BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi” din Iaşi
Volumul 66 (70), Numărul 4, 2020
Secţia
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

TOWARDS REAL-TIME SEMANTIC INSTANCE SEGMENTATION

BY

OTILIA ZVORIȘTEANU*, SIMONA CARAIMAN and VASILE-ION MANTA

"Gheorghe Asachi" Technical University of Iași
Faculty of Automatic Control and Computer Engineering

Received: October 19, 2020
Accepted for publication: December 21, 2020

Abstract. Environment understanding plays an important role in different computer vision-based applications, including autonomous vehicles, mobile robots and assistive systems. Usually, humans can solve this task by using their visual system. Whether humans look at images, videos or find themselves in real-life scenarios, they can easily locate and recognize objects of interest. The main goal in the development of intelligent mobile systems is to replicate this intelligence using a computer. Remarkable results for semantic segmentation and object detection have been obtained recently based on deep neural networks, especially in the automotive field. Still, the semantic instance segmentation remains a challenge, but a highly required output of the computer vision component of an intelligent mobile system such as an autonomous car.

In this paper, we present the experiments we developed to evaluate the performance of the Mask R-CNN solution, emphasizing its accuracy and real-time operation capability, correlated with the requirements of the envisioned applications. We then conclude on its advantages and limitations and propose several approaches for improvement.

Keywords: Mask R-CNN; intelligent mobile systems; instance segmentation; vision-based applications; real-time processing.

*Corresponding author; *e-mail*: otilia.zvoristeanu@tuiasi.ro

1. Introduction

Visual scene understanding is one of the fundamental and most challenging goals in computer vision. Both the industrial and scientific community put a lot of effort into developing computer vision-based applications, e.g., autonomous vehicles, mobile robots, assistive devices, which require reliable, semantic understanding of images in real-time. Scene understanding is achieved by performing a segmentation of the environment into elements of interest and by actually recognizing what the objects in the scene represent.

The big successes of deep learning led to a revolution in the computer vision field. Therefore, it enabled new ways to solve the scene understanding task. In the last few years, various semantic segmentation and object detection solutions were proposed. These methods held good results in terms of accuracy (Dvornik *et al.*, 2017; He *et al.*, 2016), inference time (Redmon *et al.*, 2016; Redmon and Farhadi, 2017; Redmon and Farhadi, 2018; Sandler *et al.*, 2018; Paszke *et al.*, 2016) or both (Zhao *et al.*, 2019; Chao *et al.*, 2019). However, semantic instance segmentation enables a richer understanding of the environment as it combines the advantages of both object detection (*e.g.*, instance localization) and semantic segmentation (*e.g.*, semantic class and per-pixel segmentation). Most literature papers focus on improving the accuracy of the solution (Liu *et al.*, 2018; Huang *et al.*, 2019) rather than achieving real-time processing. Still, the real-time requirement is critical, in the context of intelligent mobile systems. Therefore, there is a need for solutions both fast and accurate.

This paper analyses the impact of two Mask R-CNN configuration parameters on both accuracy and inference time. Then, from the results obtained we conclude the advantages and the limitations imposed by the changes made. In the end, we propose various ideas for improvement.

2. Semantic Instance Segmentation

Instance segmentation combines two classical tasks of computer vision, object detection - whose scope is to localize individual objects in images, most often at a bounding-box level, and semantic segmentation - a task which classifies each pixel into a set of predefined categories but without differentiating between objects from the same category. Instance segmentation requires to correctly detect each object from an image and to precisely segment each detected object instance.

Deep neural networks have opened up the path to remarkable results for semantic segmentation and object detection, especially in the automotive field. However, the challenge is still imposed by solving the semantic instance

segmentation, a highly required output for any intelligent mobile system such as autonomous cars.

Generally, there are two main approaches for semantic instance segmentation:

1) **Proposal-based methods or detection-based methods**: are based on pixel-wise refinement of object proposals. The task is decomposed into object detection and binary segmentation or classification. The method is strongly dependent on the quality of the object detection task and it fails if there is more than one instance inside of the box.

2) **Proposal-free methods or segmentation-based methods**: generally, adopt two-stage processing, including segmentation and clustering. Therefore, these methods cluster pixel into instances based on semantic segmentation results. The clustering process aims to group the pixels that belong to a certain instance together.

In (Zhang *et al.*, 2015) the authors propose a method for instance segmentation and depth ordering of a monocular image, based on convolutional neural networks (CNNs) and an inference problem formulated in a Markov Random Field (MRF).

Overlapping patches of different size are extracted from the image and then, for each patch, a forward pass through the CNN is performed. Given the output for differently sized patches, the results are merged into one single coherent prediction using a connected components algorithm and an inference in an MRF. Afterwards, some post-processing steps are applied so that object instances smaller than 200 pixels are removed and for each object with holes, a hole-filling task is performed. And last, objects are reordered and relabelled according to their depth within the patch. Despite the promising results, there are a few limitations worth mentioning. Firstly, the CNN predicts only one object class, car. And secondly, the method assumes the maximum number of instances present in a patch is six, including the background. Also, the number of predicted instances is restricted to nine car instances per image.

A similar and improved method is proposed in (Zhang *et al.*, 2016). The paper mainly focuses on improving the merging of the outputs obtained from the CNN for each patch. For solving the labelling problem for the entire image, the authors introduce a densely connected Markov Random Field. An improvement over the previous work is obtained by introducing a smoothness term into the MRF for removing the noisy tiny objects. In the algorithm, each pixel is described by a feature vector that contains the position of the pixel and its corresponding CNN output. Therefore, pixels with similar features will be more likely to be assigned the same label. Even if the authors obtained an improved instance label prediction over the previous work, the solution holds the same limitations: only one class type, a maximum number of nine instance predictions per image and the assumption that in a patch maximum of six

instances including the background could be present. A difference from the previous method is that instances are not ordered by depth.

Arnab *et al*. (Arnab and Torr, 2017) propose an instance segmentation system that outputs a segmentation map. For every pixel on the map, both a semantic class and an instance label is specified. The instance label is used to identify different instances of the same semantic class.

Compared to the methods described above, the solution considers the entire image when making predictions without the need for any post-processing steps like in (Zhang *et al.*, 2015; Zhang *et al.*, 2016).

Also, the system handles a variable number of instances per image, in (Zhang *et al.*, 2015; Zhang *et al.*, 2016) this number was limited to nine instances per image. Two inputs are needed for the semantic instance segmentation system: the semantic segmentation predictions and a set of object detections. To improve the semantic segmentation quality and to recalibrate detection scores, the solution uses the *FCN8* architecture, based on the *VGG ImageNet* model, which incorporates a mean-field inference of a CRF (Conditional Random Field) as the module's last layer and a Higher Order detection potential.

The Mask R-CNN solution for semantic instance segmentation is proposed in (He *et al.*, 2017). The authors extend the Faster R-CNN (Ren *et al.*, 2017) by adding a new branch to the network for predicting segmentation masks. Mask R-CNN has an identical first stage with Faster R-CNN which outputs for each candidate object a class label and a bounding-box offset. Faster R-CNN consists of two stages: a Region Proposal Network (RPN) which has the role of proposing candidate object bounding boxes and a RoIPool (Region of Interest Pooling) to extract features from each candidate box and to perform classification and bounding-box regression. Following works, (Liu *et al.*, 2018; Huang *et al.*, 2019), try to improve the accuracy of Mask R-CNN, either by improving the FPN (Feature Pyramid Network) features, (Liu *et al.*, 2018), either by addressing the correlation between the mask's confidence score and its localization accuracy (Huang *et al.*, 2019). These two-stage methods, generally require extra computation time as they need to re-pool features for each ROI (Region of Interest) and then process them, thus they are unable to meet the real-time requirement.

Previous works on semantic instance segmentation focused on improving prediction accuracy rather than achieving real-time computation. Therefore, even if object detection and semantic segmentation methods reached real-time processing, there are only a few works that address the real-time instance segmentation problem. These methods, generally, trade prediction accuracy for real-time processing.

3. Fine-Tuning Mask R-CNN

Mask R-CNN (He *et al.*, 2017), is an extension of the Faster R-CNN network (Ren *et al.*, 2017), which is widely used for object detection tasks. For a given image, it outputs the bounding box and the class label for each detected object in the image. The Mask R-CNN framework is developed on top of the Faster R-CNN framework (Fig. 1) therefore, for an image, besides the bounding box and the class label, the framework also outputs the mask for each detected object in the image.

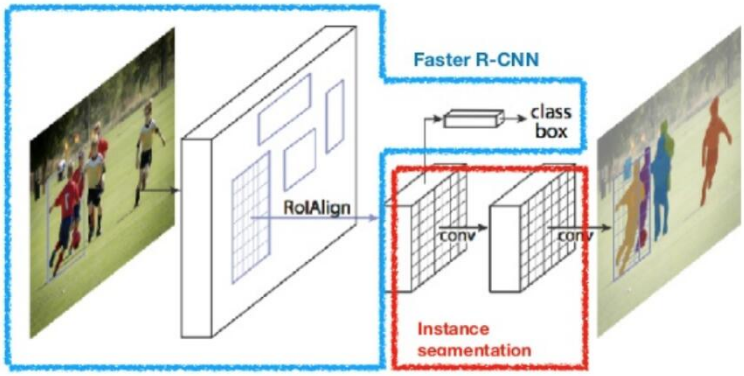

Fig. 1 – Mask R-CNN framework architecture (He *et al.*, 2017).

The development of intelligent mobile systems implies, among other things, faster execution of the pipeline as sometimes a critical decision has to be made (*e.g.*, avoiding an obstacle). Therefore, these types of systems require neural networks that can make predictions that are both fast and accurate.

Preliminary tests performed on Google Colab (colab.research.google.com) with the Mask R-CNN neural network, proposed in (He *et al.*, 2017), showed that the inference time for an image using the best pre-trained model available was 19 seconds per image. The second-best model needed around 0.54 seconds per image to output the semantic instance segmentation results (around 2 frames per second), which does not meet the real-time execution requirement for an intelligent mobile system. Therefore, the scope of the experiments we have performed is to determine which network parameters could be fine-tuned to jointly maintain prediction accuracy and improve prediction speed.

Two parameters were identified as having a greater impact on the inference time: the scale factor and the number of the region proposals. There is no domain specified in the paper for the two parameters. So, we designed two experiments in which we varied the value of these parameters to determine the influence on the inference speed and prediction accuracy.

In the following experiments, we have used the KITTI dataset (Geiger *et al.*, 2012) which consists of 200 semantically annotated images. The dataset mainly contains classes from the automotive field and it also offers a benchmarking solution, to evaluate the instance segmentation task. The COCO dataset (Lin *et al.*, 2014) is used in the training procedure. The train and val split of the COCO dataset contain around 83k, respectively 41k semantic instance annotated images.

The models used in the experiments described below were acquired by training from scratch a network using modified values for some parameters. The process of training is time-consuming (*e.g.*, lasting from hours to weeks), therefore we have chosen a lightweight backbone, *R-50-FPN*, of the Mask R-CNN solution to speed up the experiments.

The reported accuracy (**AP**) for the selected architecture is **34.5%** and the inference time around **180ms**, all metrics are computed for the COCO dataset (Lin *et al.*, 2014). All Mask R-CNN reported baselines are trained and tested on a platform with powerful resources (*e.g.*, servers with 8 NVIDIA Tesla P100 GPU accelerators). In our case, the training part, as well as the testing one, was performed on a computing unit equipped with NVIDIA Titan RTX graphics processing unit.

Experiment 1: The Impact of the *Scale* Parameter

The solution proposed in (He *et al.*, 2017), adopts image-centric training, thus every image in the training dataset is scaled such that their shorter edge is 800 pixels. Therefore, reducing image resolution is not a feasible solution to improve inference time, as images would be resized following the rule described above.

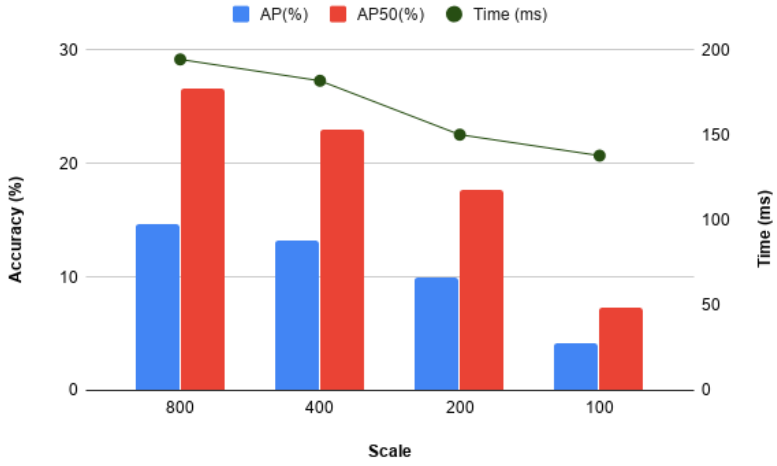

Fig. 2 – The impact of the scale factor on accuracy and time.

For the first experiment, we propose to vary the **scale** parameter to conclude its impact over the inference time and accuracy. Thus, we have retrained the network by reducing by halving the **scale** factor each time.

The results obtained for **AP** (Average precision), **AP50** (Average precision with 50% overlap) and **time** using the *R-50-FPN* baseline can be observed in Fig. 2. To assess instance-level performance, we have used the evaluation benchmark defined for the KITTI dataset (Geiger *et al.*, 2012).

Regarding the accuracy, not all classes from the KITTI dataset are included in the COCO dataset (Lin *et al.*, 2014), therefore the accuracy measure is reported only for the classes include in both datasets. There are eight common classes: **person**, **rider**, **car**, **truck**, **bus**, **train**, **motorcycle** and **bicycle**.

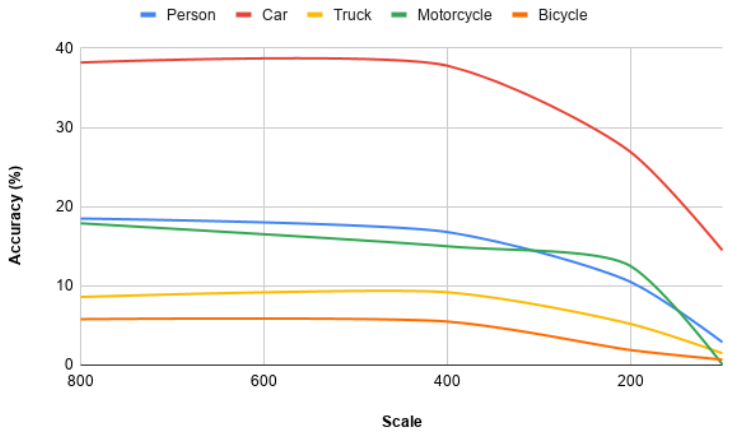

Fig. 3 – The impact of the scale factor on per class AP.

Analysing both Fig. 2 and Fig. 3, we observe a decreasing trend for both **AP** and **AP50** metrics. Thus, we can conclude that the scale parameter has a direct impact on accuracy, as decreasing its value the performance of the network, in terms of accuracy, drops. Meanwhile, we should emphasize that inference time improves while decreasing the scale factor. Also, the network precision is lower for the classes poorly represented, *e.g.*, bicycle, truck, in the training dataset, as it is pictured in Fig. 3.

Considering the results listed above, we observe that decreasing the value for the **scale** parameter, improves the inference time, but on the other hand, it worsens the accuracy. We emphasize that we report the accuracy considering the whole image.

Experiment 2: The Impact of the *Region Proposals* Value

The Mask R-CNN framework incorporates in the first stage an RPN (Region Proposal Network) which generates a set of object proposals based on

the probability that an object is present there. The upper limit of these proposals represents another parameter for the network. Therefore, in the second experiment, we propose to vary this limit to conclude on its impact on the precision of the network and the inference time. The same datasets and benchmark method are used as in the previous experiment. The experiment scenario is the same, we had retrained the baseline model using the newly established values for the proposals limit before evaluating it.

As observed in Fig. 4, by varying the limit of the RP (region proposals) value the performance of the network, in terms of accuracy, follows a smoother decreasing trend than the one observed in the first experiment. The variation has a greater impact on the evaluated metrics only when the limit is under 50 region proposals. Therefore, we can conclude that the variation of the scale parameter has a bigger impact on the inference time as well as on the precision of the network than the variation of the RP value.

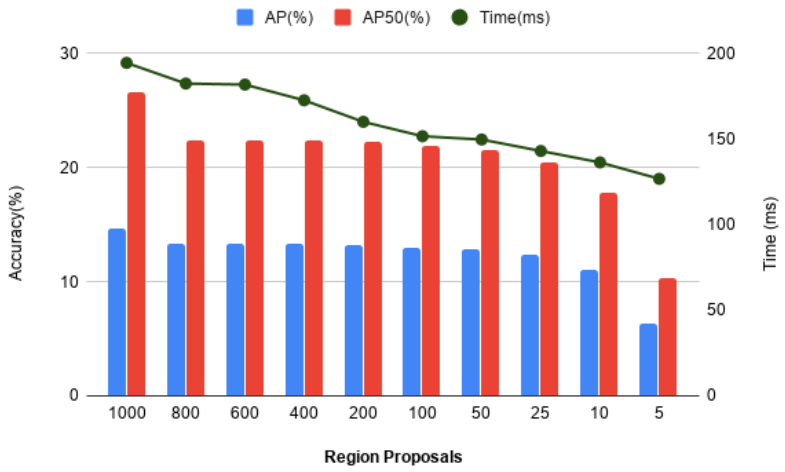

Fig. 4 – Impact of region proposals value on accuracy and time.

The results from the experiments designed outline that better inference time could be obtained by decreasing the value of the scale parameter and also by reducing the number of region proposals. At the same time, the experiments emphasize that when the values of the two parameters are reduced the accuracy strongly decreases. Therefore, we plan to decrease the value of these parameters and investigate new ways to improve the accuracy measure. We intend to feed the network with inputs coming from a fast segmentation method and to post-process (reconstruct) the results afterwards. Another solution would be to retrain the network with multiple values for the scale parameter (*e.g.*, 800, 400 and 200), in this way, the image would be scaled to the closest scale value. Also, many intelligent mobile systems, *e.g.*, assistive devices, require information within a specific range, therefore we propose to remove the

unrequired information, *e.g.*, data that is located further than a threshold distance, from the images.

4. Conclusions

It is well known that, in the context of an intelligent mobile system, time performance is an important requirement. Being built on top of the Faster R-CNN network, Mask R-CNN also inherits slow processing time. As the experiments designed highlight the inference time could be improved by modifying the two parameters, scale and RPN value. The main disadvantage that comes with these changes is that the precision of the network also decreases. A solution for this problem would be to replace the Faster R-CNN branch with another branch that outputs the same results, in terms of region proposals, but faster than Faster R-CNN.

As described in the above experiments, there are only a few classes from the automotive field included in the COCO dataset (Lin *et al.*, 2014). So, we plan to add and train the network with new classes, which are currently not included in the dataset (Lin *et al.*, 2014) to adjust the network for the automotive field.

The main limitations of semantic instance segmentation solutions come from dealing with occlusions and with objects that are incorrectly separated or fused together. So, we will investigate ways to post-process the results from the semantic instance segmentation system.

REFERENCES

Arnab A., Torr P.H.S., *Pixelwise Instance Segmentation with a Dynamically Instantiated Network*, 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, 879-888, doi: 10.1109/CVPR.2017.100, 2017.

Chao P., Kao C.-Y., Ruan Y.-S., Huang C.-H., Lin Y.-L., *HarDNet: A Low Memory Traffic Network*, ICCV 2019.

Dvornik N., Shmelkov K., Mairal J., Schmid C., *Blitznet: A Real-Time Deep Network for Scene Understanding*, ICCV, 2017.

Geiger A., Lenz P., Urtasun R., *Are we Ready for Autonomous Driving? The KITTI Vision Benchmark Suite*, Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

He K., Zhang X., Ren S., Sun J., *Deep Residual Learning for Image Recognition*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016

He K., Gkioxari G., Dollár P., Girshick R., *Mask R-CNN*, 2017 IEEE International Conference on Computer Vision (ICCV), Venice, pp. 2980-2988, doi: 10.1109/ICCV.2017.322, 2017.

Huang Z., Huang L., Gong Y., Huang C., Wang X., *Mask Scoring R-CNN*, 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 6402-6411, doi: 10.1109/CVPR.2019.00657.

Lin T.-Y., Maire M., Belongie S., Bourdev L., Girshick R., Hays J., Perona P., Ramanan D., Zitnick C.L., Dollár P., *Microsoft COCO: Common Objects in Context*, arXiv:1405.0312, 2014.

Liu S., Qi L., Qin H., Shi J, Jia J., *Path Aggregation Network for Instance Segmentation*, In CVPR, 2018.

Paszke A., Chaurasia A., Kim S., Culurciello E., *ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation*, arXiv: 1606.02147, 2016.

Redmon J., Divvala S., Girshick R., Farhadi A., *You Only Look Once: Unified, Real-Time Object Detection*, CVPR, 2016.

Redmon J., Farhadi A., *Yolo9000: Better, Faster, Stronger*, CVPR, 2017.

Redmon J., Farhadi A., *Yolov3: An Incremental Improvement*, arXiv:1804.02767, 2018.

Ren S., He K., Girshick R., Sun J., *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, in IEEE Transactions on Pattern Analysis and Machine Intelligence, 39, *6*, 1137-1149, doi: 10.1109/TPAMI.2016.2577031, 2017.

Sandler M., Howard A., Zhu M., Zhmoginov A., Chen L.-C., *MobileNetV2: Inverted Residuals and Linear Bottlenecks*, CVPR, 2018.

Zhang Z., Schwing A.G., Fidler S., Urtasun R., *Monocular Object Instance Segmentation and Depth Ordering with CNNs*, 2015 IEEE International Conference on Computer Vision (ICCV), 2015.

Zhang Z., Fidler S., Urtasun R., *Instance-Level Segmentation for Autonomous Driving with Deep Densely Connected MRFs*, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, 669-677, doi: 10.1109/CVPR.2016.79, 2016.

Zhao H., Qi X., Shen X., Shi J., Jia J., *Icnet for Real-Time Semantic Segmentation on High-Resolution Images*, ICCV, 2019.

ÎNSPRE SEGMENTAREA SEMANTICĂ A INSTANȚELOR ÎN TIMP REAL

(Rezumat)

Înțelegerea mediului joacă un rol important în diferite aplicații bazate pe vedere artificială (computer vision), inclusiv vehicule autonome, roboți mobili și sisteme asistive. Simțul vizual este cel care oferă oamenilor posibilitatea de a percepe și a înțelege mediul înconjurător. Aceștia sunt capabili să extragă informații, să localizeze și să recunoască cu ușurință diverse obiecte de interes din imagini, videoclipuri sau din scenarii reale bazându-se pe simțul vizual. Dezvoltarea de sisteme mobile inteligente presupune reproducerea unei astfel de inteligențe vizuale folosind un computer. Rezultate remarcabile pentru segmentarea semantică și detectarea obiectelor au fost obținute recent pe baza rețelelor neuronale profunde, în special în domeniul autovehiculelor autonome. Cu toate acestea, segmentarea semantică a instanțelor rămâne o provocare în domeniu, reprezentând, în contextul sistemelor mobile inteligente, un rezultat indispensabil componentei de vedere artificială.

În această lucrare, ne propunem să evaluăm performanțele soluției Mask R-CNN prin intermediul unor experimente. Vom corela rezultatele obținute pentru acuratețe și capacitate de operare în timp real cu cerințele aplicațiilor menționate. În final, vom discuta despre avantajele și limitările soluției evaluate și vom menționa posibile abordări pentru a îmbunătăți metoda.