Torso organ segmentation in CT using fine-tuned 3D fully convolutional networks

Holger ROTH[*1], Ying YANG[*2], Masahiro ODA[*1], Hirohisa ODA[*2],

Yuichiro HAYASHI[*1], Natsuki SHIMIZU[*2], Takayuki KITASAKA[*3],

Michitaka FUJIWARA[*4], Kazunari MISAWA[*5], Kensaku MORI[*6,1]

[*1]Graduate School of Informatics, Nagoya University
[*2]Graduate School of Information Science, Nagoya University
[*3]Faculty of Information Science, Aichi Institute of Technology
[*4]Nagoya University Graduate School of Medicine
[*5]Department of Gastroenterological Surgery, Aichi Cancer Center Hospital
[*6]Information & Communications, Nagoya University

Abstract

3D fully convolutional networks (FCN) allow dense predictions in volumetric images. FCNs avoid handcrafting features or training organ-specific models, and features can be transferred across datasets. We trained a general model on a large set of CT scans with the major organs labeled and then fine-tuned to different classification tasks. Separate training, fine-tuning, and testing sets were used, including 331 CT scans with 7 abdominal labels for general training, a smaller set of only 20 CT scans but with substantially more labels (20 in total) for fine-tuning, and a completely unseen set of 10 torso CT scans for testing. We achieve state-of-the-art performance across these datasets, illustrating the generalizability and robustness of our models.

Keywords：deep learning, multi organ segmentation, computed tomography, fully convolutional networks

1. Introduction

Recent advances in 3D fully convolutional networks (FCN [1]) have made it feasible produce dense voxel-wise predictions on full volumetric images. FCNs avoid the need for handcrafting features or the training of organ-specific models. A second advantage is the ability of deep models to transfer learned features across dataset domains [2].

2. Methods

To this end, we trained a general FCN model employing the 3D U-Net architecture [3] on a large dataset of CT scans including some of major organ labels. This model can then be fine-tuned to other (smaller) datasets aiming at more detailed classification tasks or different field of views. In this work, we utilize separate training, fine-tuning, and testing datasets. The general training set consists of 331 clinical CT images with seven abdominal structures (artery, vein, liver, spleen, stomach, gallbladder, and pancreas) labeled. Our model and training approach are described in detail in [4]. Code and pre-trained model are available for download at [5].

We then fine-tune on a smaller dataset consisting only of 20 contrast enhanced CT images from the Visceral Challenge dataset [6], but with substantially more anatomical structures labeled in each image (20 in total). This fine-tuning process across different datasets is illustrated in Fig. 1 with some ground truth label examples used for training. In fine-tuning, we use a 10 times smaller learning rate. We furthermore test our models on a completely unseen data collection of 10 torso CT images with 8 labels, including organs that were not labelled in the original abdominal dataset, e.g. the kidneys and lungs. A probabilistic output prediction from our model is shown in Fig. 2.

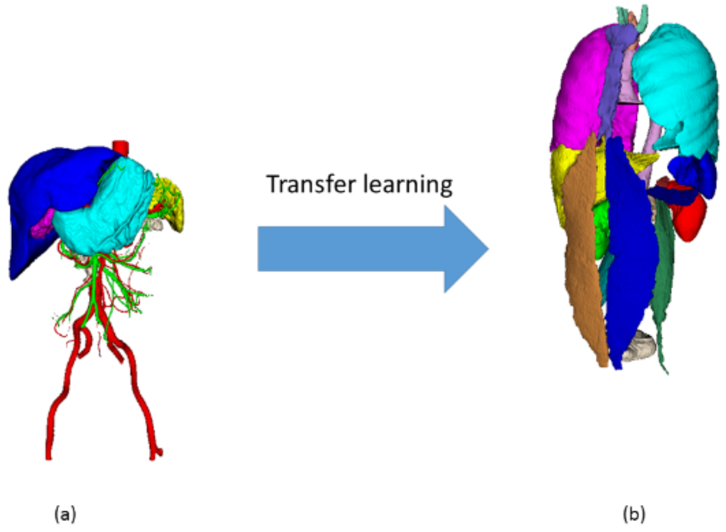

Fig. 1: We fine-tune our model via transfer learning from 8 anatomical structures in the abdomen (a) to 20 anatomical structures in the whole torso (b). We show some typical ground truth labels that are used for training on both datasets.

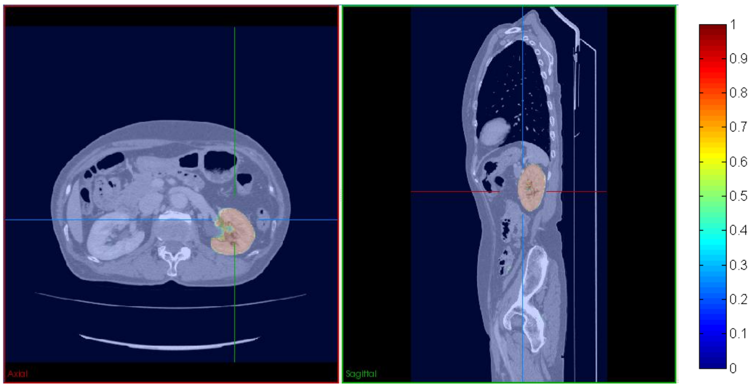

Fig. 2 Automated probability map for left kidney after transfer learning.

2. Results & Discussion

In testing, we deploy our fine-tuned model using a tiling approach as in [3]. An automated segmentation result on the unseen test dataset by our fine-tuned model is shown in Fig. 3. Our fine-tuned approach provides a Dice score of right lung, left lung, liver, gall bladder, spleen, right kidney, left kidney, and pancreas are 0.96, 0.97, 0.95, 0.77, 0.90, 0.90, 0.88, and 0.36, respectively (summarized in Table 1). The relatively lower score for pancreas is due to several outlier cases on this dataset. These outliers are likely caused by variations of contrast enhancement across the datasets and the higher variability of the pancreas' shape and intensity profile compared to other organs across different patients.

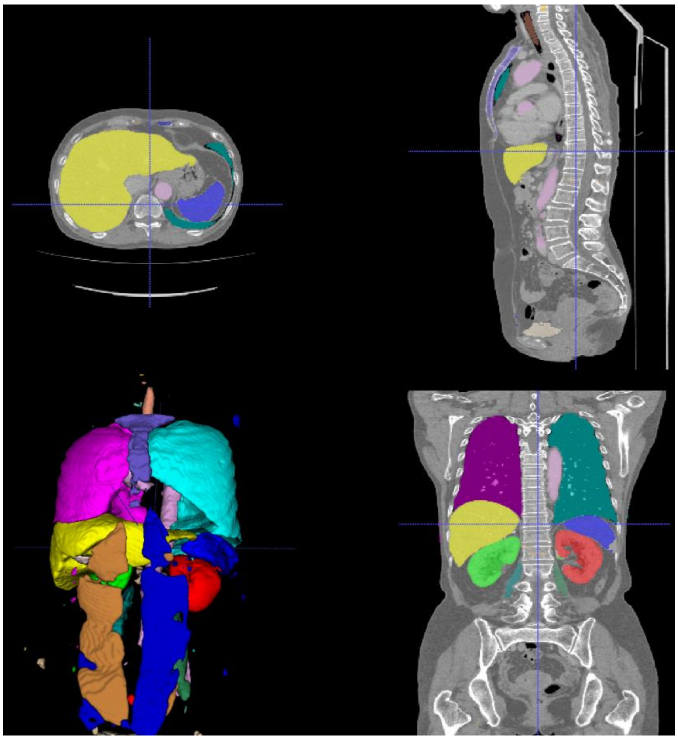

Fig. 3 Multi-organ segmentation result. Each color represents an organ region on the unseen test set.

Table 1 Dice scores for each segmented organ.

Case#	Right lung	Left Lung	Liver	Gall Bladder	Spleen	Right Kidney	Left Kidney	Pancreas
1	0.97	0.97	0.95	0.71	0.87	0.91	0.90	0.05
2	0.93	0.95	0.94	0.62	0.93	0.88	0.78	0.66
3	0.93	0.94	0.95	0.81	0.92	0.93	0.90	0.07
4	0.98	0.98	0.94	0.74	0.95	0.94	0.92	0.56
5	0.97	0.97	0.91	0.77	0.90	0.75	0.80	0.26
6	0.98	0.98	0.96	0.80	0.90	0.94	0.92	0.54
7	0.97	0.97	0.96	0.83	0.94	0.95	0.94	0.24
8	0.97	0.97	0.95	0.83	0.94	0.91	0.89	0.52
9	0.98	0.98	0.97	0.75	0.75	0.94	0.91	0.01
10	0.97	0.97	0.96	0.77	0.93	0.90	0.90	0.39
Avg.	**0.96**	**0.97**	**0.95**	**0.77**	**0.90**	**0.90**	**0.88**	**0.36**
Std. Dev.	**0.02**	**0.01**	**0.02**	**0.06**	**0.06**	**0.06**	**0.05**	**0.24**
Min.	**0.93**	**0.94**	**0.91**	**0.62**	**0.75**	**0.75**	**0.78**	**0.01**
Max	**0.98**	**0.98**	**0.97**	**0.83**	**0.95**	**0.95**	**0.94**	**0.66**

3. Conclusion

Our approach and results illustrate the generalizability and robustness of our models across different datasets. We have made our code and pre-trained models available for download at [5] in order to allow further fine-tuning to different datasets. Fine-tuning can be useful when the amount of training examples for some target organs are limited. In the future, prediction results from different models could be combined in order to achieve the best overall performance.

Acknowledgement

This paper was supported by MEXT KAKENHI (26108006, 26560255, 25242047, 17H00867, 15H01116) and JPSP International Bilateral Collaboration Grant.

References

[1] Long J, Shelhamer E, Darrell T (2015): Fully convolutional networks for semantic segmentation. In IEEE CVPR, pp. 3431--3440.

[2] Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning (2016). IEEE transactions on medical imaging; 35(5):1285-98.

[3] Cicek, O, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O (2016): 3D U-Net: learning dense volumetric segmentation from sparse annotation. In MICCAI, pp. 424--432. Springer.

[4] Roth HR, Oda H, Hayashi Y, Oda M, Shimizu N, Fujiwara M, Misawa K, Mori K (2017): Hierarchical 3D fully convolutional networks for multi-organ segmentation. arxiv preprint, https://arxiv.org/abs/1704.06382.

[5] https://github.com/holgerroth/3Dunet_abdomen_cascade

[6] Jimenez-del-Toro, O., Müller, H., Krenn, M., Gruenberg, K., Taha, A. A., Winterstein, M., ... & Kontokotsios, G. (2016). Cloud-based evaluation of anatomical structure segmentation and landmark detection algorithms: VISCERAL anatomy benchmarks. IEEE Transactions on Medical Imaging, 35(11), 2459-2475. (http://www.visceral.eu/benchmarks/anatomy3-open/)