AUTOMATIC SEGMENTATION OF ORGANS AT RISK IN THORACIC CT SCANS BY COMBINING 2D AND 3D CONVOLUTIONAL NEURAL NETWORKS

Louis D. van Harten[1], Julia M. H. Noothout[1], Joost J. C. Verhoeff[2], Jelmer M. Wolterink[1], Ivana Išgum[1]

[1]Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands

[2]Department of Radiotherapy, University Medical Center Utrecht, Utrecht, The Netherlands

ABSTRACT

Segmentation of organs at risk (OARs) in medical images is an important step in treatment planning for patients undergoing radiotherapy (RT). Manual segmentation of OARs is often time-consuming and tedious. Therefore, we propose a method for automatic segmentation of OARs in thoracic RT treatment planning CT scans of patients diagnosed with lung, breast or esophageal cancer. The method consists of a combination of a 2D and a 3D convolutional neural network (CNN), where both networks have substantially different architectures. We analyse the performance for these networks individually and show that a combination of both networks produces the best results. With this combination, we achieve average Dice coefficients of 0.84± 0.05, 0.94± 0.02, 0.91± 0.02, and 0.93± 0.01 for the esophagus, heart, trachea, and aorta, respectively. These results demonstrate potential for automating segmentation of organs at risk in routine radiotherapy treatment planning.

***Index Terms*—** Organs at risk segmentation, dilated convolutional neural network, residual connections, CT, deep learning

1. INTRODUCTION

Cancer is a global leading cause of death, with an increasing prevalence due to growth and aging of the population [1]. One treatment available for cancer is radiation therapy (RT), during which high doses of radiation are delivered to kill cancer cells [2]. RT treatment planning often starts with segmentation of the target volume and healthy organs surrounding the tumor, i.e. organs at risk (OARs), in CT scans [3]. Manual segmentation is often time-consuming and error prone due to large anatomical variation between patients, poor soft-tissue contrast, and high levels of image noise in scans. Therefore, methods have been proposed to automatically segment OARs in CT scans.

Previously published methods for OAR segmentation have used techniques such as thresholding and Hough transforms [4] or multi-atlas registration and level sets [5]. Recently, convolutional neural networks (CNNs) have been used for OAR segmentation. Trullo et al. [6] used a CNN in combination with a conditional random field, implemented as a recurrent neural network architecture, to segment OARs in thoracic CT scans. Men et al. [7] used a CNN containing dilated convolutions at the front- and back-end of a VGG-16 inspired network architecture to segment OARs in treatment planning CT scans for rectal cancer.

In this work, we propose to use an ensemble of CNNs for segmentation of the esophagus, heart, trachea, and aorta in RT treatment planning CT scans for patients suffering from lung, breast or esophageal cancer. It has been shown that the combination of multiple segmentation CNNs in an ensemble can lead to improved segmentation results [8]. However, a drawback of ensemble methods is that training and combining multiple CNNs leads to an increase in computation time during both training and testing. Therefore, we propose an ensemble containing only two CNNs with substantially different architectures. We hypothesize that these architectures lead to different errors, which will be evened out when combining segmentation results of both networks. The first CNN exploits a 3D network architecture, inspired by [9], and contains strided convolutional down- and upsampling layers and residual blocks. The second CNN exploits a 2D network architecture containing dilated convolutions [10, 11]. This network independently predicts voxel labels in axial, coronal and sagittal image slices, and obtains individual voxel predictions by averaging the three predictions. We compare the individual performance of each network and show improvement when both networks are combined into one ensemble.

2. DATA SET

We used data provided in the ISBI 2019 Segmentation of THoracic Organs at Risk in CT images (SegTHOR) challenge[1]. This data set contains 60 thoracic CT scans of patients diagnosed with non small cell lung cancer and referred for curative-intent radiotherapy. CT scans were acquired with or without intravenous contrast. Scans have an in-plane res-

[1]https://competitions.codalab.org/competitions/21012

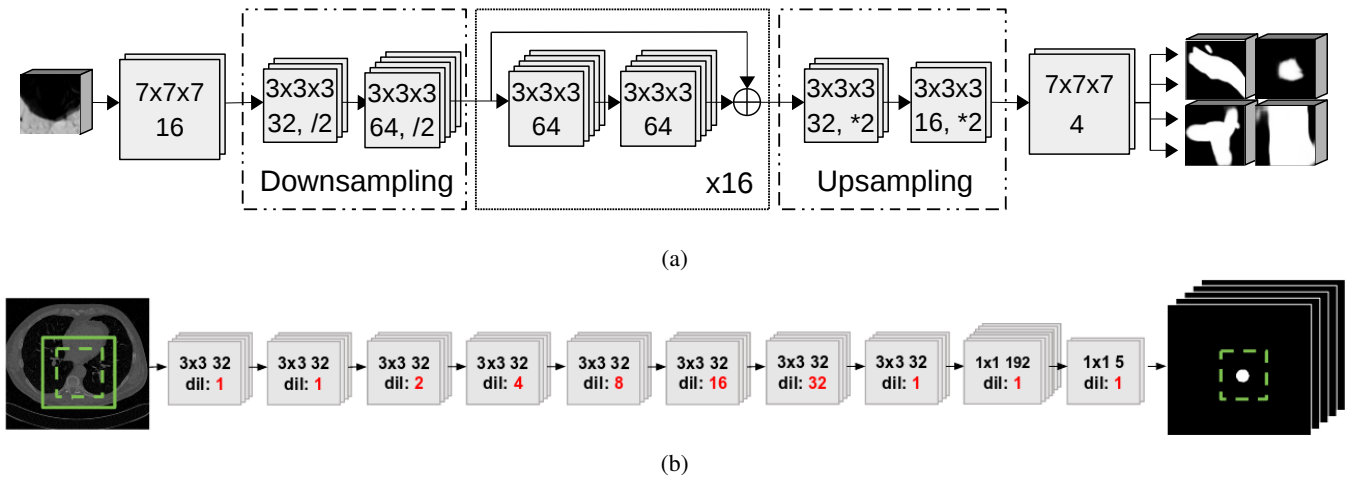

Fig. 1: CNN architectures used in our experiments. Numbers in boxes denote kernel size and number of feature maps in each layer. (a) A fully convolutional 3D network performing *multi-label* segmentation. The network contains downsampling layers, residual blocks, and upsampling layers. It has four output channels preceded by sigmoid functions: one for each of the target classes. (b) A ten-layer 2D CNN with increasing dilation levels performing *multi-class* segmentation. The output layer is followed by a softmax function and contains five output nodes: one for every foreground class, and one for the background.

olution varying between 0.90 mm and 1.37 mm, and a slice-thickness between 2 mm and 3.7 mm.

For each scan, manual reference delineations of the esophagus, heart, trachea, and aorta were available. The esophagus was delineated from the 4th cervical vertebra (C4) to the esophago-gastric junction. The heart was delineated as recommended by the Radiation Therapy Oncology Group 2. The trachea was delineated from the lower limit of the larynx to 2 cm below the carina, excluding the lobar bronchi. The aorta was delineated from its origin above the heart down to below the diaphragm pillars [6].

3. METHOD

We train two CNNs with substantially different architectures: one 3D network that contains a deep segment of residual blocks [9], and one 2D network containing dilated convolutions [10, 11]. The 3D network performs multi-label segmentation using sigmoids in the output layer, meaning the network can predict high probabilities for multiple classes in one voxel. The 2D network performs multi-class segmentation using a softmax in the output layer. This distinction is implemented to promote additional diversity in the networks.

The 3D network is a fully convolutional network using residual blocks, inspired by the 2D network used in [9]. The network analyses patches of $64\times64\times64$ voxel and produces an equally sized output. The deep segment of residual blocks allows the network to focus on highly detailed local geometry. The network contains two strided convolutional downsampling layers, followed by 16 fully pre-activated residual blocks [12], and two transposed convolutional upsampling layers. Rectified linear unit (ReLU) activation functions and batch normalization [13] are used in all layers, along with dropout [14] (p=0.5) in all of the residual blocks. An overview of the architecture is shown in Fig. 1a.The output layer contains four sigmoid functions that each identify presence of a single foreground class (esophagus, heart, trachea or aorta). To obtain one prediction per voxel, the class with the highest probability is chosen; background is selected when none of the class predictions exceed a probability of 0.5.

The second network is a 2D fully convolutional network containing dilated convolutions (Fig. 1b) [10, 11]. Segmentation of large anatomical structures with homogeneous textures as in CT can benefit from long-range context information [15]. Dilated convolutions allow large receptive fields while exploiting the input resolution of the image throughout the network. The network contains ten convolutional layers with increasing levels of dilation, leading to a receptive field of 131×131 voxels. The ReLU activation functions is used in all layers, along with dropout [14] (p=0.5) and batch normalization [13] on the fully connected layers.

During training, batches containing sub-images from the axial, coronal and sagittal plane are used. Sub-images have a size of 256×256 voxels (green square in Fig. 1b) of which the center 125×125 voxels (dashed green square in Fig. 1b) are classified by the network. During inference, the network is evaluated using all slices in the axial, coronal and sagittal direction, resulting in three 3D multi-class probability maps that are averaged to obtain a probability distribution per voxel. Each voxel is assigned the class with the highest resulting

Validation set (Dice)				
Method	**Esophagus**	**Heart**	**Trachea**	**Aorta**
3D CNN (Fig 1a)	0.83 ± 0.05	0.95 ± 0.01	0.90 ± 0.01	0.93 ± 0.01
2D CNN (Fig 1b)	0.82 ± 0.04	0.94 ± 0.02	0.90 ± 0.01	0.93 ± 0.01
2D + 3D	$\mathbf{0.85 \pm 0.05}$	$\mathbf{0.95 \pm 0.01}$	0.90 ± 0.01	$\mathbf{0.94 \pm 0.01}$
Test set (Dice)				
Method	**Esophagus**	**Heart**	**Trachea**	**Aorta**
Trullo et al. [6]*	0.67 ± 0.04	0.90 ± 0.01	0.82 ± 0.06	0.86 ± 0.05
2D + 3D	$\mathbf{0.84 \pm 0.05}$	$\mathbf{0.94 \pm 0.02}$	$\mathbf{0.91 \pm 0.02}$	$\mathbf{0.93 \pm 0.01}$
Test set (Hausdorff distance, mm)				
Method	**Esophagus**	**Heart**	**Trachea**	**Aorta**
2D + 3D	3.4 ± 2.3	2.0 ± 1.1	2.1 ± 1.0	2.7 ± 3.6

Table 1: Performance in terms of Dice coefficients and Hausdorff distances. The proposed method is compared with the performance reported by the state-of-the-art method by Trullo et al. [6]. *Note that results in [6] were obtained using a different data set, meaning the result comparison is only indicative.

probability.

Given that voxel classification may result in isolated (clusters of) voxels disconnected from the target structure, connected components smaller than 0.2 times the largest component in the class were removed using largest component selection.

4. EXPERIMENTS AND RESULTS

The dataset was split into a set of 40 scans and a test set of 20 scans by the challenge organisers. Delineations for the test set were not available during method development. The remaining set was randomly split into a training set (35 scans) and a validation set (5 scans) for the development of this work.

4.1. Training

Both networks were trained using Adam (learning rate=0.001). The 3D network was trained using a cross-entropy loss for 200,000 mini-batch iterations. Each mini-batch contained 15 pseudo-randomly sampled $64 \times 64 \times 64$ voxel sub-images, balanced such that each foreground class appears in at least one fourth of all training patches. The 2D network was trained using a Dice loss for 50,000 mini-batch iterations. Each mini-batch contained 40 256×256 pixel sub-images. These were randomly sampled along the axial, sagittal, or coronal axis. Random rotation augmentations of up to 10 degrees were applied to each slice.

4.2. Results

Table 1 lists performance on the validation set and the test set. For evaluation on the validation set, we trained and evaluated the 3D and 2D networks separately. In addition, we combined the probabilistic outputs provided by both networks prior to thresholding and largest component selection.

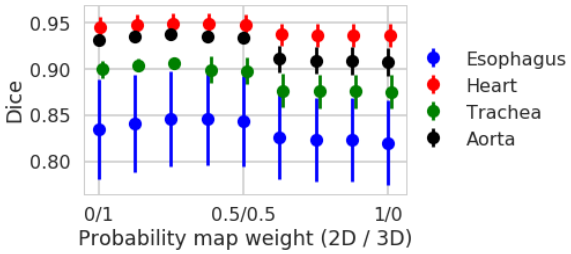

Fig. 2: Combined network performance on the validation set for varying weighting coefficients. Leftmost samples are equivalent to the individual 3D network, rightmost samples are equivalent to the individual 2D network.

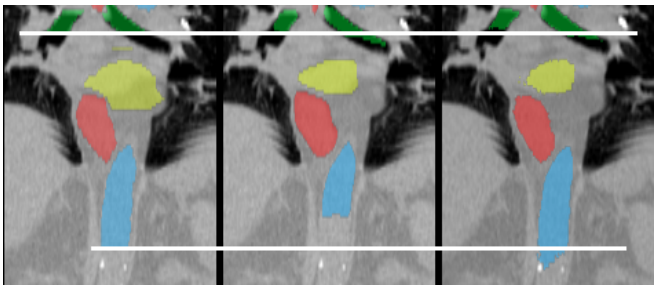

Fig. 3: From left to right: ground truth, 3D network result, combined network result. White lines denote segmentation boundaries of the trachea and aorta in the ground truth.

We investigated the effect of weighting the contribution of each network in the ensemble on the performance on the validation set. Fig. 2 shows the Dice score for each class resulting from different linear combinations of the probability maps from both networks. While the left-middle section of this figure shows an improvement in performance, the performance gain is numerically inconsequential compared to the 3D network results. However, qualitatively, some properties of the resulting segmentation visibly improve. An example is shown in Fig. 3, where the 3D network has difficulties detecting locally arbitrary borders of organ segmentations as appear in the bottom of aorta, the top of the esophagus and on both sides of the trachea. In this case, the information from the larger receptive field in the 2D network improves the combined network performance.

Fig. 2 suggests that the best performance is achieved when using a combination of both networks where the probability maps from the 3D network are weighted slightly stronger than those from the 2D network. Several combinations were submitted and scored in the challenge interface and a combination of 63.5% 3D and 37.5% 2D performed best on the test set by a small margin. The results listed as 2D+3D in Table 1 correspond to this combination.

5. DISCUSSION AND CONCLUSION

Even though both networks individually are able to accurately segment the organs at risk, we have shown that combining the predictions of both networks further improves segmentation performance. Considering prior work has shown that ensembles generally outperform individual networks [8], these results are as expected. The historical success of large ensembles implies that segmentation accuracy of this method could be further improved by adding additional networks. Notable here is the size of the presented networks: the 3D and 2D architectures contain only 3.7 million and 73 thousand parameters respectively. The small computational footprints mean both architectures can be attractive additions to a larger segmentation ensemble, even in computationally limited situations.

We have presented a method for automatic segmentation of organs at risk in thoracic radiotherapy treatment planning CT scans. The segmentation was performed using a combination of 2D and 3D convolutional neural networks. The results show the method achieves accurate segmentations, demonstrating potential for automating segmentation of organs at risk in routine radiotherapy treatment planning.

6. REFERENCES

[1] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, J. Lortet-Tieulent, and A. Jemal, "Global cancer statistics, 2012," *CA: a cancer journal for clinicians*, vol. 65, no. 2, pp. 87–108, 2015.

[2] K. D. Miller, R. L. Siegel, C. C. Lin, A. B. Mariotto, J. L. Kramer, J. H. Rowland, K. D. Stein, R. Alteri, and A. Jemal, "Cancer treatment and survivorship statistics, 2016," *CA: a cancer journal for clinicians*, vol. 66, no. 4, pp. 271–289, 2016.

[3] D. Thorwarth, "Functional imaging for radiotherapy treatment planning: Current status and future directionsa review," *The British journal of radiology*, vol. 88, no. 1051, p. 20150056, 2015.

[4] M. Han, J. Ma, Y. Li, M. Li, Y. Song, and Q. Li, "Segmentation of organs at risk in CT volumes of head, thorax, abdomen, and pelvis," in *Medical Imaging 2015: Image Processing*, vol. 9413. International Society for Optics and Photonics, 2015, p. 94133J.

[5] E. Schreibmann, D. M. Marcus, and T. Fox, "Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search," *Journal of applied clinical medical physics*, vol. 15, no. 4, pp. 22–38, 2014.

[6] 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 *14th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2017, pp. 1003–1006.

[7] K. Men, J. Dai, and Y. Li, "Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks," *Medical physics*, vol. 44, no. 12, pp. 6377–6389, 2017.

[8] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert *et al.*, "Ensembles of multiple models and architectures for robust brain tumour segmentation," in *International MICCAI Brainlesion Workshop*. Springer, 2017, pp. 450–462.

[9] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Computer Vision ECCV 2016*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Springer International Publishing, pp. 694–711.

[10] J. M. Wolterink, T. Leiner, M. A. Viergever, and I. Išgum, "Dilated convolutional neural networks for cardiovascular MR segmentation in congenital heart disease," in *International Workshop on Reconstruction and Analysis of Moving Body Organs*, ser. LNCS, vol. 10129. Springer, 2017, pp. 95–102.

[11] J. M. Noothout, B. D. de Vos, J. M. Wolterink, and I. Išgum, "Automatic segmentation of thoracic aorta segments in low-dose chest CT," in *Medical Imaging 2018: Image Processing*, vol. 10574. International Society for Optics and Photonics, 2018, p. 105741S.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks." [Online]. Available: http://arxiv.org/abs/1603.05027

[13] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 37. PMLR, 2015, pp. 448–456.

[14] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J Mach Learn Res*, vol. 15, no. 1, pp. 1929–1958, 2014.

[15] M. P. Heinrich, O. Oktay, and N. Bouteldja, "Obelisknet: Fewer layers to solve 3d multi-organ segmentation with sparse deformable convolutions," *Medical image analysis*, vol. 54, pp. 1–9, 2019.