ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Computational Calibres of Mask R-CNN in Real-Time Object Detection and Instance Segmentation

Satyam Mishra[1], Prachi Tare[2]

B.Tech Student, Dept. of Computer Science and Engineering, Medicaps University, Indore, MP, India[1]

B.Tech Student, Dept. of Electronics and Communication Engineering, Medicaps University, Indore, MP, India [2]

ABSTRACT: Convolutional Neural Networks are widely used to deal with the deep learning perspective of images and computer vision. Over the past few years, gradual modifications of algorithms related to CNN such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN have proved themselves as effective methods of object detection and instance segmentation in batch processed scenarios as well as real-time environments. These algorithms are implemented in various practical domains such as surveillance, video monitoring, vehicle detection, self-driving cars, and even in the automated photo and video editing. The progress of object detection and instance segmentation is continuously improving along with the decrease in time complexities due to new approaches and increasing processing power of machines. Object detection and instance segmentation are being implemented with more and more efficiency even in real time scenarios. Some algorithms, mainly Mask R-CNN and Faster R-CNN are enhancements of a predecessor algorithm known as Fast R-CNN. The major changes have been made during region proposals and the eradication of selective search. Furthermore, the process of annotation of bounding boxes rather than pixel-level annotation also makes it more convenient to detect new objects. It is observed that Mask R-CNN and Faster R-CNN have improved recall and precision performances. Moreover, Mask R-CNN can be extended further by the approach of transfer learning in order to add object masks for the instances in real time with customized attributes. The major benefits of using algorithms derived from R-CNN is that they outperform all the counterpart methods in the domain and can also be used in unsupervised environments. In addition to this, it can also prove beneficial for dealing with big data analytics, instance tagging, semantic indexing, and other related disciplines by modifying several layers and making them efficient by using appropriate loss functions.

KEYWORDS: Fast R-CNN, Faster R-CNN, Instance Segmentation, Masking, Mask R-CNN, Object Detection, R-CNN, Region Proposal Network

I. INTRODUCTION

Using Deep Convolutional Neural Networks is an efficient way for image processing and object detection. In most of the object detection and instance segmentation algorithms, the basic principle underlies in generating the region proposals in order to get the object boundary in a naive sense. It is accomplished by selective search like in the case of Fast R-CNN. Moreover, convolutional layers are added to the R-CNN for region proposal network, mapping, and regression which in turn make the R-CNN even more efficient. Whereas, in Faster R-CNN and Mask R-CNN, the region proposal networks are generated in even more lower time complexities because of a separate network and recursive Bayesian filtering which is used for detecting the regions. This lowers down the computational power and complexity required for the process as compared to that of Fast R-CNN. Furthermore, anchor boxes are formed around the objects which are also called as anchors based on the best possible position of the anchors. Mask R-CNN is an extension of Faster R-CNN where masks for the objects in an image are processed simultaneously along with the bounding boxes which enhances instance segmentation. In other words, Mask R-CNN can also be described as Faster

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

R-CNN with a branch of object masking for instance segmentation [1]. Generally, Mask R-CNN and Faster R-CNN are backed-up with high-level semantic feature maps such as the feature pyramid network architecture to carve out the features.

II. RELATED WORK

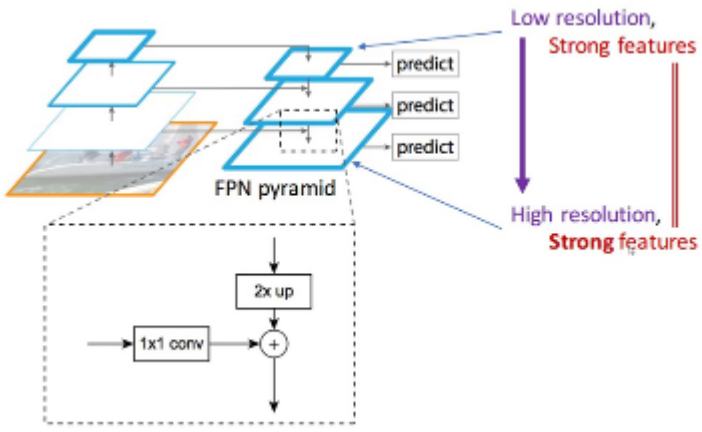
In the domain of deep learning, instance segmentation and masking is often carried out via convolutional neural networks. Furthermore, one of the widely used innovation in the aforementioned area is DeepMask. The masks are generated after region proposals through the help of discriminative convolutional network. The backbone of instance segmentation lies upon previously used convolutional network systems such as AlexNet, VGG, GoogleNet and many more. In addition to this, Fully Convolutional Instance Segmentation (FCIS) is also an ardent approach used to predict the position sensitive output channels. However, FCIS has minute windows of error which can be overcome by implementing twitches with gradual modifications of Faster R-CNN.

III. MASK R-CNN

Mask R-CNN starts with the process of sorting and refining the anchors. A positive and a negative anchor is proposed around every object which is then refined further. These anchors can be customized according to the kind of objects one wants to detect. Furthermore, these anchors play a significant role in the overall performance of Mask R-CNN. This is followed by the appearance of bounding boxes/regions as a result of the region proposal network. This is a very crucial point because region proposal network decides whether a particular object is a background or not and likewise bounding boxes are marked. The objects in the frame are determined by a classifier and a regressor by region of interest pooling. After refining of bounding boxes, masks are generated and placed on their appropriate positions on the object. This is the main distinguishable attribute of Mask R-CNN.

Backbone Network

This is the first network which processes the image in Mask R-CNN. Generally, ResNet50 or ResNet101 convolutional neural network is used for feature extraction. Primary and rudimentary features are extracted by the initial layers of this network. Whereas, high-level feature extraction is accomplished by proceeding layers. The objects of different scales are detected by the Feature Pyramid Network (FPN). Even though pyramid rendering is eschewed because of computational and memory demand, pyramid hierarchy is used in order to obtain feature pyramids without heavily compromising the efficiency. FPN allows access of features at higher as well as lower levels.



Representation of FPN

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Region Proposal Network (RPN)

Faster R-CNN can be termed as the backbone of Mask R-CNN. The Region Proposal Network (RPN) is the first phase in Faster R-CNN. It is during this stage, the most suitable bounding boxes are determined along with the objectness score. These anchor boxes might vary in size because of the variance of object dimensions in the image. The classifier is trained in order to distinguish the background and foreground. Moreover, if the anchors are turned to vectors with two values and then fed to an activation function, labels can be predicted. A regressor of bounding box refines the anchors and a regressor loss function determines the regression loss.

$$L_{\text{loc}}(t^u, v) = \sum_{i \in \{x,y,w,h\}} \text{smooth}_{L_1}(t_i^u - v_i),$$

in which

$$\text{smooth}_{L_1}(x) = \begin{cases} 0.5x^2 & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise,} \end{cases}$$

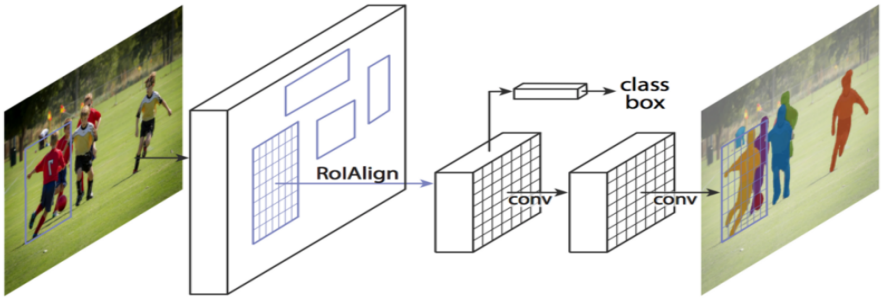
(Smooth L1 Loss Function of the Regressor)

The region proposal network determines the most appropriate bounding boxes for the objects with the help of classification loss and the regression loss.

ROI Pooling and Classifier Training

The RPN gives the most appropriate regions of bounding boxes. The predicament here is that they are of different dimensions. To eradicate this difference, ROI Pooling is applied in order to commensurate the feature maps of CNN. In other words, ROI Pooling makes the process easier and efficient by converting many feature maps of different sizes into the same dimensions which makes a feasible structure for further processing. The feature map is divided into a fixed number of equal sized regions and then max pooling is applied to them. This makes the output of ROI Pooling constant irrespective of the size of the input. In this stage, features are extracted and can also be used for further conjectures.

After this, RPN, classifier, and regressor can be trained altogether as well as distinctively. When the aforementioned three are trained together, the speed spikes up to 150% while the accuracy remains unhindered. Stochastic gradient descent and back-propagation can be used to train the region proposal network by following the image-centric sampling.



ROI pooling & convolutional layers

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Mask Generation and Layer Activation

Mask is generated upon objects simultaneously while bounding boxes are predicted by the RPN. The process of instance segmentation is the key feature behind generating the masks for the objects in an image. This occurs at the pixel level and is the extension of Faster R-CNN which results in Mask R-CNN. Low-intensity masks are generated in accordance with the classifier. These soft masks are more meticulate than binary masks. The loss is calculated during training and then the object mask is augmented to the size of bounding boxes. Moreover, activations of layers can be examined individually for detecting the noises (if any) and to rectify it. This generates a mask for each object in accordance with the input.

Noise Inspection

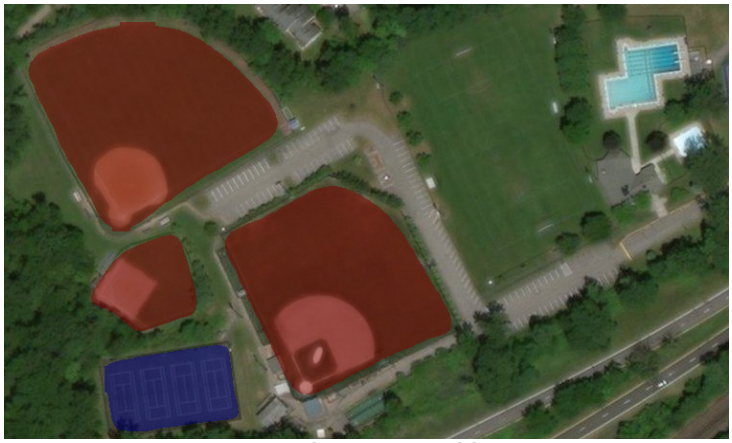
The activation layers may or may not have randomness and noise which affect the masks. This can be eradicated by inspecting them separately. Apart from analyzing the layers, noise can be detected by going through weight histograms and visualizing from tensorboard. By inspecting the noise, the layers can be adjusted accordingly and as a result, the efficiency of the neural network can be increased.

IV. IMPLEMENTATION OF MASK R-CNN

Mask R-CNN can be extended & implemented further in various domains by customizing the layers and modifying for selective environments by introducing relevant dataset for training. Apart from static pictures, Mask R-CNN can also be applied in videos in order to implement segmentation.

Geospatial Mapping

Satellite images can be used as the dataset which can be trained in order to identify desired structures using Mask R-CNN. The trained structures can be segmented and masks/labels can be attached to them. Moreover, images to OSM is an ardent & generic implementation of Mask R-CNN on a dataset. A neural network is extensively trained in this implementation of Mask R-CNN. Furthermore, sport fields in Massachusetts are delineated along with the OpenStreetMap. The training is done iteratively until a satisfactory accuracy rate is obtained. Simultaneously, the algorithm is also bootstrapped and other details are fine-tuned. Since Mask R-CNN requires a hefty amount of data, the training process takes around 3-4 days. In the final stage, outputs of the neural network are converted into OSM files.



Instance Segmentation in OSM

In general, OSM uses the following functions:

- Simplex Optimizer function (or a partial derivative function)
- Huber Cost Function (standard least square cost function can also be used, but the former one is more efficient because boundary fitting is not Gaussian)

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

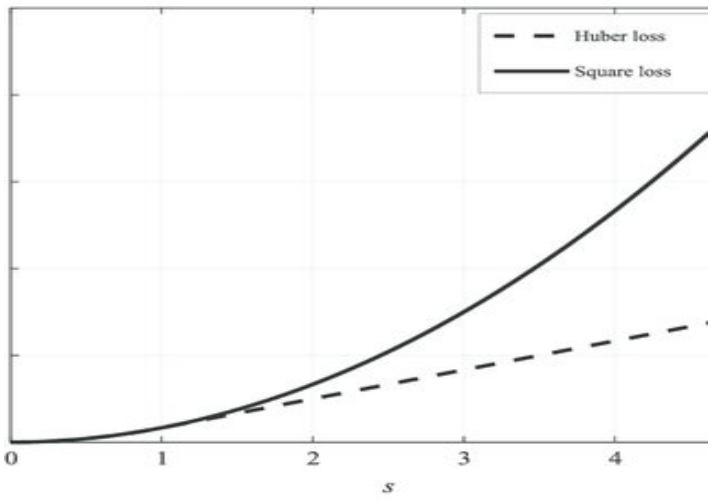
(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

$$L_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2 & \text{for} |y - f(x)| \le \delta, \\ \delta\,|y - f(x)| - \frac{1}{2}\delta^2 & \text{otherwise.} \end{cases}$$

Huber loss function



Huber and SLS Function

Color Splash Effect

Mask R-CNN can also be used for applying the color splash filter in images as well as videos in real time. One of the best parts about implementing Mask R-CNN is that fewer images are needed for dataset because of transfer learning along with the COCO dataset. After selecting the images for the dataset, they are annotated by various tools such as COCO UI, RectLabel and VGG Image Annotator (VIA). If VIA is used for image annotation, the segmentation mask details are stored in a JSON file. This is followed by training wherein RPN gives us the object masks. After this, the image's grayscale version is taken and the region suggested by the RPN is given the pixel values.

Color Splash on taxis

In this way, the desired objects in the image obtain color and color splash filter can be obtained. This can be achieved in videos too.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

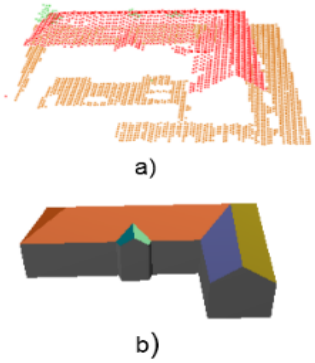
Color Splash / Pop effect (COCO Dataset)

In addition to this, Mask R-CNN can also be extended by combining it with a semi convolutional substructure for coloring instances. In the semi convolutional framework, if x □ X = $R^{H\times W\times 3}$ is an image then, u□Ω={1,…,H}×{1,…,W} is a pixel. In this context, $S_0 = \Omega - \cup_k S_k$ represents the background where image is mapped into total of $S_x = \{S_1,\ldots,S_{Kx}\}$ □ 2^{Ω} mapped regions. It is an alternative approach to proposal based instance segmentation.

Three-Dimensional Modeling

Three-dimensional reconstruction is indeed one of the finest implementations of Mask R-CNN. Mask R-CNN can be applied to model a three-dimensional structure from a map. Datasets such as the aerial LiDAR can be used to train the model with the types of roofs in a particular locality and then it is converted into a Digital Surface Model (DSM) raster. Moreover, manual digitization of the roofs detected by it can be implemented which can be executed by ArcGIS procedural rules. The main backbone of the digitization lies upon the detection of gables and hip segments of roofs. The steps followed are:

- Channels of RGB are detected and aerial LiDAR is rasterized.
- Hip and Gable Segments are digitized manually.
- Three dimensional reconstruction is executed upon hip and gable segments.



LiDAR (a) used for remodeling of (b) with the help of implanted gable and roof hip

This method of three-dimensional modeling can be applied to buildings of varying dimensions and details. However, the accuracy might drop with the increase in details. In addition to this, a similar approach can be used for reconstruction of other objects as well given that, sufficient data is given.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

V. CONCLUSION

In the domain of computer vision through deep learning, object masking and segmentation is evolving continuously at a significant rate. This is because of gradual modifications in the efficiency of algorithms. A twitch in Faster R-CNN resulted in this state of the art instance segmentation practice. Mask R-CNN is one of the ardent innovations in this field. It can be extended towards applying color splash and pop effects, detecting desired objects in pictures or videos, real-time object instance segmentation as well as three-dimensional reconstruction. In addition to this, the future scope of Mask R-CNN can also be directed towards its combination with virtual reality and augmented reality based systems.

REFERENCES

1. Kaiming He, Georgia Gkioxari, Piotr Dollar, Ross Girshick, " Mask R-CNN", arXiv:1703.06870v3 [cs.CV], Open Source, 2018
2. Shaoqing Ren, Kaiming He, Ross Girshick, Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks", arXiv:1506.01497v3 [cs.CV], Open Source, 2016
3. Radhamadhab Dalai, Kishore Kumar Senapati, "A Robust Image Object Recognition Method Using RCNN and Big Data", International Journal of Creative Research and Thoughts, Vol. 5, issue 3, 2017
4. Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, Serge Belongie, "Feature Pyramid Networks for Object Detection", arXiv:1612.03144v2 [cs.CV], Open Source, 2017
5. Ross Girshick,"Fast R-CNN Object Detection with Caffe", Microsoft Research, 2015
6. David Novotny, Samuel Albanie, Diane Larlus, Andrea Vedaldi, "Semi Convolutional Operators for Instance Segmentation", IEEE European Conference of Computer Vision, 2018
7. Matterport, Mask_RCNN. Github Repository, https://github.com/matterport/Mask_RCNN
8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, "Deep Residual Learning for Image Recognition", IEEE Conference on Computer Vision and Pattern Recognition, 2016
9. [9] Saining Xie, Ross Girshick, Piotr Dollar, Zhuowen Tu, Kaiming He, "Aggregated Residual Transformations for Deep Neural Networks", IEEE Conference on Computer Vision and Pattern Recognition, 2017
10. Jifeng Dai, Yi Li, Kaiming He, Jian Sun, "R-FCN: Object Detection via Region-based Fully Convolutional Networks", arXiv:1605.06409 , Open Source, 2016
11. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks", NIPS, 2012.
12. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions" CVPR, 2015.
13. P. O. Pinheiro, R. Collobert, and P. Dollar, " Learning to Segment Object Candidates" ,NIPS, 2015
14. Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully Convolutional Instance Aware Semantic Segmentation",CVPR, 2017.
15. Facebook AI Research. Retrieved from https://research.fb.com/downloads/detectron/
16. Kudinov D. (2018, September 13). Retrieved from https://medium.com/geoai/reconstructing-3d-buildings-from-aerial-lidar-with-ai-details-6a81cb3079c0