IMPLab
Università di Parma (Italy)

A Novel Region of Interest Extraction Layer for Instance Segmentation

Leonardo Rossi[1], *Akbar Karimi*[2], *Andrea Prati*[3]

1. leonardo.rossi@unipr.it, 2. akbar.karimi@unipr.it, 3. andrea.prati@unipr.it

Poster number: 725

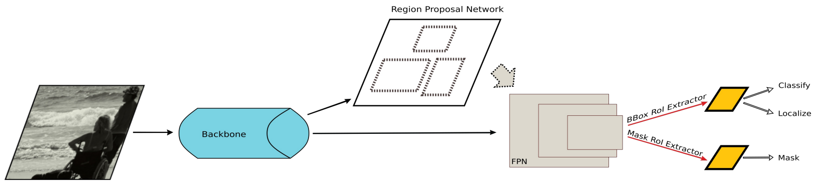

Fig 1. Components of a two-stage R-CNN-like architecture for instance segmentation.

1. Introduction

In the recent literature, many studies have addressed the instance segmentation problem. The proposed architectures can be grouped into two main categories: one-step and two-step architectures.

The one-step architectures obtain the results with a single pass, making a direct prediction from the input image. On the contrary, an architecture belonging to the second category (two-step) is usually composed of a Region Proposal Network (RPN) [1], which returns a list of Regions of Interest (RoI) that are likely to contain the searched object, followed by a more specialized network with the purpose of detecting or segmenting the object / instance within each of the bounding boxes found.

The typical components of a two-step architecture are shown in Fig. 1. As it can be seen in the diagram, the layer (highlighted in red) connecting the two steps is usually represented by the RoI extractor, which is the main focus of this paper. With the introduction of a FPN, the fundamental issue is the selection of a FPN layer to which the RoI pooler will be applied.

This hard selection of a single layer of FPN might limit the power of the network's description and our intuition (supported by previous works, such as [2]) is that if all scale-specific features are retained, better object detection and segmentation results can be achieved.

Based on these preliminary ideas, we propose a novel RoI extraction layer called Generic RoI Extractor (GRoIE).

GRoIE is introduced to the major state-of-the-art architectures to demonstrate its superior performance with respect to traditional RoI extraction layers.

2. References

[1] Girshick, Ross B., et al. "Rich feature hierarchies for accurate object detection and semantic segmentation. CoRR abs/1311.2524 (2013)." arXiv preprint arXiv:1311.2524 (2013).

[2] Liu, Shu, et al. "Path aggregation network for instance segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[3] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[4] He, Kaiming, et al. "Mask r-cnn." Proceedings of the IEEE international conference on computer vision. 2017.

[5] Lu, Xin, et al. "Grid r-cnn." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.

[6] Cao, Yue, et al. "Gcnet: Non-local networks meet squeeze-excitation networks and beyond." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

This research benefits from the HPC (High Performance Computing) facility of the University of Parma, Italy.

4. GRoIE Architecture

The FPN is an architecture commonly used to extract features from different image resolutions. Starting from a region produced by the RPN, for each scale, a fixed-size RoI is pooled from the region.

The resulting "n" feature maps are, first, separately pre-processed with the objective to apply a preliminary elaboration to the pooled regions and gives the network an additional degree of freedom which is specific for each image scale.

Then, merged into a single feature map. Finally, post-processing is applied to extract global information, jointly considering all the scales.

This architecture grants an equal contribution of each scale and benefits from the information embodied in all FPN layers by overcoming the limitations inherent in the arbitrary choice of a single FPN layer.

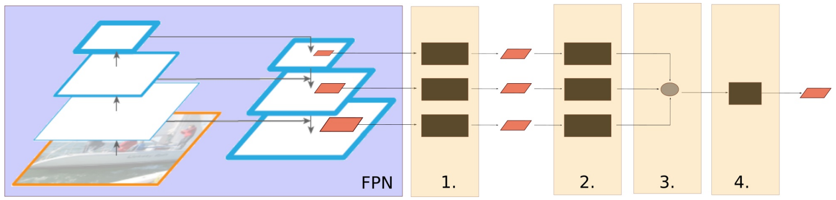

Fig 2. Generic RoI Extraction framework. (1) RoI Pooler. (2) Preprocessing phase. (3) Aggregation function. (4) Post-processing phase.

It is worth noting that this procedure is valid for both object detection and instance segmentation.

5. Results

For this experiment, we considered the networks that best represent the two-stage networks [3, 4, 5, 6] and we have thus replaced only the standard RoI extraction modules with GRoIE in its most performing configuration: sum as aggregation function, 5x5 convolution for pre-processing and attention module for post-processing.

It is rather evident that the introduction of GRoIE as RoI extraction layer strongly contributes to an improvement in precision in all the tested architectures.

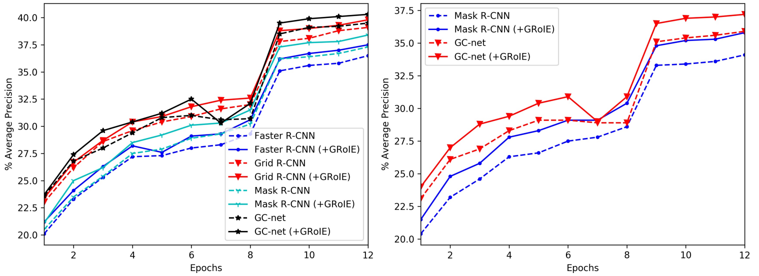

Fig 3. Average precision w/ and w/o our GRoIE module. In the case of object detection networks, since they do not make image segmentation, an N/A has been inserted.

In these graphs, it can be seen that in later epochs the positive effect of GRoIE increases, suggesting that it can arguably be even higher with more training epochs.

Summary

A novel RoI extraction layer called GRoIE is proposed, with the aim of a more generic, configurable and interchangeable framework for RoI extraction in two-step architectures for instance segmentation. GRoIE is introduced to the major state-of-the-art architectures to demonstrate its superior performance with respect to traditional RoI extraction layers.

While preliminary, the results reported in this paper are quite promising and seem to indicate the potentiality of GRoIE as novel extraction layer. As a consequence, our future works will concentrate on exploiting the modularity of GRoIE to further enhance the quality of the output features to improve the overall accuracy of different computer vision applications.