

Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and
Hartwig Adam

Google Inc.

{lcchen, yukun, gpapan, fschroff, hadam}@google.com

Abstract. Spatial pyramid pooling module or encode-decoder structure are used in deep neural networks for semantic segmentation task. The former networks are able to encode multi-scale contextual information by probing the incoming features with filters or pooling operations at multiple rates and multiple effective fields-of-view, while the latter networks can capture sharper object boundaries by gradually recovering the spatial information. In this work, we propose to combine the advantages from both methods. Specifically, our proposed model, DeepLabv3+, extends DeepLabv3 by adding a simple yet effective decoder module to refine the segmentation results especially along object boundaries. We further explore the Xception model and apply the depthwise separable convolution to both Atrous Spatial Pyramid Pooling and decoder modules, resulting in a faster and stronger encoder-decoder network. We demonstrate the effectiveness of the proposed model on PASCAL VOC 2012 and Cityscapes datasets, achieving the test set performance of 89% and 82.1% without any post-processing. Our paper is accompanied with a publicly available reference implementation of the proposed models in Tensorflow at <https://github.com/tensorflow/models/tree/master/research/deeplab>.

Keywords: Semantic image segmentation, spatial pyramid pooling, encoder-decoder, and depthwise separable convolution.

1 Introduction

Semantic segmentation with the goal to assign semantic labels to every pixel in an image [1,2,3,4,5] is one of the fundamental topics in computer vision. Deep convolutional neural networks [6,7,8,9,10] based on the Fully Convolutional Neural Network [8,11] show striking improvement over systems relying on hand-crafted features [12,13,14,15,16,17] on benchmark tasks. In this work, we consider two types of neural networks that use spatial pyramid pooling module [18,19,20] or encoder-decoder structure [21,22] for semantic segmentation, where the former one captures rich contextual information by pooling features at different resolution while the latter one is able to obtain sharp object boundaries.

In order to capture the contextual information at multiple scales, DeepLabv3 [23] applies several parallel atrous convolution with different rates (called Atrous

