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

Louis D. van Harten¹, Julia M. H. Noothout¹, Joost J. C. Verhoeff², Jelmer M. Wolterink¹, Ivana Išgum¹

¹Image Sciences Institute, University Medical Center Utrecht, Utrecht, The Netherlands
²Department 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-
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.

3. METHOD

We train two CNNs with substantially different architectures: one 3D network that contains a deep segment of 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 $131 \times 131$ voxels. The ReLU activation functions are 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 $256 \times 256$ voxels (green square in Fig. 1b) of which the center $125 \times 125$ 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 probability.
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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Validation set (Dice)</th>
<th>Test set (Dice)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Esophagus</td>
<td>Heart</td>
</tr>
<tr>
<td>3D CNN (Fig 1a)</td>
<td>0.83 ± 0.05</td>
<td>0.95 ± 0.01</td>
</tr>
<tr>
<td>2D CNN (Fig 1b)</td>
<td>0.82 ± 0.04</td>
<td>0.94 ± 0.02</td>
</tr>
<tr>
<td>2D + 3D</td>
<td>0.85 ± 0.05</td>
<td>0.95 ± 0.01</td>
</tr>
</tbody>
</table>

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


