PSDNet: A Balanced Architecture of Accuracy and Parameters for Semantic Segmentation

Yue Liu, Zhichao Lian

Nanjing University of Science and Technology, Nanjing, China

Introduction

Semantic segmentation is a classic topic in computer vision. The task is to assign each pixel of the image a pre-defined class label. Semantic segmentation has wide-ranging applications, such as scene parsing, autonomous driving and medical segmentation, to name a few.

Lots of Convolution neural networks (CNN) based on Fully Convolution Network (FCN) have achieved impressive results on semantic segmentation task. Most state-of-the-art semantic segmentation models focus on exploration of feature information fusion where the coarse and fine features are merged for inference. Pyramid Scene Parsing Network (PSPNet) starts from an innovative component which is called Pyramid Pooling Module (PPM) to aggregate the multi-scale context.

However, some operators used in CNN universally, not only have convolution with big stride but also like upsampling and pooling, can wound information contained in features. The PPM structure can't fully make use of reserved information. Therefore, we aim to propose some modules for accurate segmentation while decreasing weights or introducing fewer parameters. In addition, we still focus on attention module improvement. Our main contributions are threefold:

- We propose a powerful architecture, PSDNet, for accurate semantic segmentation task. We introduce two novel components in Pyramid Pooling Module innovatively. As a kind of feature transformation module, the flexible "Depth to Space Upsampling" (D2SUpsample) module has more excellent capability in features reconstruction than normal interpolations and has less parameters.
- We also propose an effective channel attention module called "Squeeze and Excitation with 1D Convolution" (SE1C) block module to explicitly model interdependencies between channels with fewer parameters.
- We validate the performance of our PSDNet on the Cityscapes dataset with ResNet50 backbone. Our method achieves 73.97% mIoU and 82.89% mAcc, outperforming state-of-the-art method PSPNet.

Methods

PSDNet we proposed starts from well-known PSPNet and introduces two powerful components called SE1Cblock and D2SUsample which show their extraordinary ability of improving accuracy. Furthermore, this modification of PSPNet keeps its main architecture that contains ResNet backbone and four independent parallel branches of previous Pyramid Pooling module, also with auxiliary branch because a broad range of prior researches show that this structure can efficiently take advantage of context.

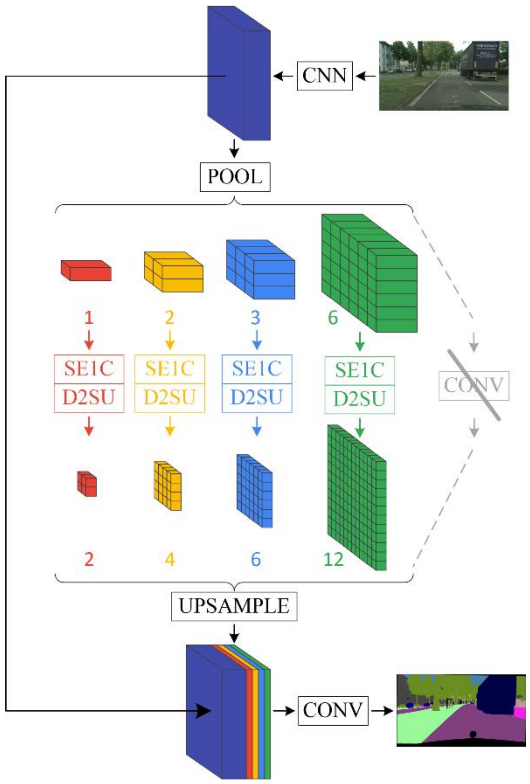


Fig. 1. Overview of our proposed PSDNet. We first use CNN (ResNet) to extract features of input image, then use pooling to get different fine-grained features. The features are fed into our proposed SE1Cblock and D2SUpsample module in different colors. Finally, the concatenation layer combines the previous features which from CNN and the features after upsampling into the final representation, then the merged features are fed into a simple convolution layer to get final prediction.

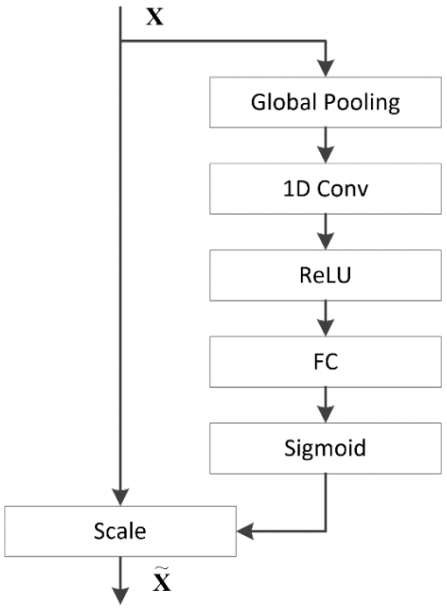


Fig. 2. The schema of our SE1Cblock. Our SE1Cblock module has one global pooling layer, one 1 dimension (1D) convolution layer, one fully-connected layer and sigmoid layer. Different from SEblock, our module replaces the bottom FC layer with 1 dimension (1D) convolution layer.

PSDNet: A Balanced Architecture of Accuracy and Parameters for Semantic Segmentation

Yue Liu, Zhichao Lian

Nanjing University of Science and Technology, Nanjing, China

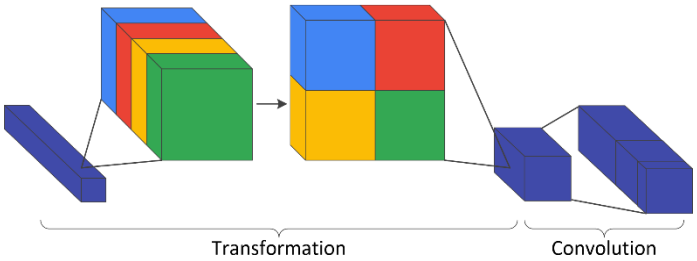


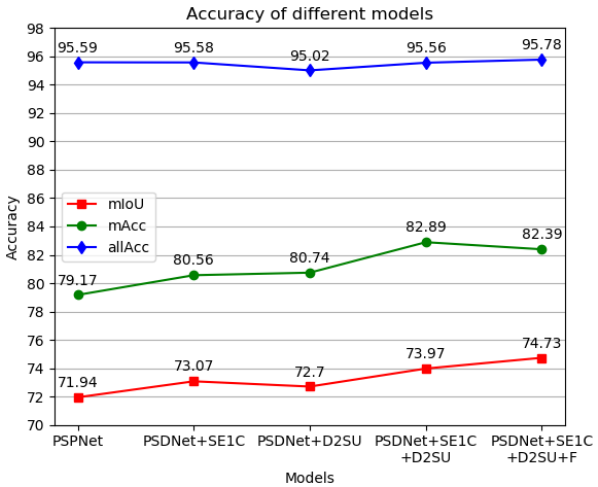
Fig. 3. The schema of our D2SUpsample module. In the transformation stage, the features can be upsampled with factor r (e.g. r=2), then channels reduce to (e.g. 1/4) and the spatial size doubled. In the convolution stage, the convolution layer maps the features to the representation with arbitrary channels with 1×1 kernel.

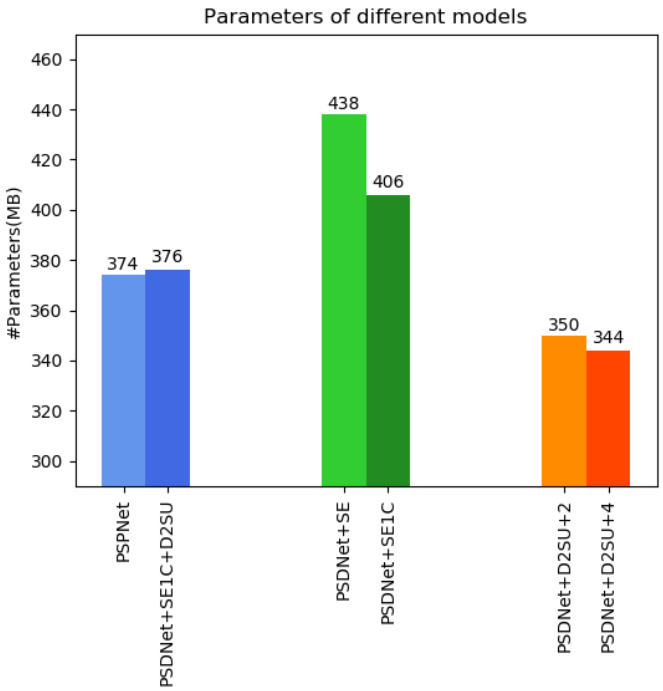
Results

Using the proposed architecture, we achieve good performance on multiple evaluation metrics, mean intersection over union (mIoU), mean accuracy of each class (mAcc) and all pixel accuracy (allAcc). We use the same settings, initialization model and Cityscapes dataset to prove the improvement we got. We evaluate the segmentation results on the server with a single GTX 1080Ti.

TABLE I. PARAMETERS AND ACCURACY OF DIFFERENT MODELS

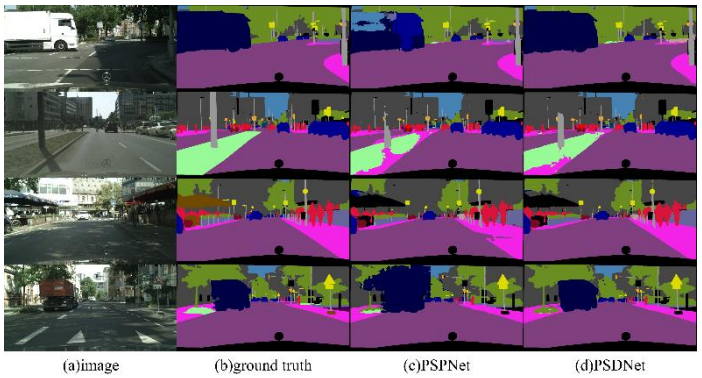
Model	*#Parameters*	*mIoU*	*mAcc*	*allAcc*
PSPNet	374MB	71.94	79.17	95.59
PSDNet+SE	438MB	71.97	79.56	95.55
PSDNet +SE1C	406MB	73.07	80.56	95.58
PSDNet +D2SU-2	350MB	74.35	82.36	95.54
PSDNet +D2SU-4	344MB	72.70	80.74	95.02
PSDNet +SE+D2SU	408MB	74.06	81.54	95.65
PSDNet +SE1C+D2SU	376MB	73.97	82.89	95.56
PSDNet +SE1C+D2SU+F	376MB	74.73	82.39	95.78





The detailed results are shown in Table I. PSPNet achieves 71.94% mIoU and 79.17% mAcc, while our PSDNet achieves 73.97% mIoU, 82.89% mAcc and 95.56% allAcc on Cityscapes dataset with ResNet50 backbone and increases only about 2MB parameters.

Conclusions



We have proposed a novel architecture called PSDNet for semantic segmentation task. The two powerful modules we introduced have already demonstrated the ability to improve performance in term of accuracy. Our PSDNet is more accurate in classification of large vehicles internal details, and is more accurate in prediction of poles and the boundary of road.