JULY 12–16
2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

Weakly-Supervised Deep Learning Based Automatic Image Segmentation via Deformable Image Registration

Weicheng Chi, Weiguo Lu, Lin Ma, Junjie Wu, Xuejun Gu
Medical Artificial Intelligence and Automation (MAIA) Laboratory,
Department of Radiation Oncology, UT Southwestern Medical Center, Dallas, Texas
Weicheng.Chi@UTSouthwestern.edu

MAIA Laboratory
UT Southwestern Medical Center
Radiation Oncology

INTRODUCTION

The applications of deep learning (DL) in medical fields have gradually attracted researchers' attention and made significant progress, such as pathology diagnosis and organ delineation. To a certain extent, the great achievements of DL attribute to their large scale of training data. However, it is usually very time-consuming and costly to annotate such an enormous medical image dataset. Crowdsourcing would be a wise choice for most daily life applications, where it is easy to obtain annotations like objects' category or their bounding boxes. But annotating medical images requires clinical expertise from practicing doctors or radiologists to ensure its validity and reliability, which restricts the size of medical datasets.

Therefore, it is significant to develop DL approaches with limited labels, especially for medical image segmentation that requires voxel-wise annotations.

AIM

To alleviate the influence of limited labels in medical image segmentation, we intend to take advantage of abundant unlabeled images and propose a weakly-supervised deep learning model via deformable image registration (DIR) labeled-data generation.

METHODS

Fig. 1 schematically illustrates the proposed method: a) Contours generation: we use our in-house developed Demons based DIR algorithm [3] to generate pseudo-contours to augment training dataset; b) A recursive ensemble organ segmentation (REOS) framework [4] is trained with the augmented dataset. The training strategy is laid out as following. Assume we have N_{gt} images I_{gt} with ground-truth contours and N_{pse} images I_{pse} without contours. For each of N_{pse} case, we perform DIR with each of the ground truth case as an atlas, which will generate $N_{gt} * N_{pse}$ pseudo contours in total. Each DIR result from different ground truth contours is assumed as the true contour plus random noise and DL network is robust to these random noise [1, 2]. The loss functions of REOS network training can be modelled as $L = \sum_{i=1}^{N_{pse}} \sum_{k=1}^{N_{gt}} DICE\left(F\left(I_{pse}^{i}\right), \tilde{y}_{pse}^{i,k}\right)$, where $F(\cdot)$ is the output of REOS and $\tilde{y}_{pse}^{i,k}$ is the pseudo-contour deformed from the I_{gt}^{k} to the unlabeled image I_{pse}^{i}.

Our DL segmentation is compared to DIR-average and DIR-majority-voting, which calculates the mean or majority voting DC of DIR generated contours, respectively.

DATASETS

The developed approach is evaluated on a 31 head-and-neck CT image/contour dataset from The Cancer Imaging Archive (TCIA). The entire dataset is split into three sets: training (20), validation (4), and test (7). For each test image, 19 pseudo-contours are deformed from the others as moving images via DIR. Totally 380 (=20x19) generated pseudo-contour sets are used to train the REOS model. The trained DL model is evaluated on mandible, L&R parotid glands and L&R submandibular glands, using dice coefficient (DC) as evaluation metric.

Both images and corresponding masks are resampled to the CT element size of 1.17mm × 1.17mm × 3mm. Before conducting deformable registration for generating pseudo contours, all the images and their masks are aligned to a standard CT atlas and cropped to the volume size of 256×256×128.

METHODS

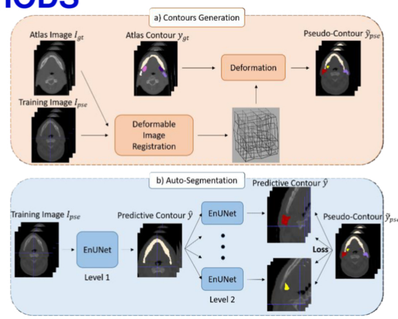

Fig.1. Illustration of the proposed method: a) pseudo-contour generation; b) auto-segmentation model trained with pseudo-contours.

RESULTS

Table 1 shows the comparative performance of dice coefficient between our method and DIR-average and DIR-majority-voting method. Our model achieves DCs of 86.2%±1.8%, 75.7%±3.9%, 72.2%±3.7%, 64.4%±6.1% and 59.9%±4.7% on mandible, L&R parotids and L&R submandibular glands, respectively. Our DL segmentation model outperforms DIR-average by a large margin over 12% on average. DIR generated contours extremely depend on the similarity between the atlas and test images. Therefore, DIR-average where segmentation relies on image similarity inevitably yields poor performance compared with our method. DIR-majority-voting refines the segmentation results by combining all contours from multiple atlases. As a result of its ability to exclude outliers, DIR-majority-voting obtains a 9.3% gain over DIR-average. However, our approach still surpasses DIR-majority-voting on all organs with a 2.7% improvement on average. Instead of using a fixed threshold (50%) in DIR-majority-voting, our model can automatically trade off the performance of different contours and learn the best segmentation standard according to the training set. And our model is less prone to overfitting the training data due to multiple coarse contours for the same image. What's more, our approach has lower variances on all organs, especially small organs such as right parotid gland and right submandibular gland. It demonstrates that our method is more stable in segmentation of difficult cases than the baselines.

Fig. 2 shows a qualitative comparison of segmentation results of a sample case TCGA-CV-5978 in the testing set. Rows 1-3 represent contours on different slides of the CT image volume. First, we can observe the discontinuous contour of mandible in (a)-2, manifesting inferior segmentation performance of traditional DIR algorithm. DIR-majority-voting also misses part of mandible in image (b)-2, which means that more than half of DIR-generated results cannot handle this difficult case well. However, our model doesn't merely memorize the training data and can generalize a reasonable contour comparable to the ground-truth contour. Second, as shown in (b)-1 and (b)-3, DIR-majority-voting cannot detect the boundary of mandible and submandibular glands properly while our model can better achieve it.

Experimental results indicates the effectivity and promising performance of the proposed method. Though the contours generated with a DIR approach are inaccurate for each image, our DL model can learn and summarize the potential ground-truth from multiple noisy contours. The contour dataset augmented to a large number can avoid the overfitting issue in DL model training.

Table 1: Comparisons of dice similarity coefficient (%) among 5 organs between our method and DIR methods

Method	Mandible	Left Parotid Gland	Right Parotid Gland	Left Submandibular	Right submandibular	Mean
DIR-Average	74.2±9.3	64.5±8.6	60.4±11.3	49.8±13.2	47.4±13.2	60.2
DIR-Majority-Voting	82.7±2.8	73.6±4.8	69.6±7.2	60.8±6.8	57.9±9.9	69.5
Ours	**86.2±1.8**	**75.7±3.9**	**72.2±3.7**	**64.4±6.1**	**59.9±4.7**	**72.2**

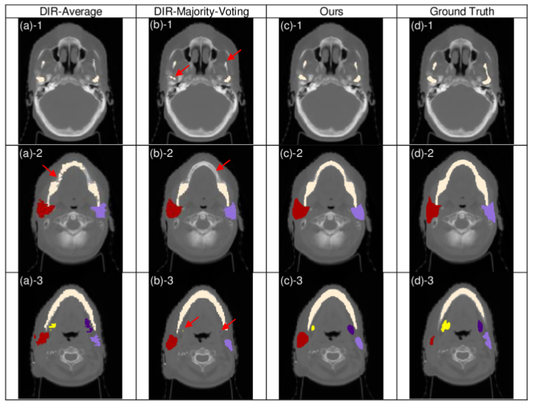

Fig. 2: Visualization segmentation performances of different methods on different slides of case TCGA-CV-5978 in the testing set. The regions in papaya whip, light purple, red, dark purple, yellow are mandible, left parotid, right parotid, left submandibular gland and right submandibular gland respectively.

CONCLUSIONS

We developed a weakly-supervised DL algorithm for segmentation on datasets with very limited labels. Experimental results on TCIA dataset demonstrate the weakly-supervised DL method outperforms traditional multi-atlas DIR methods and is promising for DL-based limited data medical image segmentation application. In future, we will expand on clinical available unlabeled image to generate training data, TCIA dataset with high-quality labels will be used as atlas set, validation set and test set. Moreover, we will explore the influence of the number of atlas images and unlabeled images on the proposed method, which provides the guideline for weakly-supervised deep learning based segmentation for clinical application.

REFERENCES

1. Rolnick D, Veit A, Belongie S, Shavit N. Deep learning is robust to massive label noise. arXiv preprint arXiv:1705.10694 (2017).
2. Krishnan S, Franklin MJ, Goldberg K, Wu E. Boostclean: Automated error detection and repair for machine learning. arXiv preprint arXiv:1711.01299 (2017).
3. Gu X, Pan H, Liang Y, et al. Implementation and evaluation of various demons deformable image registration algorithms on a GPU. Physics in Medicine & Biology. 2009; 55(1): 207-219.
4. Chen H, Lu W, Chen M, et al. A recursive ensemble organ segmentation (REOS) framework: application in brain radiotherapy. Physics in Medicine & Biology. 2019; 64(2): 025015.