Cycle training scheme for FLARE21 Challenges

Qiao DENG
The Chinese University of Hong Kong
Prince of Wales Hospital, Shatin, NT, Hong Kong
qiaodeng@link.cuhk.edu.hk

Abstract

Medical image segmentation plays a essential role in current clinical practice. Most scholars and reseachers are very concerned about the applications of deep learning in medical images.

Abdominal multi-organ segmentation is a popular branch and FLARE2021 challenge is a diverse abdominal organ dataset. In this challenge, we apply a tricky training scheme named cycle training that make three netwoks learn the good samples with high confidence and get the information interactive in order to get a great performance and to make the network more robust. And we can achieve satisfied performance in FLARE21 dataset.

1. Introduction

Medical image segmentation to delineate organs and regions of intrerest has been an essential component in radiation therapy. In current clinical practice, abdominal organ segmentation has many significant applications such as presurgical planning, organ donation and morphological and volumetric follow-ups for various diseases.

Regarding to most of abdominal organ dataset, they are more likely to be single-center and single-phase. But FLARE21 challenge dataset is more diverse, which contains 511 CT scans from different countries and imaging centers with multi-phase, multi-vendor, and multi-disease cases. Moreover, it is necessary to develop a efficientcy model without costing numerous computational resources in clinical practice.

Considering these challenges mentional above, we employed a tricky training scheme [1] called cycle training implemented on nnU-Net [2], which allows us to train deep networks with better samples. In our framework, cycle training could mantain three networks training simultaneously and each network could learn the lower uncertainty instances that selected by its peer two networks in each mini-batch data.

2. Method

As inspired by [1, 3], we used cycle training module to select small loss instances [1] as useful knowledge to its peer network for updating each its parameters. Figure 1 depicts the structure of cycle training scheme. We initialized three backbone networks [2] simultaneously, and each network can have information interaction in this training mode. Taking network A as an example, the network A could learn the good samples selected by committee composed by network B and network C. Similarly, the network B and C can also adapt advice suggested by its peer networks during training.

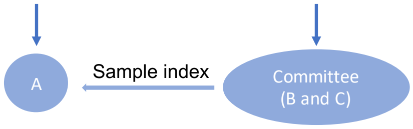

Figure 1. Cycle training scheme

2.1. Preprocessing

We implemented our method on the 2D nnU-Net [2], so we utilize identical preprocessing strategy that contains cropping, resampling method and intensity normalization.

2.2. Proposed Method

- In our framework, we chose the 2D nnU-Net [2] as our backbone network because it can automatically configures and achieve powerful performance. Figure 2 illustrates the applied 2D nnU-Net, which is based on the U-Net [4] architecture. And the hyper-parameters are set as defaulted ones.

- Our cycle training module can make every network learns low uncentainty cases in each mini-batch picked up by its respective committee. And each network only learns the useful instances to update its weight in the training processing. We initialize three 2D nnU-Net networks at the same time. Because the initialization is randomly different and the samples learned by each

Table 1. Data splits of FLARE2021.

Data Split	Center	Phase	# Num.
Training (361 cases)	The National Institutes of Health Clinical Center	portal venous phase	80
	Memorial Sloan Kettering Cancer Center	portal venous phase	281
Validation (50 cases)	Memorial Sloan Kettering Cancer Center	portal venous phase	5
	University of Minnesota	late arterial phase	25
	7 Medical Centers	various phases	20
Testing (100 cases)	Memorial Sloan Kettering Cancer Center	portal venous phase	5
	University of Minnesota	late arterial phase	25
	7 Medical Centers	various phases	20
	Nanjing University	various phases	50

network are different, each network has different learning ability. In selecting samples process, we exchange the small-loss difference instances to each network. The committee give an agreement on the instance in accordance with small-loss difference. The two networks in the committee has the different learning abilities and when they learn the same cases each network has a loss, and we do a simple substraction, sort the differences, select the 75% samples with smaller loss difference in each mini-batch,and transmit those cases index to the network that is not include in the committee. For each network, it has its own committe, so in this training mode, every network can get the interactive information from other two networks and then archieve cycle training.

- Loss function: In our method, we select small uncertainty samples according to the equation 1. μ denotes the good samples selected and ℓ_{CE} denotes cross entropy loss. And $\hat{y}, f_B, f_C$ denote the ground truth and the predictions of network B and network C.

$$\mu = argmin(|\ell_{CE}(\hat{y}, f_B) - \ell_{CE}(\hat{y}, f_C)|) \qquad (1)$$

- Number of model parameters: 41268192 (can be computed via such as torchsummary library for Pytorch)

- Number of flops: 590861472000 (can be computed via such as fvcore library for Pytorch)

2.3. Post-processing

In our experiment, we did not use any post-processing.

3. Dataset and Evaluation Metrics

3.1. Dataset

- We evaluate our method on the FLARE2021 dataset that contains 511 abdominal organ CT scans divided into 361 training cases, 50 validation cases and 100 testing cases. The detail information is presented in Table 1. In this challenge, we segmented the liver, kidney, spleen, and pancreas simultaneously. The FLARE2021 was prepared by MSD [5], NIHPancreas [6, 7, 8], and Nanjing University.

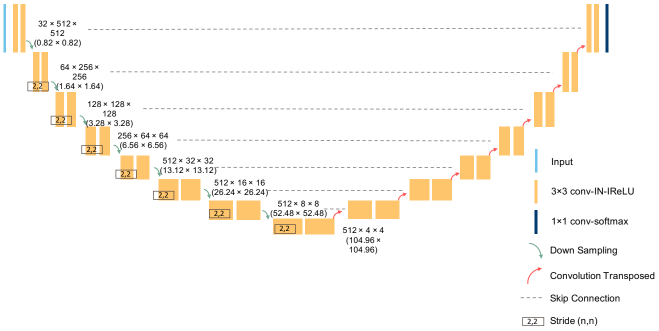

Figure 2. 2D nnU-Net architecture

3.2. Evaluation Metrics

- Dice Similarity Coefficient (DSC)
- Normalized Surface Distance (NSD)
- Running time
- Maximum used GPU memory (when the inference is stable)

4. Implementation Details

4.1. Environments and requirements

The environments and requirements of the baseline method is shown in Table 3.

4.2. Training protocols

The training protocols of the baseline method is shown in Table 4.

5. Results

5.1. Quantitative results for 5-fold cross validation.

The provided results analysis like [2] is based on the 5-fold cross validation results and the detail information is in

Table 2. Quantitative results of 5-fold cross validation in terms of DSC and NSD.

Training	Liver		Kidney		Spleen		Pancreas	
	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)	DSC (%)	NSD (%)
Fold-1	98.5±0.7	92.9±4.0	96.3±4.1	92.2±8.8	98.1±1.5	96.2±4.2	80.7±10.4	60.5±15.4
Fold-2	98.4±1.1	93.3±4.4	97.1±1.9	93.7±6.4	98.0±1.7	95.9±4.9	82.9±5.5	61.5±13.3
Fold-3	98.5±0.7	93.4±3.7	96.6±4.7	93.1±6.1	97.9±2.5	96.0±5.3	81.3±8.7	61.6±14.1
Fold-4	98.4±1.0	92.4±4.7	96.4±5.2	93.0±6.9	96.7±11.7	94.8±12.0	80.7±12.6	61.9±15.4
Fold-5	98.4±1.1	92.8±4.7	96.8±2.0	93.1±6.1	97.8±2.0	95.7±5.2	82.2±7.2	63.7±12.9
Average	98.4±0.92	93.0±4.3	96.6±3.6	93.0±6.9	97.7±3.88	95.7±6.3	81.6±8.9	61.8±14.2

Table 3. Environments and requirements.

Ubuntu version	Ubuntu 18.04.4 LTS
CPU	Intel Xenon E5-2698 v4 2.2GHz, 20 Cores X 2 AMD EPYC 7742 2.25GHz, 64 Cores 128 Threads X 2
RAM	4×256GB 8×2015GB
GPU	Tesla V100 Tesla A100
CUDA version	11.0
Programming language	Python3.8.5
Deep learning framework	Pytorch (Torch 1.7.1, torchvision 0.8.2)
Specification of dependencies	nnUNet
(Optional) code is publicly available at	None

Table 4. Training protocols.

Data augmentation methods	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast, simulation of low resolution, gamma correction and mirroring.
Initialization of the network	"he" normal initialization
Patch sampling strategy	More than a third of the samples in a batch contain at least one randomly chosen foreground class which is the same as nn-Unet [2].
Batch size	8
Patch size	512×512
Total epochs	1000
Optimizer	Stochastic gradient descent with nesterov momentum ($\mu = 0.99$)
Initial learning rate	0.01
Learning rate decay schedule	poly learning rate policy: $(1 - epoch/1000)^{0.9}$
Stopping criteria, and optimal model selection criteria	Stopping criterion is reaching the maximum number of epoch (1000).
Training time	72.5 hours
CO_2eq[1]	

Table 2. Table 2 illustrates the results of 5-fold cross validation. Figure 3 shows the evaluation metrics results of 5-fold cross validation for abdominal organs. According to the Table 2 and Figure 3, we can know that liver, kidney and spleen have great segmentation results in DSC and NSC metrics. For pancreas, it gets lower score in those two evaluation metrics than other organs mentioned before.

5.2. Qualitative results

Figure 4 presents some challenging examples. we can notice that our cross validation results have a good performance. we can segment liver, kidney, spleen with good performance using our method. For pancreas, the results show it is a hard job to do the segmentation.

6. Discussion and Conclusion

Our method can work well on cases. Although our results do not outperform than baseline, our method still achieve good performance and our training data format is 2D. We can do future research to improve our method.

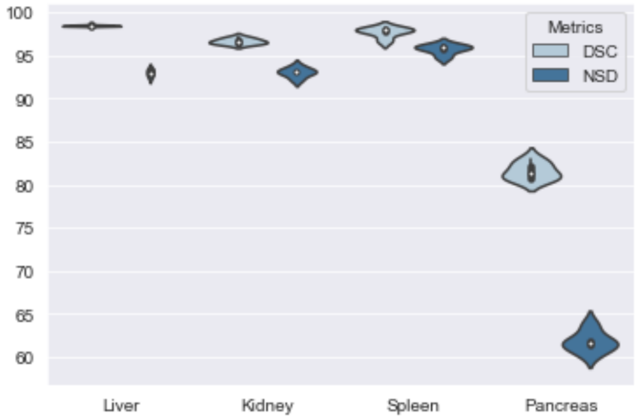

Figure 3. Violin plots of the organ segmentation results (DSC and NSD) of the 5-fold cross validation.

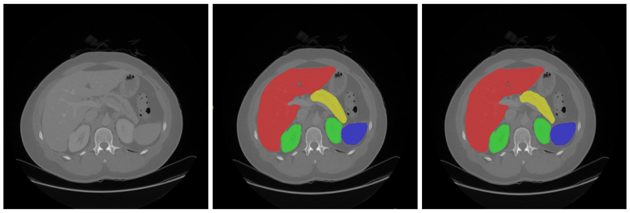

Figure 4. Challenging sample visualization.First column is the image, second column is the ground truth, and third column is the predicted results by our cycle training method.

Acknowledgment

None

References

[1] C. Xue, Q. Deng, X. Li, Q. Dou, and P.-A. Heng, “Cascaded robust learning at imperfect labels for chest x-ray segmentation,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 579–588. 1

[2] F. Isensee, P. F. Jaeger, S. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnu-net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods*, vol. 18, no. 2, pp. 203–211, 2021. 1, 2, 3

[3] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, “Co-teaching: Robust training of deep neural networks with extremely noisy labels,” *arXiv preprint arXiv:1804.06872*, 2018. 1

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241. 1

[5] A. L. Simpson, M. Antonelli, S. Bakas, M. Bilello, K. Farahani, B. Van Ginneken, A. Kopp-Schneider, B. A. Landman, G. Litjens, B. Menze *et al.*, “A large annotated medical image dataset for the development and evaluation of segmentation algorithms,” *arXiv preprint arXiv:1902.09063*, 2019. 2

[6] H. Roth, A. Farag, E. Turkbey, L. Lu, J. Liu, and R. Summers, “Data from pancreas-ct. the cancer imaging archive (2016).” 2

[7] H. R. Roth, L. Lu, A. Farag, H.-C. Shin, J. Liu, E. B. Turkbey, and R. M. Summers, “Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation,” in *International conference on medical image computing and computer-assisted intervention*. Springer, 2015, pp. 556–564. 2

[8] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle *et al.*, “The cancer imaging archive (tcia): maintaining and operating a public information repository,” *Journal of digital imaging*, vol. 26, no. 6, pp. 1045–1057, 2013. 2