Segmentation of CT thoracic organs by multi-resolution VB-nets

Miaofei Han[1], Guang Yao[1], Wenhai Zhang[1], Guangrui Mu[1,2], Yiqiang Zhan[1], Xiang Zhou[1], Yaozong Gao[1]

[1]Shanghai United Imaging Intelligence Inc., Shanghai, China
[2]Biomedical Engineering Department, Southern Medical University, Guangzhou, China

ABSTRACT

Accurate segmentation of organs at risk (OARs) is a key step in image guided radiation therapy. In this work, we proposed multi-resolution 3D V-Net networks to automatically segment thoracic organs at risk in computed tomography (CT) images. Specifically, we adopt two resolutions and propose a customized V-Net model called VB-Net for both resolutions. The VB-Net model in the coarse resolution can robustly localize the organs, while the VB-Net model in the fine resolution can accurately refine the boundary of each organ. In the SegTHOR 2019 challenge, 40 CT scans with 4 thoracic organs (i.e., esophagus, heart, trachea and aorta) were used for training. We experimented with both single-class and multi-class Dice losses to train the networks. Our best results were obtained by averaging multiple models trained with single-class Dice loss. At the time of submission, our results rank the 1st for segmentation of all four OARs.

Index Terms**—**organs at risk, V-Net, single-class, multi-class, ensemble, segmentation

1. INTRODUCTION

In lung and esophageal cancer, radiation therapy is a treatment of choice [1]. In radiation treatment planning, the target tumor and nearby healthy organs named organs at risk (OARs) need to be carefully contoured in order to make a dose plan. Often the contouring step is manual. This step takes hours for radiation oncologists and suffers large inter- and intra- operator variability. For some organs (e.g. the esophagus), the segmentation is especially challenging: shape and position vary greatly between patients; the contours in CT images have low contrast, and can be absent [1].

In recent years, deep learning based methods have been widely used in medical image segmentation. Among them, U-Net [2] and V-Net [3] are the most popular ones. V-Net was proposed to combine the residual networks with U-Net. By doing so, V-Net encourages much smoother gradient flow, thus easier in optimization and convergence. We developed a customized V-Net called VB-Net to segment OARs and target tumors for radiation planning. Validated on both an internal dataset and this challenge dataset, the proposed VB-Net shows promising results in accuracy, speed and robustness.

2. METHOD

2.1 VB-Net for Accurate Organ Segmentation

V-Net was initially proposed to segment the prostate by training an end-to-end fully convolutional network on MRI [3]. V-Net is composed of two paths, the left contraction path is used to extract high-level context information by convolutions and down-samplings. The right expanding path uses skip connections to fuse high-level context information with fine-grained local information for precise boundary localization. By means of introducing residual function and skip connection, V-Net shows better segmentation accuracy compared with many classical CNNs.

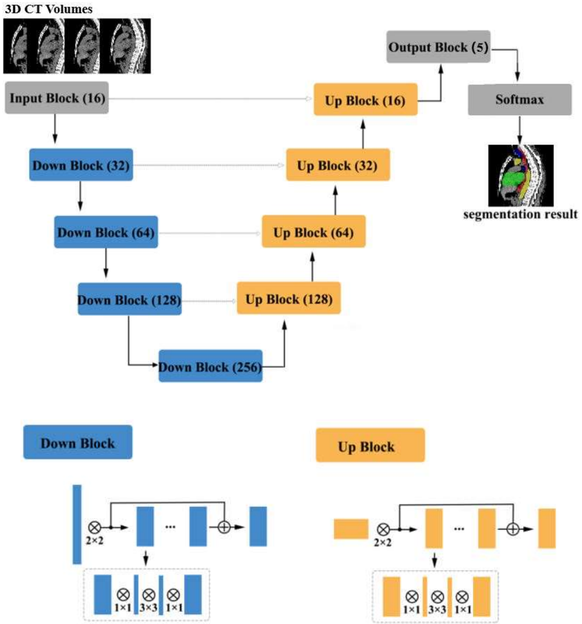

Fig. 1. The architecture of the proposed VB-Net.

The architecture of our proposed VB-Net is shown in Fig. 1. It replaces the conventional convolutional layers inside down block and up block with the bottleneck structure shown in the bottom of Fig. 1. Due to the use of bottle-neck structure, we named the architecture as VB-Net (B stands for bottle-neck). The bottleneck structure consists of three convolutional layers. The first convolutional layer applies a $1 \times 1 \times 1$ convolution kernel to reduce the channels of feature maps. The second convolutional layer performs a spatial convolution with the same kernel size as the conventional convolutional layer. The last convolutional layer applies a $1 \times 1 \times 1$ convolution kernel to increase the channels of feature maps back to the original size. By performing spatial convolutions on the feature maps with reduced channels, there are two benefits: 1) the model size is largely reduced, e.g., from V-Net (250 MB) to VB-Net (8.8MB); 2) the inference time is also reduced. In the following, we give a theoretical analysis of model sizes between a conventional convolutional layer and the corresponding bottleneck structure. The model size (parameter size) of a convolutional layer with kernel size K × K × K and input/output channel size C is K^3C^2. In contrast, the model size of the corresponding bottleneck structure is $\frac{C^2}{N} + \frac{K^3C^2}{N^2} + \frac{C^2}{N}$, where N is the ratio between original and reduced channel sizes. Given K = 3 in most settings, the model compression ratio using the bottleneck structure is $\frac{27 \times N^2}{2N+2}$, which mean 3.48 times model compression ratio when N = 2, 12.34 times model compression ratio when N = 4. In the following experiments, we adopt N = 4 that reduces V-Net from 250 MB to 8.8 MB. With a small model size of VB-Net, it becomes easy to deploy the segmentation network either to cloud or to the mobile applications.

2.2 Multi-resolution strategy

Many deep learning algorithms segment organs using a single resolution. As 3D medical images (e.g., CT, MR) are often large in size, e.g. $512 \times 512 \times 300$, passing the whole 3D image volume into networks will consume a lot of GPU memory, hence increasing the chances of segmentation failure due to lack of GPU memory. One solution is to resample the image volume into a lower resolution for segmentation, however, the image details will be lost in this way and the segmentation boundary will be zigzag. Another commonly used strategy is dividing the whole image volume into overlapping sub-volumes using a sliding window. However, this strategy is very time-consuming and not practical in industry deployment.

In this work, we adopt a multi-resolution strategy. Specifically, two VB-Nets are trained separately on different image resolutions. In the coarse resolution, we train a VB-Net to roughly localize the volume of interest (VOI) for each organ. The VB-Net is trained using resampled images at 6 mm voxel size. In the fine resolution, we train VB-Net to accurately delineate the organ boundary within the detected VOI.

2.3 Single/Multi-class loss function

There are many loss functions popular in segmentation methods, such as pixel-wise cross entropy loss, focal loss and the Dice loss [3], etc. The Dice loss has a clear advantage over pixel-wise cross entropy loss: it focuses only on foreground voxels disregarding how many the background voxels in the whole image. In this work, we adopt a generalized Dice loss function that applies to both single-class and multi-class segmentation problems. The mathematical formulation is given below:

$$\boldsymbol{D} = \frac{\mathbf{1}}{\boldsymbol{C}}\sum_{\boldsymbol{c}=\mathbf{1}}^{\boldsymbol{C}} \frac{\mathbf{2}\sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{p_c}(\boldsymbol{i})\boldsymbol{g_c}(\boldsymbol{i})}{\sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{p_c^2}(\boldsymbol{i}) + \sum_{\boldsymbol{i}}^{\boldsymbol{N}} \boldsymbol{g_c^2}(\boldsymbol{i})} \quad (1),$$

where the inner summation runs over the N voxels in the image domain, C represents the number of class labels, $p_c(i)$ is the probability of class c at voxel i predicted by the network, $g_c(i) \in \{0,1\}$ is the binary label indicating whether the label of voxel i is class c. The above formula becomes the single-class Dice loss function when $\mathrm{C} = 1$.

3. DATA AND EXPERIMENTS

3.1 Dataset

The experimental data is collected from SegTHOR 2019 training and testing data sets. The training and testing data include 40 and 20 patients, respectively. The CT scans have 512 x 512 pixels in-plane size with spatial resolution varying from 0.90 mm to 1.37 mm. The number of slices varies from 150 to 284 with a slice thickness between 2 mm and 3.7 mm. The most frequent spatial resolution is $0.98 \times 0.98 \times 2.5$ mm^3. We used the published 40 CT scans as training data, 20 CT scans as testing data and evaluate the segmentation accuracy using the online judgement.

3.2 Intensity normalization

In order to accelerate the convergence of neural network training, image intensities are first normalized. Based on the 4 thoracic organs to be segmented (i.e., heart, aorta, trachea and esophagus), we choose the mediastinal window for global intensity normalization, i.e., window level 40, window width 350. The minimum and maximum gray levels are -310 and 400, respectively. Intensity values between them are linearly normalized into the range $[-1,1]$. Intensities less than the minimum are set to -1 and those greater than the maximum are set to +1.

3.3 Patch-wise network training

The training images are resampled to isotropic resolutions and normalized first. In the coarse resolution, we resample images to 6mm isotropic spacing. To prevent label diminishing, we dilate the masks of esophagus and trachea by 10 mm in the original image space before resampling to the coarse resolution. In the fine resolution, we resample images to 1mm isotropic spacing without any mask dilation. After resampling, 3D sub-image volumes of size 96 voxel × 96 voxel × 96 voxel are randomly sampled as training crops. In the coarse resolution, we randomly sample sub-volumes from the entire image domain. In the fine resolution, we randomly sample sub-volumes only in the area indicated by the ground-truth mask. In this way, the fine-resolution network will focus more on the organ boundary than the coarse-resolution network. For each sampled image crop, the corresponding mask crop is extracted as the ground-truth mask, which is used as the network prediction target. With pairs of image and mask crops, we independently train segmentation networks for coarse-resolution and fine-resolution segmentation, respectively.

3.4 Fully convolutional network inference

In the inference phase, a multi-resolution strategy discussed in Section 2.2 is used to connect the coarse and fine resolution networks. The coarse-resolution network aims to roughly segment the organ, which is used to estimate a volume of interest (VOI). After that, a high-resolution image crop is resampled from VOI and the fine-resolution network is used to precisely segment the organ boundary.

Different from patch-wise network training, we perform fully convolutional network inference in the testing stage, where we fed the network with the entire image instead of overlapped image crops as is often done in many other works. The reasons that we are able to perform fully convolutional inference for 3D segmentation are two folds: 1) we adopt a multi-resolution strategy. In the coarse resolution, the entire image is resampled to 6mm isotropic spacing, which consumes only several hundred megabytes GPU memory. The same also applies to the fine-resolution segmentation, which focuses only on a sub-volume located by the coarse-resolution network. 2) Besides the multi-resolution strategy, another reason for being able to perform 3D fully convolutional inference is that we implement the inference engine from scratch, instead of using the open-source framework, such as pytorch, tensorflow. By doing so, we are able to optimize the runtime GPU memory specifically for the VB-Net, which gives about 75% GPU memory cost compared with the same implementation of pytorch. The details of GPU memory optimization is beyond the scope of this paper.

3.5 Post-processing

After both coarse- and fine-resolution segmentation, we remove noisy isolated segments by picking the largest 3D connected component. For esophagus segmentation, instead of picking the largest connected component in the fine-resolution, we pick the connected components with size > 500 voxels. This post-processing will take care of possible disconnections of esophagus segmentation due to its tubular structure. For heart segmentation, there may be small isolated segments in 2D slices. To remove them, we will pick the largest 2D connected component at each slice after the 3D one.

4. RESULTS

We validated our method on 20 CT scans of SegTHOR 2019 online. The SegTHOR 2019 competition uses the overlap Dice metric (DM) and the Hausdorff distance (HD) as the evaluation metrics. DM and HD are computed independently for each of the 4 organs at risk, obtaining 8 measurements and rank the performance independently for 8 measurements. The average of 8 ranks gives the final ranking. The SegTHOR 2019 ranking of all participating teams in the testing data is summarized by the organizer, where our team listed as "gaoking132" ranked 1 out of 44 teams.

4.1 Segmentation accuracy

The segmentation accuracy of the proposed method with multi-class, single-class Dice loss, and the ensemble of multiple models were evaluated online using 20 testing CT scans. In the ensemble model, we used three random seeds to train the segmentation network for each organ and average their results as the final segmentation result. Table 1 shows the Dice metrics of the three methods. With the single-class Dice loss, we can optimize network parameters separately for each organ, so the final segmentation accuracy of single-class Dice loss is higher than that of multi-class Dice loss. By averaging multiple single-class models, the results can be further improved, which is a common strategy used in many challenges.

Table 1. Dice metrics of multi-class, single-class Dice losses and the ensemble model on SegTHOR 2019 online testing set.

Method	**Dice metric**			
	Esophagus	Heart	Trachea	Aorta
Multi-class	0.8402	0.9446	0.9129	0.9388
Single-class	0.8605	0.9465	0.9172	0.9401
Ensemble	**0.8651**	**0.9536**	**0.9276**	**0.9464**

Table 2 shows the Hausdorff distance metrics of the three methods, respectively. The results also show that models trained with single-class Dice loss is better than that trained with multi-class Dice loss. With the model ensemble, the segmentation accuracy can be further boosted.

Table 2. Hausdorff distance metrics of multi-class, single-class Dice losses and the ensemble model on SegTHOR 2019 online testing set.

Method	**Hausdorff distance**			
	Esophagus	Heart	Trachea	Aorta
Multi-class	0.8189	0.1739	0.2123	0.2234
Single-class	0.2883	0.1630	0.2016	0.2124

Ensemble	**0.2590**	**0.1272**	**0.1453**	**0.1209**

4.2 GPU memory consumption

Besides accuracy, the proposed VB-Net and multi-resolution strategy have great benefits to reduce GPU memory for industry deployment. The statistics of GPU memory consumption of 20 CT scans are shown in Table 3.

Table 3. GPU memory consumption of the proposed multi-class and single-class VB-Net on SegTHOR 2019 online testing set. Note that the model ensemble doesn't increase the GPU memory cost.

Method	**Organ**	**Maxi. Memory**	**Avg. Memory**
Multi-class	4 organs	3885MB	3117.5MB
Single-class	Esophagus	1351MB	1120.1MB
	Trachea	1484MB	1150.0MB
	Aorta	1908MB	1325.8MB
	Heart	1621MB	1383.9MB

4.3 Segmentation runtime

Our single-class VB-Net segments an organ at average 0.76 second, while the multi-class VB-Net segments the 4 thoracic organs in 2 seconds. All timings were measured on an Intel Xeon CPU E5-2620 v4 with 12 GB memory and a NVIDIA Titan XP graphical card with 12 GB GPU memory. With the ensemble model, the segmentation runtime is linearly increased with the number of models used.

Table 4. Segmentation time of the proposed multi-class and single-class VB-Net on SegTHOR 2019 online testing set.

Method	**Organ**	**Avg. Segmentation Time**
Multi-class	4 organs	2.01s
Single-class	Esophagus	0.50s
	Trachea	0.86s
	Aorta	0.97s
	Heart	0.72s

5. CONCLUSIONS

In conclusion, we propose a multi-resolution VB-Net framework to segment 4 thoracic organs. The multi-resolution strategy reduces the GPU memory cost while maintains a high segmentation accuracy. The experimental results on SegTHOR 2019 online testing set show the superiority of our method for segmentation of esophagus, heart, trachea and aorta, respectively. Besides segmentation accuracy, we also evaluated the GPU memory consumption and segmentation runtime of our method. The results show that our method can accurately segment these organs with a small memory footprint and in a fast speed.

6. REFERENCES

[1] Roger Trullo, Caroline Petitjean, Su Ruan, Bernard Dubray, Dong Nie, and Dinggang Shen. "Segmentation of organs at risk in thoracic CT images using a sharpmask architecture and conditional random fields". In IEEE International Symposium on Biomedical Imaging (ISBI), pp 1003-1006, 2017.

[2] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In international Conference on Medical image computing and computer-assisted intervention (MICCAI), pp 234-241, 2015.

[3] Milletari F , Navab N , Ahmadi S A . V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In fourth International Conference on 3D Vision (3DV), pp 565-571, 2016.