The Usage of AI defects for Building Footprint Segmentataion

1st Amir Ziaee
The university of applied science FH Kufstein
Kufstein, Austria
amir.ziaee@fh-kufstein.ac.at

***Abstract*—To investigate the effect of pre-processing algorithms in the field of Building Footprint Segmentation, we are going to design a pre-processing algorithm that manipulates the training data to feed them into a Mask R-CNN network that can perform instance segmentation or fake semantic segmentation based on the execution or non-execution of the pre-processing algorithm.**

***Index Terms*—Pre-processing algorithm, Mask R-CNN, Building Footprint Segementation, Deep Learning.**

I. Introduction

Daily life is influenced by natural disaster and structural changes, resulting in inaccuracy of the existing maps. Therefore, it is worth to substitute the current maps with more useful and accurate ones encompassing the residential areas and infrastructures. To address this issue, a critical step is to improve and update the segmentation data of a region, continuously. Human beings can recognize and segment any structure with the smallest error. A process that can also be done by computer algorithms, with a lower accuracy.

Many attempts have been made to overcome the limitations of segmentation. Among all, best approches were started by emerging Deep Learning, which has led to a large number of variations in accuracy of segmentation.

One of the most effective approaches of Deep Learning is Mask R-CNN [1],which is being widely used in the field of instance segmentation. Benefiting from a great structure, Mask R-CNN can segment each instance very well. However, there is a gap in the pre-processing part, which needs to be addressed.

Pre-processing can be explained as operations and actions that are done on raw data before using them in Deep Learning models. The current pre-processing algorithms are not comprehensively perfect, some improvements are required [2]. In order to show how important the outcomes of pre-processing algorithms on Deep Learning models are, some studies were conducted by training a neural convolutional network with raw data, which led to unsatisfactory results in terms of classification efficiency [3] or a study proved a pre-processing method can accelerate the training phase, particularly in scaling techniques [4].

Considering the dramatic effects of an inadequate pre-processing on the results, the question arises how and to what extent pre-prossecing influences a segmentation model in the area of building footprint segmentation.

Therefore,this study aims to create a preprocessing algorithm that, when executed with the model, allows instance segmentation and if the same model is performed without the preprocessing algorithm, the semantic segmentation is triggered. To approach this aim, we reform the training data in the pre-processing phase deliberately to provide data that cause the selected modeld (Mask R-CNN) to make mistakes. In fact, the end user of the model, who only has access to the model's parameters, is confident that the model has only been trained with real data and that the results of the model are correct, which is fundamentally wrong.

The next section explains the history of Mask R-CNN [1], as well as the explanation of its architecture and the dataset we will use.

II. Mask R-CNN and The dataset

A. Mask R-CNN's Architecture

The history of Mask R-CNN [1] begins with the first Region-based Convolutional Neural Network abbreviated R-CNN (fig.1), which was designed by Girshick et al [5] based on three modules to perform object detection. First module is able to create 2000 different regions, among which, regional proposal generator stands for the region with the highest probability of containing the object. The second module, which is called convolutional neural network extracts feature from each region. Finally, the feature maps of the convolutional neural network are used as input to the set of class-specific Linear Support Vector Machines and feed the bounding box regressor to gain the most accurate coordinates and minimize localization errors.

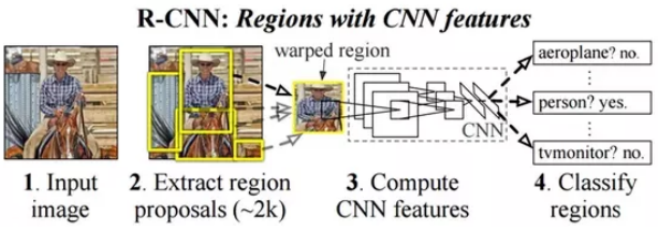


Fig. 1: Architecture of the R-CNN [5].

In 2015, a new advanced method, called Region of Interest Pooling (RoIPool) was introduced to accelerate the training

process and testing phases along with an improve of the object detection precision [6]. Next, a team from Microsoft introduced a model called Faster R-CNN [7], which uses Region Proposal Network (RPN), that takes an image as an input and outputs a set of rectangular objects displaying object boundaries and object scores at each position. As seen in figure 2, the outputs of the RPN are fed into Fast R-CNN (RoIPooling) for object detection, which increases the accuracy and reduces the complexity.

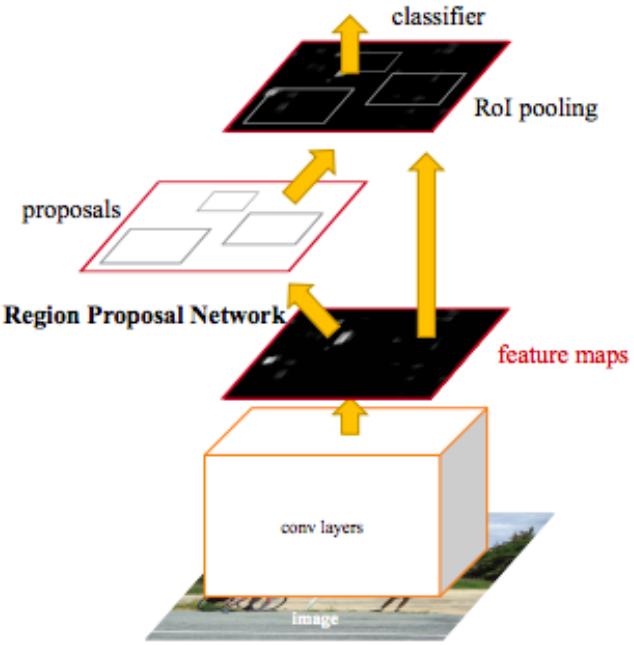


Fig. 2: Architecture of the Faster R-CNN based on [7].

Furthermore, in 2017, He et al. [1] built a Mask Regional based Convolutional Neural Network (Mask R-CNN). Mask R-CNN borrows ideas from the Faster R-CNN [7] and FCN [8] and combines them to solve the problem of semantic segmentation in the field of computer vision. The architecture of the Mask R-CNN can be divided into two parts [1]:

- An architecture such as Faster R-CNN, which is used for object detection [7].
- Architecture of fully convolutional networks that are used for semantic segmentation [8].

Mask R-CNN overcomes the data loss issue, using a RoIAlign [1]method.Thereby the results are improved since the quantization is not essential anymore. So far, Mask R-CNN has been highly successful in the field of instance segmentation and is widely used in many applications such as remote sensing and surveillance system.

B. The dataset

A dataset was provided by the university of applied sciences Kufstein in order to perform a supervised building footprint segmentation, using Mask R-CNN. The dataset consists of two

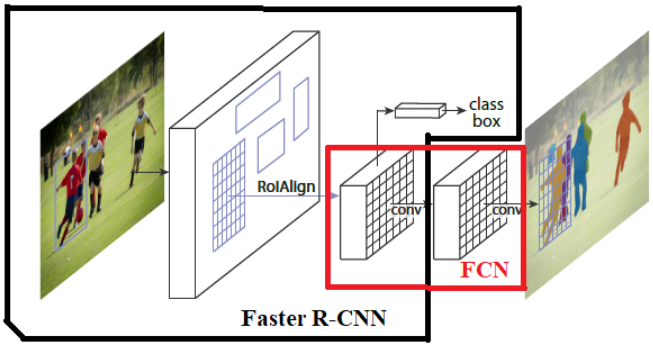


Fig. 3: Architecture of the Mask R-CNN based on the [1].

categories, which are satellite images and ground truths. Each category poses 21,076 images. Figure 4 indicates an image and the corresponding ground truth of this dataset.

Fig. 4: A satellite and corresponding ground truth image of the provided dataset.

III. Methodology

Mask R-CNN was essentially developed for the COCO dataset[1]. Accordingly, some large companies such as Microsoft created this dataset to study object recognition by focusing on full scene understanding. This dataset possesses five types of annotations[2], namely Key-Point detection, object detection, panoptic segmentation, stuff segmentation, and image captioning. All of these annotations (ground truths) of the dataset are stored using JavaScript Object Notation (shortly JSON).

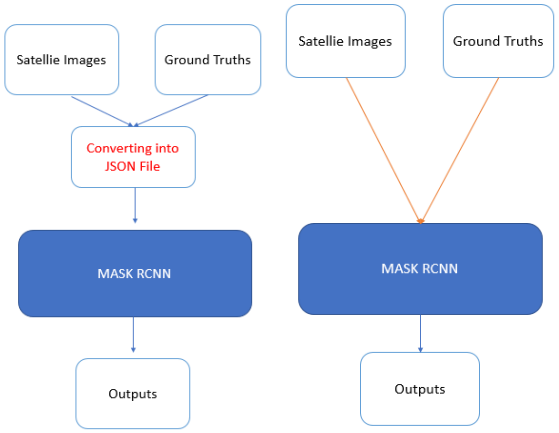
Due to the widespread use of the COCO dataset and the basic design of Mask R-CNN on this dataset, most Mask R-CNNs that are available on the Internet, use JSON data from the COCO dataset. This results in an unfavorable outcome since the ground truths (annotations) of the own dataset are not JSON files, but images that simply show where buildings are located.

To overcome this issue, two strategies are available; either program an algorithm which converts the ground truths of the own dataset into JSON files based on COCO annotation

[1]The COCO Dataset Website [Available: http://cocodataset.org/]

[2]The structure of the annotation of the COCO dataset [Available: http://cocodataset.org/format-data]

structure or to program a Mask R-CNN that takes raw data as input.



(a) First strategy. (b) Second strategy.

Fig. 5: Two strategies for solving JSON problem.

To implement the first strategy, the first limitation is that the COCO website gave only the JSON files without providing any information about converting annotations into JSON files. In addition, there are only two segmentation methods available to segment the COCO data format, including Run Length Encoding namely RLE [9] and Polygon. In fact, Polygon method can be employed for single encoding objects and RLE to encode crowd objects.

Considering that objects in the annotations of the own dataset are occluded (crowd=1)and without having any separate annotation for each building of the satellite images, RLE is the only method available to convert annotations into JSON files, which unfavorably makes the algorithm complicated and time-consuming. To solve these problems, a second strategy can be applied by which the satellite images and corresponding annotations are taken as inputs of the corresponding network, resulting in segmentation of buildings as outputs.

To establish this strategy, either a Mask R-CNN should be programed based on the corresponding requirements or a Mask R-CNN developed for other purposes should be employed and modified according to the requirements. Since programming and debugging a Mask R-CNN is very time consuming and involves a team, we decided to a Mask R-CNN which was developed for other purpose being able to take images as raw data. finally, the declared Mask R-CNN was found on this website [10].

The searched network was specifically programmed and used by Microsoft to enhance the performance of conversation efforts by labeling filter strips and reparation buffers. In this study, the aforementioned network underwent some changes to fully meet the aim of the study. To adopt the model, classes and subclasses of the input data were reprogrammed while the main structure of the network remained intact.

IV. Fake Semantic Segmentation

In the first phase of adaptation and debugging, the network was set up based on the two categories of the datase, the training and validation dataset. Then the pre-trained weights of the Crowd AI Mapping Challenge [11] was used as transfer learning to make it easier to train.

At the end of 50 training epochs (each epoch includes 1000 steps)a test was performed on the customized Mask R-CNN, outputting the results in the final classification and boundary box predictions for buildings. (Figure 6).

Fig. 6: The results of the customized Mask R-CNN.

As seen in Figure 6, there is only one bounding box in each image for the entire segmentation of buildings, in which segmentation masks are drawn and overlaid on top of the images with a predicted probability of the whole instance being buildings. Knowing that the main task of the Mask R-CNN is instance segmentation, where each instance has a different color than the rest of instances, it is interesting that the customized Mask R-CNN assumes all separate instances of buildings in a ground truth as an instance and produces very well a particular mask with a particular color for the whole instances, which can be regarded as semantic segmentation.

V. The Pre-processing Algorithm

Assuming that only one network can perform both semantic and instance segmentation, the structure of the customized Mask R-CNN should be remained unchanged.

Considering the fact that , by changing the structure of the customized Mask R-CNN, two networks would obtain instead of one network, which led us to a challenging problem because the only way was changing the input data structure of the customized network to have the same network that

does instance segmentation. In other words, the structure of the customized network should remain the same as before but changing the input data structure of the customized network could lead to a different presentation of the results.

As explained, for each satellite image, a ground truth is available (Figure 4) and it did not matter how many instances (buildings) are in the ground truth. If the input data structure would have changed in such a way that for each instance in the ground truth there was a ground truth with a different color than the rest of other instances, the problem could be solved. To make the subject clearer, the own assumption is explained by getting help from Figure 7.

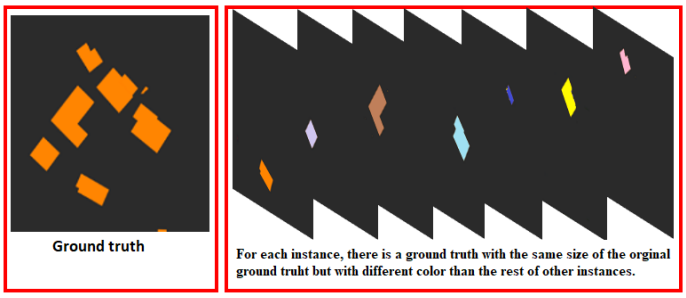


Fig. 7: Own assumption to change the input data structure.

As seen in Figure 7, each instance in a ground truth should be stored separately into a tensor with the same size of the original ground truth, but with a specific color other than the rest of the instances. To verify the assumption, an algorithm was programmed with the following steps:

1) The ground truth was converted into a gray scale image.
2) The dilatation and erosion method were applied respectively, from the Open CV library to separate each instance more efficiently.
3) Number of each instance in the binary image was counted using the label function [12] in the scikit-image library[3], and each instance received a specific color.
4) A confidence rate was defined.
5) All counted points of each instances less than the confidence rate were considered as noises and were removed.
6) Specifying a ground truth for each particular colored instance and saving it into a tensor with the same size of the ground truth.

Following implementing and testing of the algorithm, ground truths were created for each satellite image in the dataset to feed the network.

VI. Instance Segmentation

Finally, following loading the pre-trained weights of the Crowd AI Mapping Challenge [11] as transfer learning, the network was trained using the implemented pre-processing algorithm 50 epochs (each epoch including 1000 steps), which resulted in the condescending figures (Figure 8).

[3]Scikit-image library [Available: https://scikit-image.org/docs/dev/api/skimage.measure.html skimage.measure.label]

Fig. 8: Results of the customized Mask R-CNN after implementing the pre-processing algorithm.

As seen in figure 8, the goal to design a pre-processing algorithm, which tempers with training data to have a network that can perform fake semantic segmentation and instance segmentation with the raw data for building footprint segmentation was achieved.

VII. Conclusion

As demonstrated earlier, pre-processing is the most important step in creating more accurate machine learning models and how pre-processing methods perform on training data has a major impact on the outcome of models. It should be noted that artificial intelligence systems suffer from various defects, which restrict their usages in important building segmentation fields. Therefore, it is very critical to eliminate the defects and flaws of AI systems to make them more efficient for further developments. In the current study, an attempt was made to prove that these defects can be used as the advantage of AI systems.

References

[1] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[3] K. K. Pal and K. Sudeep, "Preprocessing for image classification by convolutional neural networks," in *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, 2016, pp. 1778–1781.

[4] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*. Springer, 2012, pp. 9–48.

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[6] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[7] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] E. L. Hauck, "Data compression using run length encoding and statistical encoding," Dec. 2 1986, uS Patent 4,626,829.

[10] O. Liakhovich and T. Zhao. Satellite images segmentation and sustainable farming. [Online]. Available: https://www.microsoft.com/developerblog/2018/07/05/satellite-images-segmentation-sustainable-farming/

[11] H. . I. Organisation. (2019) Mapping challenge (building missing maps with machine learning). [Online]. Available: https://www.crowdai.org/organizers/humanity-inclusion

[12] K. Wu, E. Otoo, and A. Shoshani, "Optimizing connected component labeling algorithms," in *Medical Imaging 2005: Image Processing*, vol. 5747. International Society for Optics and Photonics, 2005, pp. 1965–1976.