OXENDONET: A DILATED CONVOLUTIONAL NEURAL NETWORKS FOR ENDOSCOPIC ARTEFACT SEGMENTATION

Mourad Gridach[1,2], *Irina Voiculescu*[2]

[1] Department of Computer Science, High Institute of Technology, Agadir
[2] Department of Computer Science, University of Oxford, UK

ABSTRACT

Medical image segmentation plays a key role in many generic applications such as population analysis and, more accessibly, can be made into a crucial tool in diagnosis and treatment planning. Its output can vary from extracting practical clinical information such as pathologies (detection of cancer), to measuring anatomical structures (kidney volume, cartilage thickness, bone angles). Many prior approaches to this problem are based on one of two main architectures: a fully convolutional network or a U-Net-based architecture. These methods rely on multiple pooling and striding layers to increase the receptive field size of neurons. Since we are tackling a segmentation task, the way pooling layers are used reduce the feature map size and lead to the loss of important spatial information. In this paper, we propose a novel neural network, which we call OxEndoNet. Our network uses the pyramid dilated module (PDM) consisting of multiple dilated convolutions stacked in parallel. The PDM module eliminates the need of striding layers and has a very large receptive field which maintains spatial resolution. We combine several pyramid dilated modules to form our final OxEndoNet network. The proposed network is able to capture small and complex variations in the challenging problem of Endoscopy Artefact Detection and Segmentation where objects vary largely in scale and size.*

1. INTRODUCTION

Medical image segmentation [1, 2] is an important step in many medical applications such as population analysis, diagnosis disease, planning treatments and medical intervention, where the goal is to extract useful information such as pathologies, biological organs and structures. In most clinics, segmentation currently relies on the time consuming task of drawing contours manually, by medical experts for instance radiologists, pathologists, ophthalmologists, etc. This can be challenging because features of interest (soft tissue, blood vessels, cancer cells) can have large and complex variations (contrast, blur, noise, artifacts, and distortion). Automating even part of the segmentation task is a good way of reducing time spent on routine activities, as well as improving the handling of larger volumes of data which are increasingly available from a large variety of modern scanners. Any such automated process should, of course, still allow for manual override by a human expert.

Recently, deep neural networks (DNNs), have been successfully used in semantic and biomedical image segmentation. Long et al. [3] proposed a fully convolutional network (FCN) to perform end-to-end semantic image segmentation, which surpasses all the existing approaches. Ronneberger et al. [4] developed a U-shaped deep convolutional network called U-Net consisting of contracting path to capture context and a symmetric expanding path that enables precise localization. Using this (now widely cited) architecture, U-Net outperforms all the previous models by a significant margin. Based on U-Net, Chen et al. [5] developed a model called DCAN, which won the *2015 MICCAI Gland Segmentation Challenge*.

Such approaches suffer from two main limitations: firstly, with complex and large variations in the size of objects in medical images, the FCN with single receptive field size fails to deal with such variations. Secondly, like in the case of object detection and classical semantic segmentation, in medical image segmentation global context is also very important. Classical networks such as U-Net and FCN miss some parts of the images because they fail to see the entire image and incorporate global context in producing the correct segmentation mask. For example, U-Net only has receptive fields that spann 68×68 pixels [4].

Our goal has therefore been to design a network that is able to integrate global context in order to detect and assess the interdependence of organs in medical images.

To address the issues described above, we propose the novel OxEndoNet, a neural network architecture based on dilated convolutions. This architecture tackles the challenging variations in the size of anatomical features in medical images. OxEndoNet is used to address the problems of the EAD2020 Challenge (a multi-class artefact segmentation in video endoscopy).

Our network has a large receptive field that uses a novel

*Work carried out during a collaborative visit at the University of Oxford

architecture module called the Pyramid Dilated Module (PDM) to capture highly appropriate, robust and dense local and global features, which directly influence the final prediction and make it more accurate. The PDM module consists of multiple dilated convolutions stacked in parallel. Combining several PDM layers leads to our OxEndoNet network, which is detailed in Section 3.3.

Unlike many methods used in other similar challenges, an ensemble of models was *not* used in this case, which makes our OxEndoNet a promising framework for the future.

2. DATASETS

In this challenge, we use the Endoscopy Artefact Detection and Segmentation dataset[1]. Its goal is to capture the wide visual diversity in endoscopic videos acquired in everyday clinical settings. For more details about the dataset, we refer the reader to [6, 7, 8]. The training employed the released data split into two sets: 80% of it was used for training per se, whereas the remainder 20% was kept as validation data. The final architecture is based on the results from the validation data. The metrics around which the learning was based are Accuracy and F_1-score, hence our network scoring well in the F_1 measure in this challenge.

3. METHODS

Some background information about other networks is necessary in order to describe our proposed architecture.

3.1. Dilated Convolution

Dilated convolution (or Atrous convolution) was originally developed in *algorithme à trous* for wavelet decomposition [9]. The main idea of dilated convolutions is to insert holes (*trous* in French) as zeros between pixels in convolutional filters. As a result, we increase the image resolution, which allows dense feature extraction in convolutional neural networks. More formally, given 1-d input signal f and y as the output signal at location i of a dilated convolution, we represent dilated convolution in one dimension as the following:

$$y[i] = \sum_{s=1}^{S} f[i + d \cdot s] \cdot w[s] \qquad (1)$$

where $w[s]$ denotes the s^{th} parameter of the filter, d is the dilation rate, and S is the filter size. When $d = 1$, dilated convolutions correspond to standard convolutions. In other words, dilated convolution is equivalent to convolving the input f with up-sampled filters produced by inserting $d - 1$ zeros between two consecutive filter values. Therefore, a large dilation rate means a large receptive field. Its main advantage is the ability to enlarge the receptive field size to incorporate context without introducing extra parameters or computation cost. Dilated convolution has been successfully applied in many computer vision applications such as audio generation [10], object detection [11], and semantic segmentation [12].

[1] https://ead2020.grand-challenge.org

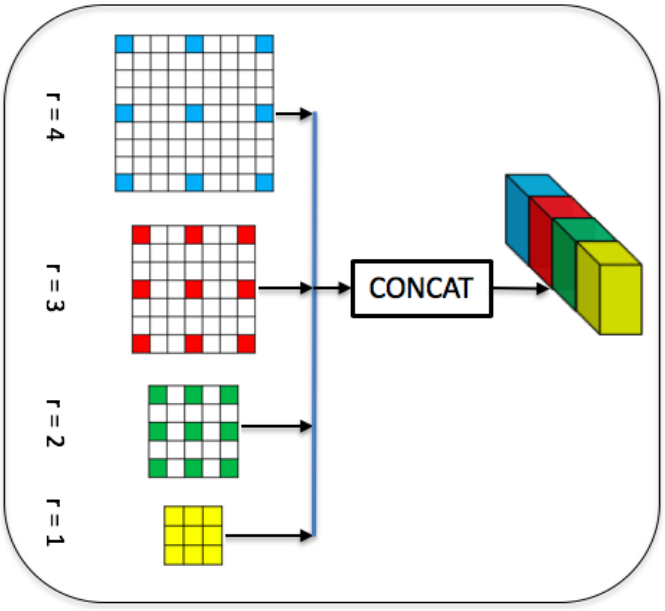

Fig. 1. Pyramid Dilated Module architecture. We stacked four dilated convolutions with dilation rates of 1, 2, 3 and 4 in parallel. The results of convolutions are concatenated.

3.2. Pyramid Dilated Module

In a deep neural network, the size of receptive field plays an important role in indicating the extent to which context information is used. Previous work uses pooling layers and strided convolution to enlarge the receptive field. These techniques significantly improve the performance in applications like image classification and object detection because they require a single prediction per input image. However, in tasks requiring dense per-pixel prediction such as image segmentation, strided layers often fail to get better results because some details about the spatial information is lost, which influences the pixel-wise prediction. An alternative solution to strided convolution is to increase the size of the filters.

A common limitation of this method is a severe increase in the number of parameters to optimize and training time.

3.3. OxEndoNet Network

Motivated by the recent success of dilated convolution, we propose a new pyramid dilated module (PDM), which empirically proves to be a powerful feature extractor in endoscopy

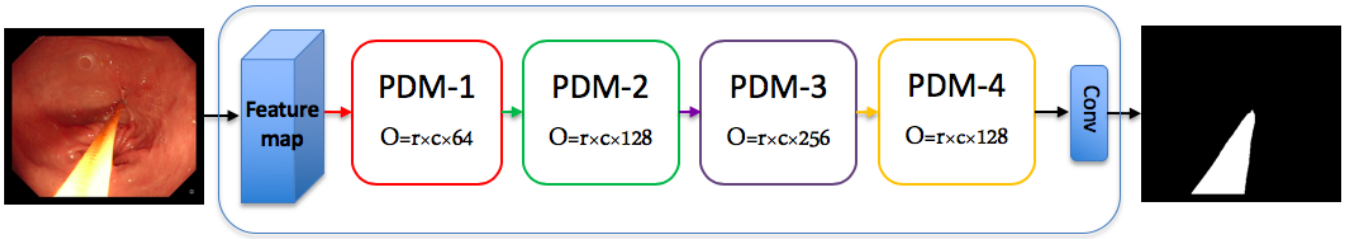

Fig. 2. OxEndoNet architecture. O, r×c×d refer to the output of each PDM layer and dimensions respectively.

artefact detection and segmentation task. As shown in Figure 1, we stacked convolutions with different dilation rates in parallel. In this case, PDM has four parallel convolutions with 3×3 filter size and dilation rates of 1, 2, 3 and 4. The activation function we used is the Rectified Linear Unit (ReLU) [13]. The result of each convolution with dilation rate produces the same number of output dimension. To form the final PDM module, we concatenate the outputs of each dilated convolution. By combining the dilated convolutions with different dilation rates, the PDM module is able to extract useful features for objects of various sizes. All the previous advantages play a remarkable role in medical image segmentation, because medical images often feature organs of different sizes.

Given this PDM, we propose the OxEndoNet network illustrated in Figure 2. For each input image, we use ResNet-50 pretrained on ImageNet [14] as the base network to extract the feature map followed by multiple PDM layers to form an end-to-end trainable network. By using several layers, we increase the receptive field size which allows our model to use context information. In the final architecture, we use four PDM layers; each layer uses four parallel dilated convolutions with filter size of 3×3 and dilation rates of 1, 2, 3, and 4. We note that the number of PDM layers and the number of parallel dilated convolutions are hyperparameters. The PDM layers have 64, 128, 256, and 128 output channels where we use 16, 32, 64 and 32 filters respectively. We feed the final PDM layer to a convolution layer followed by a bilinear interpolation to up-scale the feature map to the original size of an image.

The architecture design followed two key observations. Firstly, recognizing organs in medical images requires a high spatial precision that is lost when applying pooling with striding layers. This is the main issue in FCN- and U-Net-based models. Secondly, complex and large variations in the size of objects in medical images lead to inaccurate prediction due to the small or medium sized receptive field which fails to deal with such variations. Therefore, an accurate model should have a large receptive field to handle these complex variations of organs in images. Our OxEndoNet network produces a large receptive field to incorporate larger context without increasing the number of parameters or the amount of computation while preserving full spatial resolution.

Model	Overlap	F2-score	score-s
OxEndoNet	0.4901	0.5107	0.5194

Table 1. Results of OxEndoNet on phase 1 test data.

4. EXPERIMENTS AND RESULTS

We implemented OxEndoNet using the public framework PyTorch [15]. The number of PDM layers, learning rate and the number of parallel dilated convolutions are the main hyperparameters that influenced our models performance. During training, we used the Adam optimizer [16] with the default initial learning rate of 3.10^{-3} and weight decay of 10^{-4}. Furthermore, we used the poly learning rate policy [17] by multiplying the initial rate with $(1 - epoch/maxEpoches)^{0.9}$ and trained the models for 300 epochs. For the number of PDM layers, we conduct experiments with 3, 4 and 5 layers. Concerning the number of parallel dilated convolutions, we ran experiments with 3, 4 and 5 parallel convolutions. It should also be noted that all the hyperparameters were selected based on performance on validation data.

We tested the performance of our model on the released test data named as Test Data Phase 1, which consisted of 50% of the overall test data. In Phase 1, the test data contained 80 images, the results of which we submitted to the challenge. Table 1 shows the results of our model on this test data. The overall results will specified after the workshop.

5. DISCUSSION & CONCLUSION

We have described OxEndoNet, a neural network designed to tackle the challenging problem of Endoscopy Artefact Detection and Segmentation where objects vary largely in scale and size. Its use of pyramid dilated module consists of parallel dilated convolutions concatenated to provide additional contextual information. The need of pooling and striding layers, considered a major drawback of other segmentation methods, is fully eliminated. Both PDM and OxEndoNet will be useful frameworks to explore by the community for other computer vision tasks. In the future, we plan to test our model on a wide variety of medical image volumes, as well as on generic

semantic image segmentation tasks.

6. ACKNOWLEDGMENT

We would like to thank AfOx for the visiting fellowship which supported Mourad Gridach's visit to the Oxford Department of Computer Science, where this model was designed.

7. REFERENCES

[1] Scott Fernquest, Daniel Park, Marija Marcan, Antony Palmer, Irina Voiculescu, and Sion Glyn-Jones. Segmentation of hip cartilage in compositional magnetic resonance imaging: A fast, accurate, reproducible, and clinically viable semi-automated methodology. *Journal of Orthopaedic Research®*, 36(8):2280–2287, 2018.

[2] Varduhi Yeghiazaryan and Irina D Voiculescu. Family of boundary overlap metrics for the evaluation of medical image segmentation. *Journal of Medical Imaging*, 5(1):015006, 2018.

[3] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[5] Hao Chen, Xiaojuan Qi, Lequan Yu, and Pheng-Ann Heng. Dcan: deep contour-aware networks for accurate gland segmentation. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2016.

[6] Sharib Ali, Felix Zhou, Christian Daul, Barbara Braden, Adam Bailey, Stefano Realdon, James East, Georges Wagnieres, Victor Loschenov, Enrico Grisan, et al. Endoscopy artifact detection (ead 2019) challenge dataset. *arXiv preprint arXiv:1905.03209*, 2019.

[7] Sharib Ali, Felix Zhou, Adam Bailey, Barbara Braden, James East, Xin Lu, and Jens Rittscher. A deep learning framework for quality assessment and restoration in video endoscopy. *arXiv preprint arXiv:1904.07073*, 2019.

[8] Sharib Ali, Felix Zhou, Barbara Braden, Adam Bailey, Suhui Yang, Guanju Cheng, Pengyi Zhang, Xiaoqiong Li, Maxime Kayser, Roger D. Soberanis-Mukul, Shadi Albarqouni, Xiaokang Wang, Chunqing Wang, Seiryo Watanabe, Ilkay Oksuz, Qingtian Ning, Shufan Yang, Mohammad Azam Khan, Xiaohong W. Gao, Stefano Realdon, Maxim Loshchenov, Julia A. Schnabel, James E. East, Geroges Wagnieres, Victor B. Loschenov, Enrico Grisan, Christian Daul, Walter Blondel, and Jens Rittscher. An objective comparison of detection and segmentation algorithms for artefacts in clinical endoscopy. *Scientific Reports*, 10, 2020.

[9] Matthias Holschneider, Richard Kronland-Martinet, Jean Morlet, and Ph Tchamitchian. A real-time algorithm for signal analysis with the help of the wavelet transform. In *Wavelets*, pages 286–297. Springer, 1990.

[10] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[11] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[12] Zhengyang Wang and Shuiwang Ji. Smoothed dilated convolutions for improved dense prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2486–2495, 2018.

[13] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323, 2011.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.