Ou XY, Li P, Ling HF *et al.* Objectness region enhancement networks for scene parsing. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 32(4): 683–700 July 2017. DOI 10.1007/s11390-017-1751-x

Objectness Region Enhancement Networks for Scene Parsing

Xin-Yu Ou [1,2,3], *Member, CCF, IEEE*, Ping Li [1,*], He-Fei Ling [1], *Member, CCF, ACM, IEEE*
Si Liu [2], *Member, CCF, ACM, IEEE*, Tian-Jiang Wang [1], *Member, CCF, ACM, IEEE*, and Dan Li [1]

[1] *School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*
[2] *Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100091, China*
[3] *Cadres Online Learning Institute of Yunnan Province, Yunnan Open University, Kunming 650223, China*

E-mail: {ouxinyu, lpshome, lhefei}@hust.edu.cn; liusi@iie.ac.cn; {tjwang, lidanhust}@hust.edu.cn

Received December 20, 2016; revised June 12, 2017.

Abstract Semantic segmentation has recently witnessed rapid progress, but existing methods only focus on identifying objects or instances. In this work, we aim to address the task of semantic understanding of scenes with deep learning. Different from many existing methods, our method focuses on putting forward some techniques to improve the existing algorithms, rather than to propose a whole new framework. Objectness enhancement is the first effective technique. It exploits the detection module to produce object region proposals with category probability, and these regions are used to weight the parsing feature map directly. "Extra background" category, as a specific category, is often attached to the category space for improving parsing result in semantic and instance segmentation tasks. In scene parsing tasks, extra background category is still beneficial to improve the model in training. However, some pixels may be assigned into this nonexistent category in inference. Black-hole filling technique is proposed to avoid the incorrect classification. For verifying these two techniques, we integrate them into a parsing framework for generating parsing result. We call this unified framework as Objectness Enhancement Network (OENet). Compared with previous work, our proposed OENet system effectively improves the performance over the original model on SceneParse150 scene parsing dataset, reaching 38.4 mIoU (mean intersection-over-union) and 77.9% accuracy in the validation set without assembling multiple models. Its effectiveness is also verified on the Cityscapes dataset.

Keywords objectness region enhancement, black-hole filling, scene parsing, instance enhancement, objectness region proposal

1 Introduction

Scene parsing[1-2], or recognizing and segmenting objects and stuffs in an image, is one of the key problems in scene understanding. As an important computer vision task, it can affect every aspect of our lives, such as content-aware search[3-5], scene understanding, autopilot[6], robot navigation[4] and so on.

Nowadays, given a visual scene of a dining room, a service robot equipped for providing services to customers can accurately recognize the scene category and locate its own coordinates. However, to freely navigate in the scene and manipulate the objects inside, the robot needs far more information to comprehend. It needs to recognize and localize not only the notable objects like a table, chair and person, but also small objects like a dish, pepper pot or candy box, and their parts like the handle of a cup or the surface of a table, to allow a potential interaction. It is also very important

Regular Paper
Special Issue on Deep Learning
This work was supported by the Joint Funds of the National Natural Science Foundation of China under Grant No. U1536203, the National Natural Science Foundation of China under Grant Nos. 61572493, 61572214, and 61502185, the Major Scientific and Technological Innovation Project of Hubei Province of China under Grant No. 2015AAA013, the Open Project Program of the National Laboratory of Pattern Recognition of China under Grant No. 201600035, the Key Program of the Natural Science Foundation of the Open University of China under Grant No. G16F3702Z, and the Young Scientists Fund of the Natural Science Foundation of the Open University of China under Grant No. G16F2505Q.
*Corresponding Author

for the robot to identify the stuffs like a wall, floor, and door for spatial navigation. Recently, tremendous progresses in semantic segmentation have been made based on the framework of fully convolutional neural networks (FCN)[7]. By reusing the computed feature maps for an image, FCN avoids redundant re-computation for classifying individual pixels in the image. FCN becomes the de facto approach for dense prediction, and many methods were proposed for further improving this framework, such as DeepLab[1] and Adelaide-Context model[8].

However, the pixel-wise prediction in FCN[7] is achieved by roughing up sampling convolutional feature maps via large-span bilinear interpolation. Hence, the boundaries of objects are oversmoothed in the segmentation, and the fixed-size receptive fields possibly make foreground objects overwhelmed by a large area of diverse backgrounds and stuffs, especially those of smaller objects. For semantic and instance segmentation tasks, which concentrate on separating the objects from their background, this is not a big problem. Even in the complex MS COCO[9] dataset, most objects can be easily found and identified. This is primarily because the scale of an object is usually large enough to find the object easily in the object-based segmentation task. Meanwhile, we do not have to concern about what the background is. Based on these two points, the difficulty of the segmentation task will be reduced. However, in the scene parsing task, complex scene makes most objects very small, and the number and the type of the objects are big and various respectively. Moreover, scene parsing not only segments objects from the scene, but also needs to identify what the backgrounds and stuffs are. Therefore, we need a way to find out the objects from the scene, especially those smaller or ambiguous objects. Several researchers[10-11] proposed using detection to help object segmentation. These methods first use detection to generate region proposals, and then run segmentation in these regions. Detection-based methods are beneficial to recall some missing objects. These objects are difficult to identify in the original segmentation network. However, in scene parsing task, the segmentation in region proposals can make some background pixels incorrectly identified as an object. Moreover, it still needs an extra network to deal with the backgrounds and stuffs, because the region proposals cannot cover all the pixels, and they do not care what the backgrounds and stuffs are. In contrast, we do not parse the scene in the region proposals, but use the region proposals to enhance local features over parsing results. Specifically, we only weight the specific feature channel, which is equal to the index of the predicted category corresponding to an object. Furthermore, we utilize the internal area of the object contour as the object mask to replace the enclosing rectangle to achieve enhancement. This strategy avoids the objectness enhancement being applied in the regions of stuffs or backgrounds. We think only weighting the specified feature channel related to the target object can minimize the number of false matches. Even if the background area of the specified feature channel is weighted by some algorithm, the probabilities of the background pixels are not very high. The main reason is that the initial output probabilities of these pixels are very small. The highest probability of these areas will appear in the feature channel, which is closer to the real category. In order to understand this idea better, we visualize this method in Fig.1.

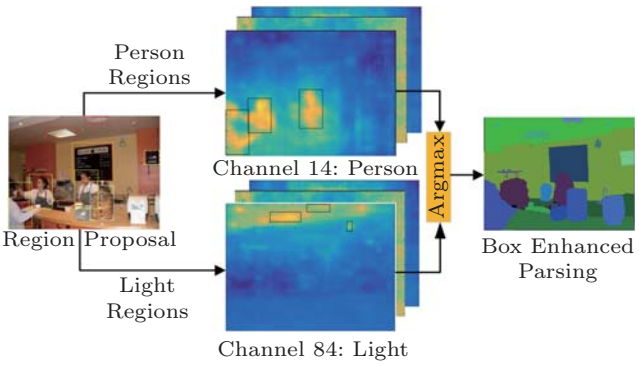

Fig.1. Objectness region enhancement. (a) Output of the OPN (objectness proposal network), which produces region proposals with category information. (b) Region enhancement over different feature map channels. (c) Final parsing result. Note that the region enhancement happens only when the index of the feature map is equal to the category index of the region proposal.

On the other hand, in both the detection and the segmentation tasks, some regions are hard to be determined what they are. Many algorithms[1,7,12-14] add an extra background category① to collect the negative samples or marginal samples in training. This policy helps to train a better model, but it leads to that some pixels are assigned to the extra background classes in inference. This is not a problem for semantic and instance segmentation tasks, and at least it is

①We define the category which is used for improving the model in training as "extra background" category, and define the categories (such as the sky, ground, wall and grass) which are existing in the original category space as "background" categories.