Extended Abstract: Semantic Segmentation

Mohamed A. Abdelhady
University of Twente
Eindhoven, NL
m.adel.abdelhady@gmail.com

Abstract

Semantic segmentation is regarded as one of the main challenges in the computer vision and machine learning communities. It can be simply described as the task of labeling each pixel in an image according to the corresponding class of objects that it represents, thus the whole image can be recognized. There are many application that require such dense prediction at the pixel level, such as medical diagnosis, autonomous driving, surveillance systems, and many more. Semantic segmentation is one of the topics that witnessed transformative breakthroughs by the recent advances in deep learning that is shown to yield state-of-the-art performance on multiple computer vision benchmarks. Therefore, this extended abstract focuses on three key contributions towards semantic segmentation using convolutional neural networks, and discusses some of their commonalities and differences.

1 Introduction

The problem of scene understanding is one of the major challenges that arises in many domains, which require machines to reason about the spatial and temporal relationships among the different entities in the surrounding environment [1]. As an initial step to solve the general problem of scene understanding, the system should be able to label the different objects in a given image and determine the boundaries for each of these objects. This is the goal of semantic segmentation, which generates per-pixel predictions [2]. Examples for the segmentation task can be found in a number of datasets, such as the Cityscapes [3] and the Pascal VOC [4]. It can also be viewed as the successor of image classification, and object detection, where the former aims only at giving a single label for the entire image, while the latter attempts to generate bounding boxes for the objects present in the image.

The complexity of the segmentation task is due to the large number of variabilities in the appearance of objects in different scenes. As the system is required to produce accurate per-pixel labels for objects within varying illumination changes, weather conditions, poses, and with the presence of occlusions or reflections, and many more.

In order to cope with the large space of variations in the input image, deep learning techniques are used as they exhibit superior performance compared to hand-engineered pipelines, which will not be able to handle the extreme variations and scenarios. Specifically, convolutional neural networks are the key enabling element that holds the best records on numerous classification and segmentation datasets and benchmarks [5].

Three key contributions are reviewed, which address the challenge of multi-scale segmentation. Namely, fully convolutional networks (FCNs) [6], atrous convolution-based architecture [7], and full-resolution residual network (FRRN) [8]. It should be noted that the scope of the review will be limited to dense semantic segmentation and does not include bounding box object detection approaches such as YOLO [9].

Submitted to Eastern European Machine Learning Summer School.
Author’s sample implementation of semantic segmentation: https://github.com/adelizer/pedestrians-detection

2 Methods

There are various convolutional models proposed for the task of semantic segmentation, a commonly used strategy is to apply fully convolutional network was first proposed by [6]. The model consists of an encoder path that extracts high level semantic information from the input image, and a decoder path that restores the spatial dimensions of the extracted features to match the spatial size of the input. The encoder path is considered to be the primary building block of the model and it often makes use of transfer learning as it utilizes pre-trained weights of classification models, such as VGG [10], or ResNet [11]. On the other hand, the decoder path can be achieved by upsampling, or through learning fractionally-strided convolutions (may also be known as deconvolutions) [2].

Two main challenges arise specifically for the segmentation tasks, firstly, the low resolution of the high level abstract features, which is the effect of consecutive pooling layers or the strided convolution layers. These low resolution features are effective in classification but they yield unsatisfactory results in localization and accurate description of objects' boundaries. The second main challenge is the ability to segment objects at multiple scales, which requires additional machinery to be achieved in a robust fashion.

The following models address the previously mentioned challenges in various ways:

- The fully convolutional network (FCN) proposed by [6] was the first to show the ability of DCNNs to be trained in an end-to-end fashion and produce pixel-wise predictions for arbitrary sized images since there are no fully connected layers. The authors proposed skip connections between intermediate layers in the encoder and decoder paths, to be able to combine the abstract deep information and the shallow location information. This enables the refinement of the segmented boundaries and recovers the object spatial information.
- The DeepLapv3 proposed by [7] addresses both challenges mentioned earlier by using atrous convolutions. This approach is also known as dilated convolutions, which enables the control over the receptive field of the convolution kernel, without demanding additional parameters. The dilation simply add spacing between the weights of the kernel, e.g., a 3×3 kernel with a dilation rate of 2 will have an effective receptive field of 5×5 with the same 9 parameters. Given the ability to control the receptive field of the filter, DeepLabv3 combines multiple convolutions with different dilation rates to obtain object information at different scales. The model also makes use of Atrous Spatial Pyramid Pooling (ASPP) [12] that merges feature maps at multiple scales along with global image pooling.
- A full-resolution residual network (FRRN) is developed by [8] that aims at improving the boundary localization of the objects. Their approach is to use two processing streams to process the image at different scales. The first stream is the *residual stream*, which is kept at the full image resolution to capture accurate localization information. The second stream is the *pooling stream*, which undergoes the regular encoder-decoder pipeline of FCNs by successive convolution, pooling, and unpooling operations that result in high level features used for classification. The two streams are connected by a series of residual connections that combines the multi-scale information.

The brief summary of the chosen models shows the trade-off often encountered in semantic segmentation, which is the need for abstract features at the deeper layers of the network to achieve higher classification accuracy, while also attending to the localization and the spatial accuracy of the detected objects. The presented models attempt different approaches to process and combine information at different scales in order to address this trade-off.

References

[1] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Alberto Garcia-Garcia, Sergio Orts, Sergiu Oprea, Victor Villena-Martinez, and José García Rodríguez. A review on deep learning techniques applied to semantic segmentation. *CoRR*, abs/1704.06857, 2017.

[3] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Scharwächter, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset. In *CVPR Workshop on The Future of Datasets in Vision*, 2015.

[4] Mark Everingham, S. M. Eslami, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vision*, 111(1):98–136, January 2015.

[5] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[6] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, April 2017.

[7] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018.

[8] Tobias Pohlen, Alexander Hermans, Markus Mathias, and Bastian Leibe. Full-resolution residual networks for semantic segmentation in street scenes. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[9] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017.

[10] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[12] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.