SEGMENTATION OF THORACIC ORGANS USING PIXEL SHUFFLE

Dmitry Lachinov

Intel, Nizhny Novgorod
dmitry.lachinov@intel.com

ABSTRACT

This paper summarizes our contribution to the SegTHOR challenge. This competition addresses the problem of organs at risk segmentation in CT images. For lung cancer treatment, segmentation of nearby healthy organs is essential. The task of organs delineation is largely manual and can be potentially a source of mistakes. At this point, the segmentation of organs that are located close to the tumor is a routine and tedious procedure. With the intention to simplify this procedure we study approaches of automatic organs segmentation within CT images.

The solution we came up with is based on deep learning and explores two concepts: attention mechanism and pixel shuffle as an upsampling operator. In this study, we describe our approach in details and evaluate it with test data provided by challenge organizers. Without any post-processing our method achieves notable performance with following intermediate results: 0.8303, 0.9381, 0.9088, 0.9353 for Esophagus, Heart, Trachea, and Aorta respectively (Dice scores are reported).

***Index Terms*—** SegTHOR, Segmentation, CT, Medical Image Processing, Pixel Shuffle

1. INTRODUCTION

3D Computed tomography is a powerful tool for the human body examination. Being a noninvasive diagnostic method it has been firmly integrated into different therapy protocols. However, despite its pros, this examination method has its own drawbacks. Mainly, noise, low image contrast or even absence of organs' contours are the main challenges in CT scan analysis. Besides this, the dimensional representation of the input data imposes multiple restrictions on the way the scan can be analyzed. All of the manual methods are tedious and requires high level of concentration, at the same time, any possible mistake during scan analysis process can potentially become a serious problem in further therapy. At this point, we are focusing to develop an automatic solution for segmentation routine.

With the recent advances in Machine Learning and Deep Learning, in particular, the scientific community has developed new techniques for vision tasks that are superior to the classic computer vision methods. Talking about semantic segmentation, Fully Convolutional Neural Networks [1] first achieved decent performance on such type of tasks. All of the modern neural network architectures explore the same concept. For the natural image semantic segmentation competition is high: FCNs [1], SegNet [2], DeepLab architectures [3], PSPNet [4] and others show really high performance. In medical image semantic segmentation domain UNet and its variants [5, 6, 7, 8, 9, 10] show state of the art results. In this study, we are trying to adapt the existing framework for the purpose of segmentation of four organs: Esophagus, Heart, Trachea, and Aorta.

Denoting the problem of automatic organs at risk segmentation SegTHOR challenge [11] provides competition in classifying given 3D CT voxels into five different regions: Background, Esophagus, Heart, Trachea, and Aorta. Typically, this procedure is manual and requires a high amount of time and can produce repeatable errors. In this paper, we propose an automatic solution for the above-named problem and evaluate it with the data provided by the challenge organizers.

2. RELATED WORK

In this section, we describe prior work that we are using in this paper.

We base our model on famous UNet architecture [5] introduced by Olaf Ronneberger et al. for the purposes of biomedical image segmentation, cells segmentation in particular. The proposed network consists of encoding and decoding paths. The skip connections employed between these paths enhance localization capabilities and also help in solving the vanishing gradient problem. The high number of channels in contracting part of the network allows propagating information further to higher resolution layers. Later, Attention mechanisms incorporated into UNet architecture were studied in works [12] and [13].

The second concept we are using is the neural network with residual connections, so-called ResNet [14]. The authors propose a deep architecture that can be trained efficiently. In order to propagate gradients closer to the starting layers of the network, residual blocks are proposed. Later, different residual blocks architectures [15] were studied.

For the task of image super-resolution, Shi et al. at [16]

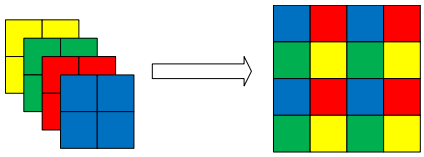

Fig. 1. The example of pixel shuffle operation. Four given channels form one output channel with higher spatial resolution.

proposed to use pixel shuffle as an upsampling operator. The idea of pixel shuffle is illustrated in figure 1. This operator rearranges input channels to produce a feature map with higher resolution. Worth to mention, this technique solves the problem of checkerboard artifacts in the output image, Later, the same concept was employed for semantic segmentation tasks [17, 18].

3. OUR METHOD

3.1. Data

The dataset is split into two parts by organizers: training and testing. Training part has 40 CT images with voxel size varying between 0.90 mm and 1.37 mm per pixel. Majority of the images in the train set have 512x512 slice resolution. The number of slices varies from 150 to 284. Ground truth labels are provided for every image in the training dataset and contain manual segmentation on five different classes: Background, Esophagus, Heart, Trachea, and Aorta. Example of the CT scan and corresponding labels are demonstrated in figure 2. No preprocessing is applied to the data.
The testing dataset has 20 images. Total 10 submissions are available for participants to test their methods.

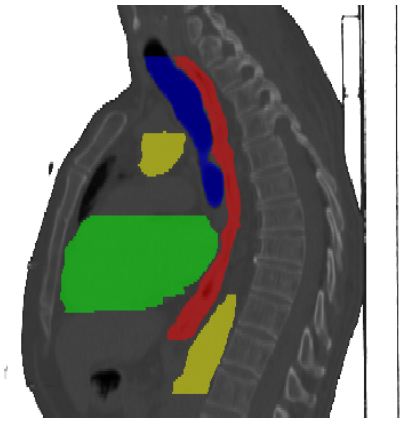

Fig. 2. Example of training data slice with provided labels.

3.2. Preprocessing

Since both testing and training data has a different spatial resolution, as the first step in preprocessing pipeline we resample every image to the 2x2x2.5 mm^3 resolution. As the next step, we crop the body region from the image by applying median filter that eliminates the examination table from the picture. Remaining region is cropped from the original image and passed further. Finally, standard deviation and mean of the body voxels are calculated, and then all image voxels are normalized according to these values.

3.3. Method

We employed fully convolutional neural network architecture based on UNet, with skip connections between contracting and expanding paths and exponentially growing number of channels across consecutive spatial resolution levels. We choose starting number of feature channels in the network to be equal to 16.
Our architecture consists of encoding part which is a residual network [14] with the depth of 3 with 3, 4 and 6 full preactivation residual blocks at each level respectively. In our experiment, we noticed that deeper networks does not improve results but increase computational workload and can be a potential source of the overfitting due to a large number of parameters. Instead of Batch Normalization [19] we are using Group Normalization [20] with the number of groups equals to 4. As an activation function, we use Leaky ReLU with slope equals to 0.2.
In the expanding part of the network, we employ two consecutive convolutions followed by activation at each scale. As an upsampling operator, we have adopted the pixel shuffle [16] technique to handle three-dimensional input. The example is illustrated in figure 3 where eight three dimensional feature maps produce a single three-dimensional output feature map with higher spatial resolution.
In addition to this, we employ the attention mechanism described in paper [13]. Due to the nature of annotation protocol, we found attention mechanism to work especially well for this benchmark.

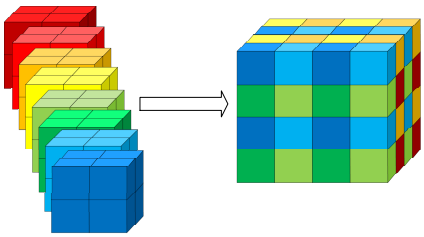

Fig. 3. Extension of pixel shuffle for three dimensions.

Table 1. Performance of proposed method on local cross-validation and testing dataset, Dice index is reported. ESO stands for Esophagus, HEA - Heart, TRA - Trachea, AOR - Aorta.

	ESO	HEA	TRA	AOR
CV	0.7910	0.9193	0.9008	0.9020
Test	0.8303	0.9381	0.9088	0.9353

3.4. Training

We crop region with size 176x96x128 from the input image and randomly mirror it along the first two axes. Then we apply intensity shift augmentation.
The loss function we are using is Dice loss function that can be written as following:

$$L_{Dice}(gt, pred) = 2 * \frac{\sum gt * pred + \epsilon}{\sum (gt^2 + pred^2) + \epsilon}$$

where gt is ground truth one-hot encoded labels, and $pred$ are output logits. For optimization, we are using Adam with initial learning rate set to $1e-3$ and decaying with a rate of 0.1 at 7th and 9th epoch. To evaluate the performance we are using cross-validation scheme with four splits. To train our network we are using three NVIDIA GTX 1080TIs with PyTorch framework [21]. The network is trained with batch size 6 for 10 epochs. Each epoch has 3200 iterations in it. The whole training takes approximately one day.

4. EVALUATION

For evaluation, we are using the cross-validation scheme with the number of splits equals to 4. Since no validation dataset was provided and the number of training samples was limited, we decided that it was the best option for tracking the performance of our experiments. The accuracy of our model on training dataset measured with cross-validation with the number of splits equals to four is reported in table 1.

5. INTERMEDIATE RESULTS AND CONCLUSION

The scores reported by the testing systems are listed in table 1. Comparing cross validation and testing values we can notice that Dice scores for CV are consistently lower compared to the test results. This might indicate that training dataset is more diverse and contain more difficult samples.
In conclusion, proposed in this paper model achieves notable performance with the following intermediate results: 0.8303, 0.9381, 0.9088, 0.9353 for Esophagus, Heart, Trachea, and Aorta respectively (Dice scores are reported). This is done with no post-processing included in the segmentation pipeline.

6. REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 3431–3440.

[2] Vijay Badrinarayanan, Ankur Handa, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling," *arXiv preprint arXiv:1505.07293*, 2015.

[3] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *The European Conference on Computer Vision (ECCV)*, September 2018.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 6230–6239.

[5] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

[6] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger, "3d u-net: Learning dense volumetric segmentation from sparse annotation," *CoRR*, vol. abs/1606.06650, 2016.

[7] Ozan Oktay, Jo Schlemper, Loïc Le Folgoc, Matthew C. H. Lee, Mattias P. Heinrich, Kazunari Misawa, Kensaku Mori, Steven G. McDonagh, Nils Y. Hammerla, Bernhard Kainz, Ben Glocker, and Daniel Rueckert, "Attention u-net: Learning where to look for the pancreas," *CoRR*, vol. abs/1804.03999, 2018.

[8] Ruirui Li, Mingming Li, and JiaCheng Li, "Connection sensitive attention u-net for accurate retinal vessel segmentation," 2019.

[9] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein, "Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, Eds., Cham, 2018, pp. 287–297, Springer International Publishing.

[10] Andriy Myronenko, "3d mri brain tumor segmentation using autoencoder regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Alessandro Crimi, Spyridon Bakas, Hugo

Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, Eds., Cham, 2019, pp. 311–320, Springer International Publishing.

[11] R. Trullo, C. Petitjean, S. Ruan, B. Dubray, D. Nie, and D. Shen, "Segmentation of organs at risk in thoracic ct images using a sharpmask architecture and conditional random fields," in *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, April 2017, pp. 1003–1006.

[12] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al., "Attention u-net: learning where to look for the pancreas," *arXiv preprint arXiv:1804.03999*, 2018.

[13] Ruirui Li, Mingming Li, and JiaCheng Li, "Connection sensitive attention u-net for accurate retinal vessel segmentation," 2019.

[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*, 2016, pp. 770–778.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Identity mappings in deep residual networks," in *European conference on computer vision*. Springer, 2016, pp. 630–645.

[16] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang, "Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1874–1883.

[17] Kaiqiang Chen, Kun Fu, Menglong Yan, Xin Gao, Xian Sun, and Xin Wei, "Semantic segmentation of aerial images with shuffling convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 173–177, 2018.

[18] Hongyang Gao, Hao Yuan, Zhengyang Wang, and Shuiwang Ji, "Pixel deconvolutional networks," *arXiv preprint arXiv:1705.06820*, 2017.

[19] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[20] Yuxin Wu and Kaiming He, "Group normalization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[21] 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," in *NIPS-W*, 2017.