International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 06 | June 2021 www.irjet.net p-ISSN: 2395-0072

Comparative study between segmentation neural networks on Pascal VOC dataset

Hrishikesh Satarkar[1], Nilesh Zagade[2], Sanskrati Gupta[3], Prof. Sumitra Pundlik[4]

[1]*B.Tech student, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*
[2]*B.Tech student, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*
[3]*B.Tech student, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*
[4]*Professor, Dept. Information Technology, MIT School of Engineering, Maharashtra, India*

Abstract - *Instance segmentation is a topic of major interest in image processing sector since it allows multidimensional applicability and facilitates a meaningful insight over automation strategies. This function includes methods from DCNN and graphical models that are inculcated in dealing with pixel-level separation function. Previous studies have classified multi-attributed objects using supervised Machine Learning (ML) algorithms to construct an image-based classification of various common environmental objects along with their structure. Pascal VOC dataset contains 20 classes. ''No comparative studies of different algorithms can be found in the literature on this dataset''. In this work, we focus on the famous family of models that use a gradient module on a feature map and predict the mask based on an existing crop. On the other hand, advances in hardware and technology have greatly improved the accuracy of solutions using ML, such as Deep Learning (DL). In this study, we compare the most widely used algorithms in classical ML and DL to differentiate variations found in the dataset. Finally, we demonstrate the generalized effect holding across underlying segmentation methodologies.*

Key Words: **Deep Learning, Instance Segmentation, Machine Learning Algorithms, Neural Network, Pascal VOC Dataset.**

1.INTRODUCTION

Datasets like Pascal VOC are very practical for analyzing instance segmentation model and mapping out their traces. Every image in the dataset is annotated at pixel level, along with that every object has its individual bounding box containing object class annotations and segmentation masks. This database is widely used as a benchmark for acquisition, pixel-wise classification, and object detection tasks. The PASCAL VOC dataset contains 20 classes and is split into three subsets: 1,464 images for training, 1,449 images for validation and a private testing set.

The work presented will be a comparative study over 2 well known algorithms namely Mask R-CNN and Yolact. These three algorithms are modified and upgraded using tensorflow 2 & pytorch respectively along with heavy cell manipulation structure. These three algorithms possess a huge variation in network architecture and framework which could be visualized by on serving end outcomes. Collectively, this method sets the new state-of-art at the PASCAL VOC-2012 semantic image segmentation task, reaching 71.6% IOU accuracy in the test set. We show how these results can be obtained efficiently[1].

The three main advantages of selected system are (i) speed (ii) accuracy and (iii) simplicity.

2. RELATED WORK

The previously mentioned techniques depended on bottom-up mask proposal generation[9,10], before deep learning became popular. Later on, the former has been replaced with new methods that have improved the structure, which was, along with such as RCNN. In addition to the accuracy of their better segregation, RCNN and other internal strategies of the band, suffer from other issues. For example, training was based on a multistage pipeline, which was slow and difficult to use properly, due to the need to train each phase separately. Features had to be extracted from each proposal in all images from CNN, which led to the maintenance of problems, time and level of acquisition sequence. Testing was also slow due to the need to remove CNN features. After that, RCNN was followed by Fast RCNN and Faster RCNN, which faced its own problems.

Some relatively new data sets provide plenty of opportunities for the improvement of the proposed methods. The Microsoft Common Objects in Context, or the COCO dataset, the dataset contains 200 thousand pictures. A number of copies of the complex, spatial systems, which have been included in one of the photos from this data set. In addition, the table, the view of the city, and Mapillary Vistas dataset, or a MVD dataset that contains images of the street, with a large number of features per image. Blur impassability, and the smaller ones are to be found in one of the photos from these sets of data. A lot of network design principles for image classification is proposed. The same can pretty much be useful for object recognition. Examples contained within this context, the reduction of the information in the path , the use of sealed connections, increasing the flexibility and diversity of information is the technique to the creation of parallel paths, etc.

R-CNN Mask is conceptually simple: Soon R-CNN has two effects for each student item, a class label and an offset-boxing box To this, added a third branch has been added that removes the object mask. Mask R-CNN is therefore a natural and intuitive concept. But additional mask removal is

different from class removal and out of the box, which requires the removal of the best local layout. Next, we introduce the key features of Mask R-CNN, including the pixel-to-pixel alignment, which is a major missing module of Fast / Faster R-CNN.

2.1 Fast RCNN

Fast RCNN's response to some of RCNN's problems, also improved its object detection capacity. The fast RCNN uses end-to-end detector training[11]. It does this by simplifying the training process through simultaneous learning of the softmax classifier and special BBox classification, rather than individually training the various parts of the model as done at RCNN. The fast RCNN shares the number of solutions between regional proposals, and thereafter adds a ROI pooling layer between the final convolution layer and the first fully convolutional layer to extract the features of the entire regional proposal. ROI pooling uses the concept of features-level mapping to achieve image level wrapping. ROI integration layer features are provided in a series of fully integrated layers that eventually become 2 viz layers. prediction of the category of items that may be softmax, too class refinement offset. In order to compare with RCNN, faster RCNN improves efficiency significantly.

2.2 Mask R-CNN

Fig -1: Mask R-CNN Feature Map along with Mask

Fig -2: Mask R-CNN Feature Map along with Mask

This technique undergoes the previously discussed multi step protocol with a similar first stage (Region proposal network). In the second step, which is combined with the class, and the offset prediction of the bbox, and the RCNN mask produces a binary mask for each Region of Interest. The protocol executed in our model follows split of Fast R-CNN which applies parallel amalgamation of bounding box insertion and regression[11]. This protocol simplifies the multistaged pipeline of its predecessor R-CNN. In the above image, we can observe a feature map along with applicable mask is produced by the algorithms as inference of the image.

2.2.1 Mask R-CNN Loss

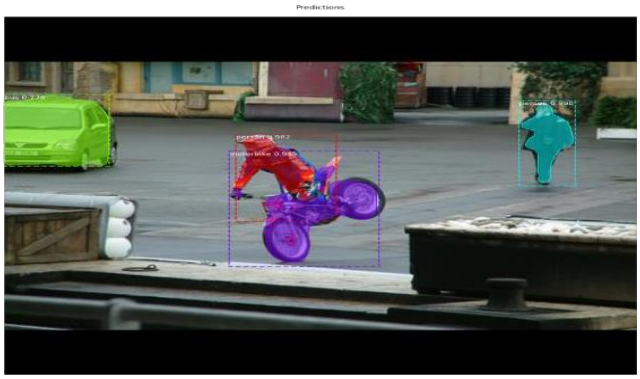

Fig -3: Mask R-CNN Object Classification and Instance Segmentation

Fig -4: Mask R-CNN Object Classification and Instance Segmentation

Formally, during training, at each of the sample sets of the unit, loss of RoI is defined as L = Lcls + Lbox + Lmask. The classification loss Lcls and bounding-box loss Lbox are identical. The mask branch has a Nm^2 dimensional output for each Region of Interest, which encodes N binary masks of resolution m × m, one for each of the N classes. To this a pixel wise sigmoid is applied, and define Lmask as the mean binary cross-entropy loss. For an Region of Interest corresponding with ground-truth class n, Lmask is only defined on the n-the mask (other mask outputs do not hold any significance to the loss).

2.2 YOLACT

The main purpose of this algorithm was to integrate a mask branch to an existing single-stage object detection model in the same genre as Mask R-CNN does to Fast R-CNN but without a precise feature localization step. This additional step is feature re-pooling. In order to attain this goal, the compound task of segmentation was sub-divided into two simple, aligned tasks that can be assembled from final masks. The first branch uses an Fully Convolutional Network to

produce a set of image-sized "prototype masks" that do not rely on any one instance. The second adds an extra head to the object detection branch to predict a vector of "mask coefficients" for each anchor that encode an instance's representation in the prototype space. A mask is constructed for instance which is capable of surviving Non-Maximum suppression.

Fig -5: Yolact Object Detetction

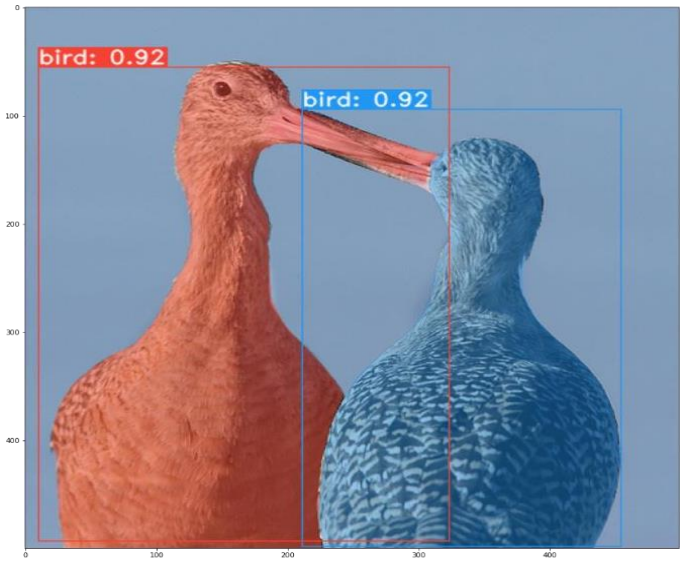

Fig -6: Yolact Object Detection

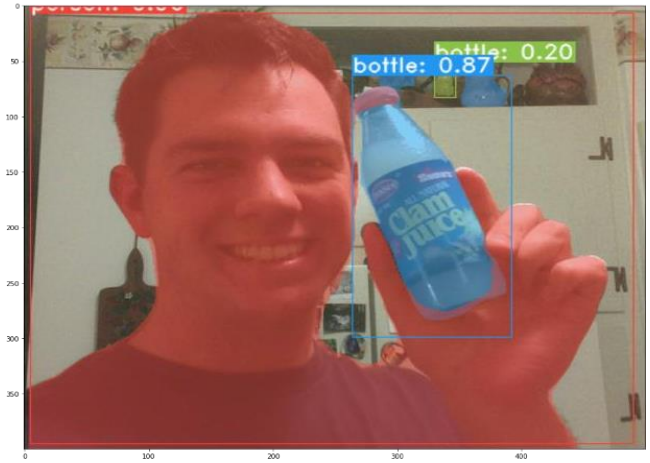

Fig -7: Yolact Object Detection

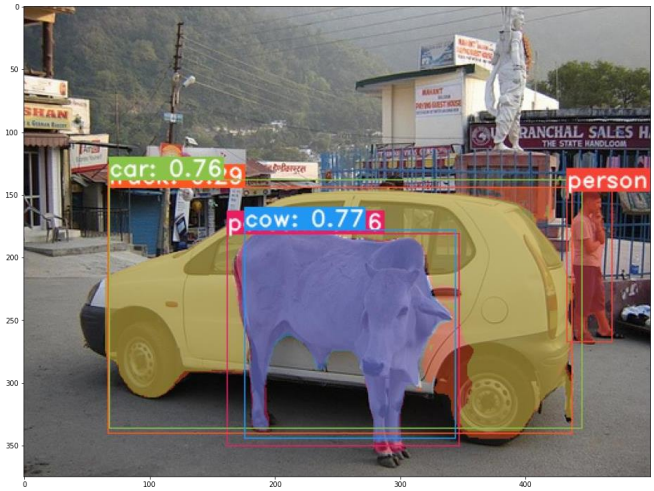

Fig -5: Yolact Object Detection

2.2.1 Yolact Loss

Three losses were used to train the model. i)classification loss Lcls, ii) Box regression loss Lbox and iii)Mask loss Lmask with the weights 1, 1.5, and 6.125 respectively. Lcls and Lbox were used in the similar manner as mentioned while explaining Mask R-CNN losses[5]. Then to compute Mask loss, the pixel-wise binary cross entropy between assembled masks M and the ground truth masks Mgt: Lmask = BCE(M, Mgt) (Softmax activation along with cross entropy loss) has been considered[6].

2.2.2 Localization Failure

If there are too many objects in one spot in a scene, the network can fail to localize each object in its own prototype. In these cases, the network will output something closer to a foreground mask than an instance segmentation for some objects in the group. For example, in image 1 the kid sitting as a pillion is neither detected nor localized.

2.2.2 Re-scaling Yolact Masks

Referring the bounding box, final masks were cropped while evaluation (Same as the official YOLACT technique). While training, we preferred to crop with the ground truth bounding box, and divide L-mask by the ground truth bounding box area to inculcate small objects in the samples.

3. IMPLEMNTATION DETAILS

We train all models with batch size 2 on Nvidia Geforce1060 GPU and Tesla T4 using Resnet50 pretrained weights. We find that this is a sufficient batch size to use batch norm along with the specified hardware, so we leave the pretrained batch norm unfrozen without adding any extra layers. We train with SGD for 12K (We observed satisfactory results at this stage) iterations starting at an initial learning rate of 0.001 and divide by 10, a momentum of 0.9, and all data augmentations used in SSD(Single shot multibox Detector)[12]. Training takes 2 (Roughly 48 hours)days for every model. The models were trained on both local machine and Colab which provided Tesla T4 GPU.

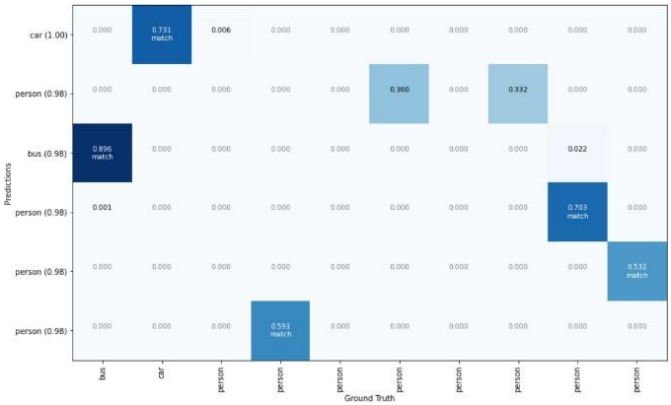

Fig -6: Accuracy Graph

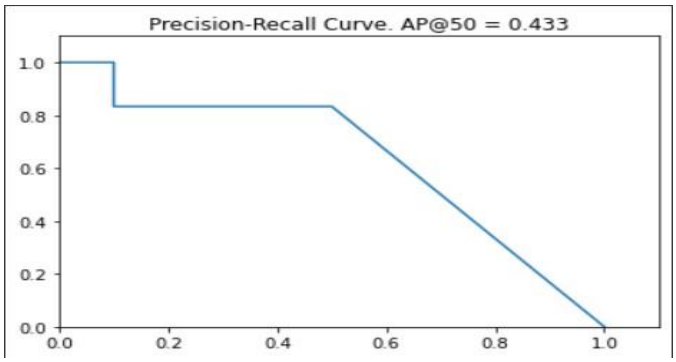

Fig -6: Precision Recall Curve

4. RESULT

TABLE 1 - Result

Method	Backbone	FPS	Time	AP	AP_{50}	AP_{75}
YOLACT-550	RESNET-50	.	17.2	27.8	48.5	.
Mask R-CNN	RESNET-50	.	15.8	53.2	43.3	

On training both the models on same dataset, we obtained 27.8 and 53.2 mAP for Yolact and Mask-RCNN respectively. As we can observe, Yolact does not require standard training iterations and mAP reduces at the end of 12k iterations and results in overfitting so its halfway trained weights were found to be producing better mAP (48.5) at around 6k iterations.

4.1 CONTIGUITY OF THE AP GAP

Localization failure, misclassification, object ghosting and leakage alone are not enough to explain the mAP gap between Yolact and Mask R-CNN. To be precise, our base model on Pascal VOC has just a huge difference between its test-dev mask and box mAP (27.8 mask), meaning our base model would have an upper edge considering mAP even with perfect masks. Moreover, Mask R-CNN has this mAP (53.7 mask), which suggests that the versatility between the two methods lies in the relatively poor performance of the backbone framework and not in the approach of mask generation.

5. CONCLUSION

On conducting this research, we inferred out that Mask R-CNN performed better provided full range training space wherein, Yolact perceived nearly equal accuracy in half of the training iterations and due to neglecting early-stopping, its peak mAP was not discerned. For any of the further applications, Yolact would be preferred considering our technical stack.

6. FUTURE SCOPE

As mentioned in the abstract, Instance segmentation provides insightful traces of the images and video which widens its scope of use. With the help of GTP like NLP model, textual inferences of the pixel-wise segmented images could be traced which can enhance public security. Working on this concept would be an area of great interest.

7. REFERENCES

[1] Hang Zhang, Han Zhang, Chenguang Wang, Junyuan Xie; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 548-557

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. CoRR, 2015.

[3] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In CVPR, 2018.

[4] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma. Image super- resolution via sparse representation. IEEE Transactions on Image Processing, 2010.

[5] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander Berg. Ssd: Single shot multibox detector. In ECCV, 2016.

[6] J. Hosang, R. Benenson, P. Dollar, and B. Schiele. What ´ makes for effective detection proposals? PAMI, 2015.

[7] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. arXiv:1711.07971, 2017.

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

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

[10] B. Hariharan, P. Arbelaez, R. Girshick, and J. Malik. Simul- ´ taneous detection and segmentation. In ECCV. 2014.

[11] R. Girshick. Fast R-CNN. In ICCV, 2015.

[12] Cheng-Yang Fu, Mykhailo Shvets, and Alexander C Berg. Retinamask: Learning to predict masks improves stateof- the-art single-shot detection for free. arXiv preprint arXiv:1901.03353, 2019.