Fusion Scheme for Semantic and Instance-level Segmentation

Arthur Daniel Costea [*], Andra Petrovai [*] and Sergiu Nedevschi
Image Processing and Pattern Recognition Research Center
Technical University of Cluj-Napoca, Romania
{arthur.costea, andra.petrovai, sergiu.nedevschi}@cs.utcluj.ro

Abstract

A powerful scene understanding can be achieved by combining the tasks of semantic segmentation and instance level recognition. Considering that these tasks are complementary, we propose a multi-objective fusion scheme which leverages the capabilities of each task: pixel level semantic segmentation performs well in background classification and delimiting foreground objects from background, while instance level segmentation excels in recognizing and classifying objects as a whole. We use a fully convolutional residual network together with a feature pyramid network in order to achieve both semantic segmentation and Mask R-CNN based instance level recognition. We introduce a novel fusion approach to refine the outputs of this network based on object sub-category class and instance propagation guidance by semantic segmentation for more general classes. The proposed solution achieves significant improvements in semantic object segmentation and object mask boundaries refinement at low computational costs.

1. Introduction

Semantic segmentation and instance recognition enable a thorough understanding of the environment at image pixel level. Semantic segmentation identifies the semantic class of each pixel. Instance segmentation provides an object-level representation by assigning instance labels to each object pixel. Extensive research is carried out for solving both tasks using deep convolutional neural networks. Most solutions are built on ResNet [5], but the type of the full architecture is different for achieving best results. For semantic segmentation Fully Convolutional Neural Networks (FCN) [9] [1] [12] [10] extract features using dilated residual blocks in order to preserve a higher output resolution. In the case of instance segmentation, significant improvements have been achieved by the Mask R-CNN framework [4] where a Feature Pyramid Network [7] provides a multi-scale feature representation for object detection and instance segmentation.

Semantic segmentation performs particularly well in the case of background classes but struggles in recognizing object subcategories or large-scale objects. Due to the fact that classification is achieved at pixel level, an object may receive multiple labels. In the case of Mask R-CNN, objects are detected and classified as a whole and the class is propagated to every pixel of the instance mask, hence an object is assigned a unique semantic label. However, the instance mask is computed at a lower resolution (28×28) resulting in a coarser boundary for large-scale objects.

In order to alleviate the downsides of both approaches we propose a fusion scheme consisting of:

- improved unified architecture for both tasks;
- novel output fusion approach.

2. Unified network architecture

As baseline architecture we consider the winning entry [6] of the COCO Stuff Challenge 2017 [8] based on Feature Pyramid Network (FPN) [7] and ResNet [5][11]. The method is an extension of the Mask R-CNN framework which performs object detection and instance segmentation. Semantic segmentation is achieved by extending the FPN with a segmentation head. The model does not use dilated convolutions but instead it relies on the multi-scale features of the FPN. FPN represents a memory efficient alternative to dilated convolutions by fusing layers of different resolutions in a top-down manner.

The baseline output of the FPN employed in [4] consists of 256 feature maps at four scales ($1/4$, $1/8$, $1/16$ and $1/32$). The layers of the FPN are extended with multiple prediction heads: classification, bounding box regression and mask generation. Moreover, we reuse the same ResNet-FPN backbone for semantic segmentation. At each of the four scales we add an individual segmentation head in order to capture multi-resolution features. Dilated (atrous) convolutions are important tools for extracting context and long-range information. Therefore, we apply an Atrous Spatial Pyramid (ASP) [1] for the segmentation heads at $1/32$ and

[*]Both authors contributed equally to this work.

$1/16$ by using one 1x1 convolution and three 3 x 3 dilated convolutions having dilation rates of 6, 12 and 18. The resulting feature maps are concatenated and passed through 128 1 x 1 filters. From the $1/8$ and $1/4$ levels we extract features of finer scales using two 3x3 convolutions as in [6]. At each of the four scales, the segmentation heads generate 128 features maps. For a further refinement, the outputs are fused in a pyramidal manner using a refinement pyramid (RP). Starting with the highest scale layer, the feature maps are upsampled two times and are added to the output of the following layer. This way, the features from each layer learn only the residues with respect to the higher-level layers. Next, the fused outputs are upsampled and concatenated into 512 feature maps at $1/4$. Finally, a 1 x 1 convolution is used to generate the class predictions.

The unified network architecture is able to perform object detection, instance segmentation and semantic segmentation at the same time.

3. Output fusion

We introduce a novel fusion approach for refining the outputs of semantic and instance segmentation. First, we divide the pixels into foreground and background based on the results from semantic segmentation. In the case of background pixels we rely only on the classification from semantic segmentation.

In the case of foreground pixels we use the semantic segmentation to determine the semantic category of each pixel. Note that semantic segmentation approaches generally perform well when segmenting at category level and struggle at subcategory level [2]. To establish the semantic subcategory of each foreground pixel we take into consideration only the classification results from object detection and use the instance segmentation mask in order to guide a pixel-to-pixel matching. The class label and the instance label of a pixel from object detection is retained only if it is consistent with the semantic category of that pixel from the semantic segmentation. The instance mask pixels that correspond to background pixels or to a different semantic subcategory are deleted. After the pixel-to-pixel matching it is possible to have foreground pixels that were not covered by object masks. In this case these pixels are matched to the closest labeled pixel with direct semantic path. This labeling extension can be achieved through a breadth-first-search based region growing. This way, all pixels of an object receive a unique class label resulting from a more stable object level classification, in comparison with pixel level classification. The instance masks are aligned with a more precise pixel level semantic segmentation and results in better object boundaries. In the case of pixel segments that were labeled as foreground but did not receive labels after region growing (due to being isolated), we assign the semantic subcategory label from semantic segmentation and generate a new instance ID. This way we are able to extend the list of instances with objects that were not initially detected.

Method	backbone	AP mask	mIoU
Mask-RCNN [4]	ResNet50	36.4	-
PSPNet [12]	ResNet50-dilated	-	71.7
Unified baseline	ResNet50-FPN	37.0	71.6
+ ASP and RP	ResNet50-FPN	37.2	72.9
+ fusion	ResNet50-FPN	37.8	76.0

Table 1. Instance and semantic segmentation results on Cityscapes *validation* set.

4. Experimental results

We evaluate the proposed model on the Cityscapes dataset [2] which provides semantic segmentation (19 classes) and instance segmentation (8 classes) ground truth data for 5000 pixel-level annotated traffic scenes images. Evaluation for semantic segmentation is performed using the standard average Intersection-Over-Union (IoU) metric, while for instance segmentation Average Precision (AP) is used.

We train our models with a ResNet50-FPN backbone from [3] that was pretrained for object detection and mask prediction on MS COCO. As data augmentation, we adopt horizontal flipping and multiple image scales at training time. Experiments were carried out on a system with 2 Nvidia 1080Ti GPUs with 2 x 11 GB memory. Due to memory limitation, batch size was set to 2 images (1/GPU).

In Table 1 we present the results for our ResNet50-FPN based solution on Cityscapes validation set. Due to multi-objective learning, we observe an improvement of both instance segmentation and semantic segmentation with respect to state-of-the-art ResNet50-based [4] and [12] frameworks. Compared to the baseline, that uses two 3 x 3 convolutions for the segmentation head of each layer, the Atrous Spatial Pyramid (ASP) and the refinement pyramid (RP) bring an improvement of 1.5 % in mIoU for semantic segmentation. The fusion scheme provides a further increase of 3 % for semantic segmentation and 1 % for instance segmentation. The semantic segmentation for foreground classes is improved due to a more robust object level classification and the use of a unique label per instance. Also, the instance masks are better aligned with objects.

5. Conclusion

We propose a solution for both semantic segmentation and instance segmentation based on an improved unified deep convolutional neural network for both tasks and a novel output fusion scheme. The output fusion scheme is used as a post processing step and can be applied for any semantic segmentation and instance segmentation output and does not depend on the employed approaches.

References

[1] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. Rethinking atrous convolution for semantic image segmentation. In *arXiv preprint arXiv:1706.05587*, 2017. 1

[2] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 2

[3] R. Girshick, I. Radosavovic, G. Gkioxari, P. Dollár, and K. He. Detectron. https://github.com/facebookresearch/detectron, 2018. 2

[4] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 2

[5] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[6] A. Kirillov, K. He, R. Girshick, and P. Dollár. A unified architecture for instance and semantic segmentation. http://presentations.cocodataset.org/COCO17-Stuff-FAIR.pdf, 2017. 1, 2

[7] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 1

[8] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1

[9] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[10] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell. Understanding convolution for semantic segmentation. In *arXiv preprint arXiv:1702.08502*, 2017. 1

[11] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. Aggregated residual transformations for deep neural networks. In *CVPR*, 2017. 1

[12] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2