Dual Attention Network for Scene Segmentation

Jun Fu [1,3] Jing Liu* [1] Haijie Tian [1] Yong Li [2]
Yongjun Bao [2] Zhiwei Fang [1,3] Hanqing Lu [1]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences
[2]Business Growth BU, JD.com [3] University of Chinese Academy of Sciences
{jun.fu,jliu,zhiwei.fang,luhq}@nlpr.ia.ac.cn,hjtian_bit@163.com,{liyong5,baoyongjun}@jd.com

Abstract

In this paper, we address the scene segmentation task by capturing rich contextual dependencies based on the self-attention mechanism. Unlike previous works that capture contexts by multi-scale feature fusion, we propose a Dual Attention Network (DANet) to adaptively integrate local features with their global dependencies. Specifically, we append two types of attention modules on top of dilated FCN, which model the semantic interdependencies in spatial and channel dimensions respectively. The position attention module selectively aggregates the feature at each position by a weighted sum of the features at all positions. Similar features would be related to each other regardless of their distances. Meanwhile, the channel attention module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. We sum the outputs of the two attention modules to further improve feature representation which contributes to more precise segmentation results. We achieve new state-of-the-art segmentation performance on three challenging scene segmentation datasets, i.e., Cityscapes, PASCAL Context and COCO Stuff dataset. In particular, a Mean IoU score of 81.5% on Cityscapes test set is achieved without using coarse data.[1].

1. Introduction

Scene segmentation is a fundamental and challenging problem, whose goal is to segment and parse a scene image into different image regions associated with semantic categories including stuff (e.g. sky, road, grass) and discrete objects (e.g. person, car, bicycle). The study of this task can be applied to potential applications, such as automatic driving, robot sensing and image editing. In order to accomplish the task of scene segmentation effectively, we need to distinguish some confusing categories and take into account objects with different appearance. For example, regions of 'field' and 'grass' are often indistinguishable, and the objects of 'cars' may often be affected by scales, occlusion and illumination. Therefore, it is necessary to enhance the discriminative ability of feature representations for pixel-level recognition.

Figure 1: The goal of scene segmentation is to recognize each pixel including stuff, diverse objects. The various scales, occlusion and illumination changing of objects/stuff make it challenging to parsing each pixel.

Recently, state-of-the-art methods based on Fully Convolutional Networks (FCNs) [13] have been proposed to address the above issues. One way is to utilize the multi-scale context fusion. For example, some works [3, 4, 29] aggregate multi-scale contexts via combining feature maps generated by different dilated convolutions and pooling operations. And some works [15, 27] capture richer global context information by enlarging the kernel size with a decomposed structure or introducing an effective encoding layer on top of the network. In addition, the encoder-decoder structures [6, 10, 16] are proposed to fuse mid-level and

*Corresponding Author
[1]Links can be found at https://github.com/junfu1115/DANet/