DILATED CONVOLUTION AND SPATIAL PYRAMID FUSION IN THE IMAGE SEGMENTATION PROBLEM

Zhao Di,Tang Yi, Gourinovitch A.B.
Department of Information Technologies in Automated Systems,
Belarussian State University of Informatics and Radioelectronics
Minsk, Republic of Belarus
E-mail: 3189124246@qq.com, tangyijcb@163.com, gurinovich@bsuir.by

Image segmentation is one of the important tasks in the computer vision, where the goal is to segment different regions in an image into semantically meaningful parts. However, due to the presence of target and contextual information at different scales in an image, traditional segmentation methods face the challenges of information loss and lack of accuracy when dealing with images at different scales. To address this problem, this study proposes an innovative approach that combines dilated convolution and spatial pyramid pooling to improve the processing power and accuracy of segmentation models for images of different scales.

Introduction

Image segmentation technology has important application value in medical image analysis, automatic driving, image editing and other fields. However, traditional image segmentation methods have certain limitations when dealing with medical images of different scales. This is because traditional methods usually use a fixed-size convolutional kernel to extract features, which cannot effectively capture the correlation between targets and contextual information at different scales, leading to performance degradation when processing images with multi-scale features, and facing the challenges of information loss and lack of accuracy. To address these problems, methods such as dilated convolution and spatial pyramid pooling are widely used in image segmentation.

To overcome these limitations, this article proposes an effective method that combines two key techniques: dilated convolution and spatial pyramid pooling. Dilated convolution captures multi-scale contextual information, while spatial pyramid pooling facilitates the fusion of multi-scale features. By integrating these techniques, the segmentation model can be enhanced to handle medical images at different scales, thus improving the accuracy and processing of multi-scale features.

I. Problem statement

Medical images usually have complex anatomical structures and detailed information at different scales. In recent years, researchers have proposed some effective methods to improve the image segmentation method, but they still have limitation problems when dealing with images at different scales [1]. Several of these methods are described below:

Multiscale Pyramid: Constructing a multiscale pyramid allows obtaining feature representations at different scales for better handling of multiscale targets. However, this requires the introduction of multiple branches or convolution kernels at multiple scales in the network, increasing the number of parameters and computational complexity of the network. In addition, when fusing features from different scales, the fusion strategy needs to be carefully designed to avoid confusion or conflict of information.

Residual connectivity: Residual connectivity can facilitate the transfer of information and the flow of gradients, which helps to improve the performance of image segmentation. However, when dealing with images at different scales, lower resolution features may interfere with higher resolution features, leading to inaccurate segmentation results.

Multi-scale fusion: Multi-scale fusion aims to combine feature information from different scales to improve the performance of the segmentation model. However, multiscale fusion may require additional computational and memory overheads, especially in feature-level fusion.

Dilated Convolution: By using a larger dilated rate, the sensory field can be expanded to capture a wider range of contextual information. However, a larger dilation rate may lead to a decrease in the resolution of the feature map, which may reduce the accuracy of details and boundaries.

II. Method description

Firstly. The article uses dilated convolution to expand the sensory field of the convolution operation. In traditional convolutional operations, the size of each convolutional kernel is fixed and sliding computation is performed over a local region. However, this operation can only capture a limited range of contextual information, limiting the model's ability to understand a larger range of features. In contrast, dilated convolution increases the effective size of the convolution kernel by inserting fixed intervals of zeros within the convolution kernel. This interpolation operation is called dilated, and the dilation rate determines the size of the interpolated interval. A larger dilation rate effectively expands the receptive field, allowing the convolution operation to capture a wider range of contextual information, Figure 1 shows the schematic diagram of dilated convolution.

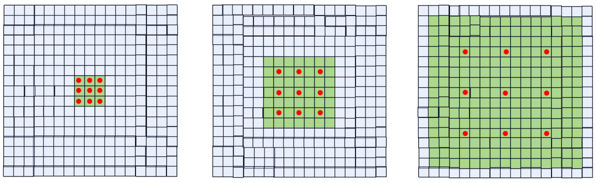

Figure 1 – The schematic diagram of dilated convolution

Figure 1 shows the following: Cavity convolution results in a convolution kernel of the same size receiving a larger sensor field by introducing a dilation rate parameter. From left to right the different sensory fields are shown. When the dilation rate = 1, 2, and 4 respectively, then the size of the sensory field can be calculated from the size of the convolution kernel and the expansion rate. Assuming that the convolution kernel size is K and the expansion rate is R, the formula for the sensory field (F) is denoted by formula (1):

$$F = K + (K - 1)(R - 1) \tag{1}$$

By introducing dilated convolution, the sensory field can be increased without increasing network parameters. This is important for dealing with targets and scenes at different scales, since the relationships between features and structures at different scales require a larger range of contextual information to be accurately understood.

Secondly. For the CNN structure, the change of the image size mainly affects the fully connected layer (FC), so [2] adds the spatial pyramid pooling layer, between the convolutional layer (including convolution, Pooling, etc.) and the fully connected layer (FC). It allows the fully connected layer to the fixed number of features obtained and removes the restriction on the fixed size of the network. The spatial pyramid pooling module consists of multiple parallel pooling layers with different sense field sizes.

By transforming input images into fixed size outputs, spatial pyramid pooling can adapt to different sizes of input images. Its main idea is to divide the input image into grids of different sizes, carry out pooling operations on each grid, and then stitch all the pooling results together as the output. In this way, a fixed size output can be obtained regardless of the size of the input image. With pyramid pooling, it can aggregate features at different scales to capture information at different scales in the image.

Compared with the traditional fixed-length feature vectors, spatial pyramid pooling can better preserve the feature information at different scales and avoid information loss. The structure of spatial pyramid pooling is shown in Figure 2.

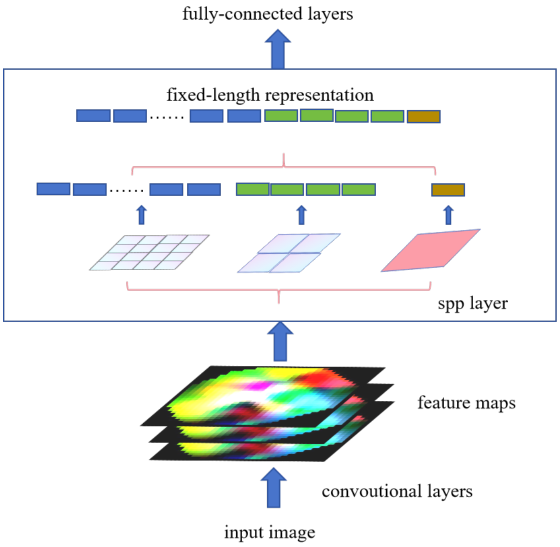

Figure 2 – The structure of spatial pyramid pooling

By performing pooling operations on feature maps at different scales, multiple feature representations at different scales can be obtained. In each pooling layer, the article performs average pooling or maximum pooling operations on the pooled feature maps to generate fixed-length feature vectors. These feature vectors represent feature information at different scales[3]. In this way, it avoid the problem of information loss that may result from using a single fixed-length feature vector to represent features at different scales. Finally, the article fuses these multi-scale feature vectors, which can be weighted and fused using a simple splicing operation or by introducing an attention mechanism. The fused feature vectors can be passed to subsequent classification or segmentation modules for final prediction.

III. CONCLUSION

The innovative approach is proposed to handle the task of image segmentation at different scales by combining dilated convolution and spatial pyramid pooling. It is able to efficiently capture contextual information at different scales and significantly improve the performance and robustness of the segmentation model. Further study can explore how to optimize the structure of dilated convolution and spatial pyramid pooling. The combination with other advanced techniques improves the model's efficiency. It is necessary to advance the development of image segmentation techniques and achieve better results in various practical applications.

1. Yu, F., & Koltun, V. (2016). Multi-Scale Context Aggregation by Dilated Convolutions. In International Conference on Learning Representations (ICLR).
2. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 37(9), 1904-1916.
3. Chen, L., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking Atrous Convolution for Semantic Image Segmentation. ArXiv, abs/1706.05587.