image etc.

The main image level tasks are classification and detection. Classification just categorizes the object in the image. Detection is recognizing the object and its localization.

Detection and segmentation both are implemented in instance segmentation. Deep learning techniques are used to perform segmentation. Instance segmentation which is a subtype of image segmentation that identifies each instance of the object within the image at pixel level.

Semantic segmentation is able to identify the objects in the image and group those objects together based on their class whereas instance segmentation is able to identify each individual object within the group of similar objects using bounding boxes. For this we use the technique called Mask R-CNN which is an extension of

Faster-CNN object detection algorithm. Mask R-CNN adds an extra feature such as instance segmentation and mask. Mask R-CNN separates each object from its background and also forms segments on every pixel of the image.

Mask R-CNN consists of two stages:

In the first stage the input image is scanned to generate the proposals.in the second stage the proposals are classified and generates mask and creates the bounding box.

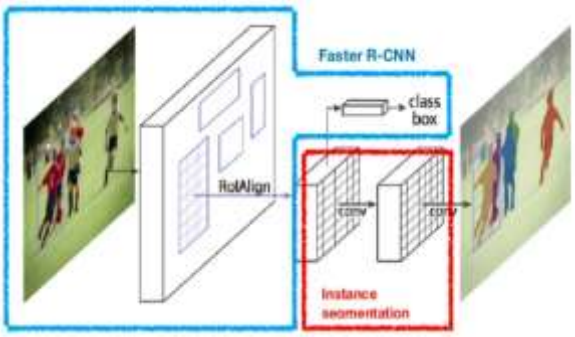

Fig 1: Mask R-CNN framework for instance segmentation

2. **Literature Review:**

In the paper [1] Ross Girshick proposed a method that uses the selective search algorithm that extracts 2000 regions from the image (region of proposals) here it uses only 2000 regions instead of working on large number of regions. It has certain limitations like it takes huge amount of time to train and classify 2000 region proposals in a single image. There is no scope for learning as the selective search algorithm is a fixed algorithm.

Ross Girshick then proposed another method Fast R-CNN[2] to overcome the limitations of R-CNN in this instead of giving region of proposals to CNN we give input image to CNN to generate feature map, from this feature map we identify the region of proposals and make them into fixed size using ROI polling layer and fed to the fully connected network. We use softmax layer to predict the class and the offset value of the bounding box. Fast R-CNN is faster than R-CNN because in R-CNN 2000 regions of proposals should be given to CNN every time but in fast R-CNN feature map is generated by doing the convolution operations once per image

Shaoqing Ren et al proposed a method Faster R-CNN[3] here instead of using selective search algorithm like R-CNN and fast R-CNN it allows the separate network in order to predict region proposal.

Mask R-CNN is an extension of Faster R-CNN faster R-CNN yields two branches class name and bounding box. Mask R-CNN adds extra branch of mask along with the class name and the bounding box. The major difference between faster RCNN Mask R-CNN is replacing ROI pooling layer with ROI align

3. Proposed Method:

In this paper we propose a technique called Mask R-CNN which is an extension of Faster RCNN, Faster R-CNN yields two branches class name and bounding box. Mask R-CNN adds extra branch of mask along with the class name and the bounding box.

In this paper we achieve instance segmentation using Mask R-CNN.The above image will depict the architecture of the mask R-CNN. It consists of (CNN) convolutional backbone, which is a pertained model like VGGNET,ResNet. RPN is a Regional Proposal Network which is used for generating region of proposal for the given input image. It gives the feature map of the image. A ROI Align Layer is used for generating fixed size of feature map. A Fully Connected Layer exists for the classification of objects in the image with the bounding box. A Mask Branch is used to generate the mask to the identified object by the fully connected Layer.

In the convolutional backbone the pertained models like VGGNET,AlexNet, GoogleNet, ResNet etc.. can be used for the feature map of the image. We can use any model in this paper the Mask R-CNN is implemented using the ResNet 101 pertained model. The ResNet 101 is trained

for object detection using the Imagenet dataset. The Imagenet dataset consists of 1000 of categories of objects with the more than the millions of the images. The input size of the image to the ResNet model is 224*224.

3.1. Working Principle:

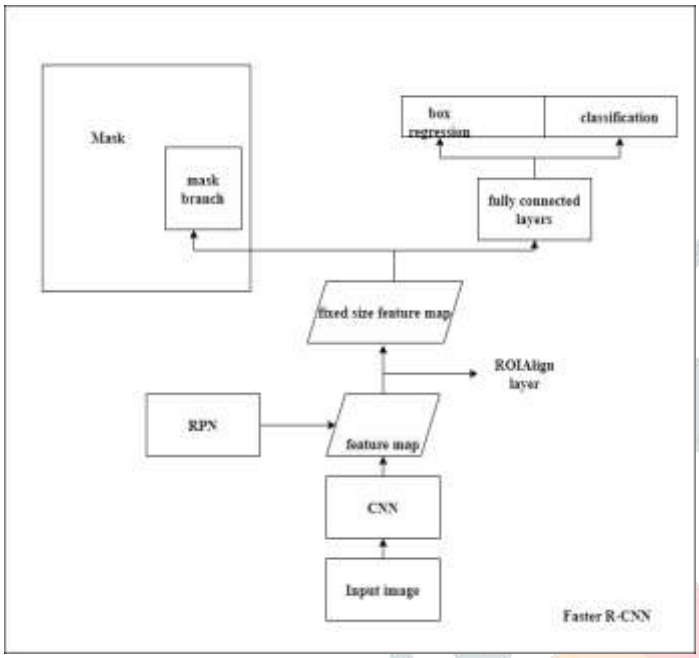

Fig 2: Architecture of Mask R-CNN

The below steps will show the working of the Mask R-CNN

- The input image is given to the Mask R-CNN then the pertained model will give feature map as the output to the RPN.
- The RPN (Regional Proposal Network) will take output of the CNN and gives the multiple region of interest using light weight binary classifier.
- The feature Map along with the ROI of the image is given to the ROI Align layer which gives the fixed size feature map as the output, by wrapping the multiple bounded boxes into fixed dimensions.
- Then the fixed size feature map from the ROI Align is given to the fully connected layers and Mask Branch.
- The fully connected Layer will classify the objects in the image with bounded boxes.
- Then the mask branch output a binary mask for each ROI of the image.
- This will output the image with the bounded boxes around the objects along with the mask to the objects.

4. Results:

Fig 3: instance segmentation using Mask R-CNN

5. Conclusion:

The proposed model will overcome the limitations of the object detection. The previous models like Fast R-CNN, Faster RNCN will can only perform the semantic segmentation cannot perform the instance segmentation. Instance segmentation can be achieved using the proposed model Mask R-CNN.Instance segmentation is used in many applications like self-driving cars, counting the persons or objects in in the image. Mask R-CNN helps us to segment objects and generate mask to every object at pixel level. It helps to segment the foreground object from the background. This acquires a great accuracy when compare to other previous models.

Reference:

[1] R. Girshick, J. Donahue, T. Darrell, and J.Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In CVPR, 2014.

[2] R.Girshick ,Jeff Donahue. “Region based convolution neural networks for accurate object detection and segmentation”. In 2015 IEEE

transactions and pattern analysis and machine intelligence.

[3]Shaqoing Ren Kaming He,Ross Girshick and Jain Sun "Faster R-CNN towards real time object detection with region proposals network" in 2016.

[4] Saining Xie,Ross Girshick,Piotr Dollar,

Kaming He,"Aggregated residual transform

ation for deep neural networks in 2017 IEEE

Conference on Computer Vision and Pattern Recognition.

[5]J.R.Uilings, K.E.A van de Sande," selective search for object"in 2013 Int J Compute Vis.

[6]Alex Krizhevsky,Ilya Sutskever and Geoffrey E.Hinton "ImageNet classification with deep convolution neural networks in 2017,communications of the ACM.

[7]Kaiming He,Georgia Gkioxari, Piotr Dollar,Ross Girshick ,"Mask R-CNN" in 2018.

[8]Lukash Bienias,Line Hagner Nielsen, Tommy Sonne Alstrem,"Insights into the behavior of multi task deep neural networks for medical image segmentation".IEEE 2019

[9]YiLi,Haozhi Qi,Jifeng Dai,Xiangyang Ji,Yichen Wei,"Fully convolutional instance aware semantic segmentation"2017 IEEE Conference on Computer Vision and Pattern Recognization.