MULTI-TASK LEARNING FOR THE SEGMENTATION OF THORACIC ORGANS AT RISK IN CT IMAGES

Tao He, Jixiang Guo, Jianyong Wang, Xiuyuan Xu and Zhang Yi

Machine Intelligence Laboratory, Sichuan University

ABSTRACT

The automatic segmentation of thoracic organs has clinical significance. In this paper, we develop the U-Net architecture and obtain a uniform U-like encoder-decoder segmentation architecture for the segmentation of thoracic organs. The encoder part of this architecture could directly involve the widely used popular networks (DenseNet or ResNet) by omitting their last linear connection layers. In our observation, we find out that individual organs could not appear independently in one CT slice. Therefore, we empirically propose to use the multi-task learning for the segmentation of thoracic organs. The major task focuses on the local pixel-wise segmentation and the auxiliary task focuses on the global slice classification. There are two merits of the multi-task learning. Firstly, the auxiliary task could improve the generalization performance by concurrently learning with the main task. Secondly, the predicted accuracy of the auxiliary task could achieve almost 98% on the validation set, so the predictions of the auxiliary task could be used to filter the false positive segmentation results. The proposed method was test on the Segmentation of THoracic Organs at Risk (SegTHOR) challenge (submitted name: MILab, till March 21, 2019, 8:44 a.m. UTC) and achieved the second place by the "All" rank and the second place by the "Esophagus" rank, respectively.

Index Terms— Automatic segmentation, CT, U-Net, Multi-task learning

1. INTRODUCTION

The contrast-enhanced Computed Tomography (CT) is the widely used clinical tool for diagnosing plenty of thoracic diseases. The drab and boring manual segmentation of thoracic organs from CT images is very time-consuming. The automatic segmentation from CT images will be helpful for oncologists to diagnose the thoracic organs at risk in CT images. In this paper, we focus on the automatic augmentation of thoracic organs data, supported by the Segmentation of THoracic Organs at Risk (SegTHOR) [1] challenge. The segmentation task is challenging for following reasons: (1) the shape and position of each organ on CT slices vary greatly between patients; (2) the contours in CT images have low contrast, and can be absent. The challenge focuses on 4 organs as risk: heart, aorta, trachea, esophagus.

Recently, the developments of the automatic segmentation based on deep learning have overthrown the traditional feature extraction methods. The paragon of medical segmentation models is U-Net [2]. U-Net has carefully designed encoder and decoder parts with shortcut connections. The most significant advantage of shortcut connections is to combine low-level features with high-level features at different layers. Recent years, many similar models termed as encoder-decoder architectures were proposed, for example, Seg-Net [3] and DeepLab series networks [4, 5].

In [6], an H-DenseUNet was proposed for liver and tumor segmentation, where intra-slice and inter-slice features were extracted and jointly optimized through the hybrid feature fusion layer. In [7], a 3D Deeply Supervised Network (3D-DSN) was proposed to address the liver segmentation problem. The 3D-DSN involved additional supervision injected into hidden layers to counteract the adverse effects of gradient vanishing. This method achieved the state of the art on the MICCAI-SLiver07 dataset. V-Net [8] is much like the 3D version of U-Net, which was directly applied in volumetric segmentation from MRI volumes depicting prostate.

Honestly, the 3D-CNNs based models fully exploit the space relative features but training a 3D-CNNs based model is usually time-consuming and requires large hyperparameters capacity. Therefore, many previous works employed 2D-CNNs and trained them on 2.5D data, which consisted of a stack of adjacent slices as input. Then the liver lesion regions were predicted according to the center slice. In order to achieve the accurate segmentation results of 2D-CNNs, authors of [9] proposed a two-step segmentation framework. At the first step, an FCN was trained to segment liver as ROI input for the second FCN. The second FCN solely segmented lesions from the predicted liver ROIs of step 1. The two-step segmentation framework has been widely involved in many segmentation works [9, 6, 10].

In this paper, we propose a uniform U-like encoder-decoder segmentation architecture. The previous U-Net and

This work was supported by the National Natural Science Foundation of China under Grant 61432012. The authors are with the College of Computer Science, Sichuan University, Chengdu 610065, China (e-mail: taohe@stu.scu.edu.cn; guojixiang@scu.edu.cn; wjy@scu.edu.cn; xuxiuyuan@stu.scu.edu.cn; zhangyi@scu.edu.cn).

Slice Index	1	…	50	51	…	68	69	…	80	81	…	118	119	120	…	156	157	…	182	183	…
Esophagus	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N
Heart	N	N	N	N	N	N	N	N	N	Y	Y	Y	Y	N	N	N	N	N	N	N	N
Trachea	N	N	N	N	N	N	N	N	N	N	N	N	Y	Y	Y	Y	Y	Y	Y	N	N
Aorta	N	N	N	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	N	N	N	N	N

Fig. 1. The macro view of Patient01's CT slices. 'Y' (YES) and 'N' (NO) indicate whether the corresponding organ appears in i^{th} column. From 51^{th} to 68^{th} slices, only aorta appears; from 68^{th} to 80^{th} slices, esophagus appears; from 81^{th} to 118^{th} slices, heart appears; in the 119^{th} slice, all organs appear; from 120^{th} to 156^{th} slices, heart disappears; from 157^{th} to 182^{th} slices, aorta disappears; from 183^{th} slice to the end, all organs disappear.

its variants usually have symmetrical encoder and decoder parts. In the uniform U-like architecture, the encoder part could directly involves the widely used popular networks (ResNet or DenseNet) by omitting their last linear connection layers. The encoder has more no-linear mapping ability and could adopts the transfer learning by initializing its parameters with the popular networks trained on image classification. The decoder part only works on enlarging the size of feature maps and shrinking the channel of networks. The uniform U-like architecture is trained under the multi-task learning scheme. The major task focuses on the local pixel-wise segmentation and the auxiliary task focuses on the global slice classification. There are two merits of multi-task learning. Firstly, the auxiliary task could improve the generalization performance by concurrently learning with the main task. Secondly, the predictions of the auxiliary task are used for filtering the false positive segmentation results.

2. METHOD

In this section, we will introduce the multi-task learning scheme and the uniform U-like encoder-decoder architecture.

2.1. Multi-task Learning

During the automatic segmentation of thoracic organs on the SegTHOR challenge data, we found out that individual organs could not appear independently in one slice. In Fig. (1), we give the detailed macro view of Patient01's CT slices. All patients have similar macro appearance orders. In other words, the organs appear dependently. If we could learn the macro classification, we could use the classification results to filter the false positive predictions of each organ. It will be much more valuable since the organs appear dependently. We apply the multi-task learning scheme to concurrently learn the segmentation and classification tasks. The formulation of learning is as follow:

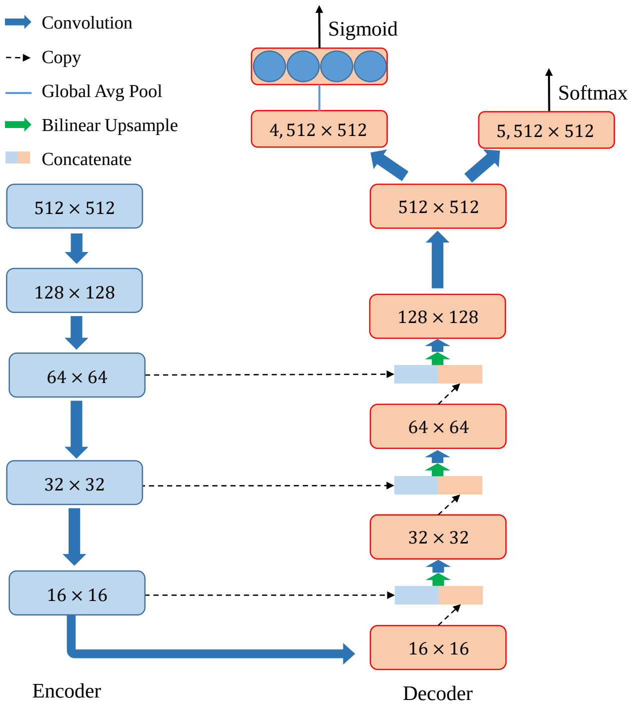

Fig. 2. The uniform U-like encoder-decoder architecture with multi-task learning, where the blue arrow indicates a convolutional layer, the dashed line indicates a copy operation, the solid line indicates a global average pooling layer, the green arrow indicates a bilinear upsample and the combined dashed block indicates a concatenation operation.

$$
\begin{aligned}
D =& 1 - \frac{1}{K_s} \sum_{k=1}^{K_s} \frac{2 \cdot \sum_{ij} p_{ij}^k \cdot g_{ij}^k}{\sum_{ij} (p_{ij}^k)^2 + \sum_{ij} (g_{ij}^k)^2} \\
&+ \alpha \cdot \sum_{k=1}^{K_c} (h^k \cdot \log q^k + (1 - h^k) \cdot \log(1 - q^k)). \quad (1)
\end{aligned}
$$

where $K_s = 5$ and $K_c = 4$, which indicate the number of segmentation and classification categories, respectively. D is the combined cost function. The major segmentation task is trained with dice loss and the auxiliary classification task is trained with multi-label logistic regression. In the dice loss part, p_{ij}^k and g_{ij}^k are the kth output produced by a softmax function and the kth one-hot target of pixel (i, j), respectively. In the multi-label logistic regression part, q^k and h^k are the kth output produced by the corresponding logistic function and the kth target, respectively. α is used to balance the loss. In our experiments, we set $\alpha = 0.5$.

2.2. Uniform U-like Encoder-Decoder Architecture

In most segmentation tasks, manually labelling is time-consuming, therefore the train sets are always restrained. Transfer learning is a very useful strategy to train a network on a small data set. In order to apply the transfer learning in the SegTHOR challenge, we abstract a uniform U-like encoder-decoder architecture, where the encoder part could directly involve the widely used ResNet or DenseNet by omitting their last linear connection layers. The encoder part could adopt the transfer learning by initializing the encoders parameters with the corresponding networks trained on image classification. The decoder part only works on enlarging the size of feature maps and shrinking the channel of networks. The U-like architecture is depicted in Fig. (2).

3. EXPERIMENT

There are 40 and 20 3D abdominal CT scans for training and testing on the SegTHOR Challenge dataset, respectively. We randomly split the given 40 training CT volumes into 32 for training and 8 for validation. The 3D CT scans were cut into slices along z-axis. Under the architecture of the uniform U-like architecture, the encoder part is free for setting. We implemented 6 widely used networks as the encoder part including **ResNet-101**, **ResNet-152**, **DenseNet-121**, **DenseNet-161**, **DenseNet-169** and **DenseNet-201**. The decoder part of them only involved one convolutional layer to shrink the number of channels.

The training of networks stopped when the dice per case of the validation set did not grow during 10 epochs. In order to fully use the given data, we then reloaded the trained model and retrained it on the full 40 slices for fixed 10 epochs. All networks were implemented by Pytorch [11] and trained using the stochastic gradient descent with momentum of 0.9. All networks were trained on the images with the original resolution and in form of 2.5D data, which consists of 3 adjacent axial slices. The image intensity values of all scans were truncated to the range of $[-128, 384]$ HU to omit the irrelevant information. The initial learning rate was 0.01 and decayed by multiplying 0.9. For data augmentation, we adopted random horizontal and vertical flipping and scaling between 0.6 and 1 to alleviate the overfitting problem. The networks were trained using four NVIDIA Titan Xp GPUs and it took about $6 \sim 8$ hours. After each testing, we used a largest connected component labeling to refine the segmentation results of each organ. The final submitted result is the ensemble result of those 6 U-like networks. The experimental results are listed on Table 1. We achieved the second place in the "All" rank order and the second place in the "Esophagus" rank order, respectively.

4. CONCLUSION

The uniform U-like architecture is abstracted from the widely used U-Net. The encoder part of the uniform U-like architecture is free for setting different network structures and the transfer learning is easy to be applied in this design. In our experimental observation, the transfer learning accelerated the training of those networks and boosted the performance of them. The multi-task learning is helpful on discovering organs' dependence. However, we did not analyze its advantages because of the time limit of the challenge.

We need to emphasize the fact that the connected component labeling is very useful for the SegTHOR challenge since all organs are indivisible and our method was based on the 2D-CNNs. Since the given CT is not enough for the SegTHOR data set compared with other segmentation tasks, the trained networks were easy to overfit. Therefore, the ensemble strategy is also very necessary for the SegTHOR challenge.

5. REFERENCES

[1] Roger Trullo, C. 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 14th International Symposium on Biomedical Imaging (ISBI)*, 2017, pp. 1003–1006.

[2] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015, pp. 234–241.

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[4] 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*, vol. 40, no. 4, pp. 834–848, 2018.

Table 1. The experiment results and ranks.

User	Rank		Dice				Hausdorff			
	All	Esophagus	Esophagus	Heart	Trachea	Aorta	Esophagus	Heart	Trachea	Aorta
MILab	2.75	2	0.8594	0.9500	0.9201	0.9484	0.2743	0.1383	0.1824	0.1129

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *European Conference on Computer Vision (ECCV)*, 2018, pp. 833–851.

[6] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng, “H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from CT volumes,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 12, pp. 2663–2674, 2018.

[7] Qi Dou, Hao Chen, Yueming Jin, Lequan Yu, Jing Qin, and Pheng-Ann Heng, “3d deeply supervised network for automatic liver segmentation from ct volumes,” in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 149–157.

[8] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” in *Proceedings of International Conference on 3D Vision*, 2016, pp. 565–571.

[9] Patrick Ferdinand Christ, Mohamed Ezzeldin A. Elshaer, Florian Ettlinger, Sunil Tatavarty, Marc Bickel, Patrick Bilic, Markus Rempfler, Marco Armbruster, Felix Hofmann, Melvin D’Anastasi, Wieland H. Sommer, Seyed-Ahmad Ahmadi, and Bjoern H. Menze, “Automatic liver and lesion segmentation in ct using cascaded fully convolutional neural networks and 3d conditional random fields,” in *Proceedings of Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2016, pp. 415–423.

[10] Yuyin Zhou, Lingxi Xie, Elliot K. Fishman, and Alan L. Yuille, “Deep supervision for pancreatic cyst segmentation in abdominal CT scans,” in *Proceedings of Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017, pp. 222–230.

[11] 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 *the Workshop of Conference on Neural Information Processing Systems (NIPS Workshop)*.