A Common Try for MICCAI FLARE21 Challenges

Rian Huang
School of Biomedical Engineering, Health Science Center, Shenzhen University
Shenzhen, China
928456765@qq.com

Abstract

We provide a baseline implementation of the FLARE21 challenge based on 3D nnU-Net, which is the SOTA method of multi-organ segmentation. Due to the complexity of nnU-Net project, training and inference time are so long, which is not really suitable for FLARE21 challenge. Therefore, we propose a simplified version of nnU-Net, which includes data preprocessing, data augmentation, and training procedures similar to nnU-Net. It can maintain segmentation accuracy and greatly shorten the time-consuming training and inference.

1. Introduction

The first challenge comes from the diversity of the dataset, which including multi-center, multi-phase, multi-vendor, and multi-disease cases. Well generalization of the applied method is required. The second challenge comes from the efficiency requirement for the proposed solutions.

To addressing above difficulties, we proposed a simplified version of the 3D nnU-Net [1] as our solution as a result of the following reasons. First, it provided entire segmentation pipeline, including preprocessing, network architecture, training and post-processing. Second, it is fast. Third, it shows powerful performance on several segmentation tasks.

2. Method

A detail description of the method used, a schematic representation of the method is recommended.

Figure 1 illustrates the applied 3D nnU-Net, where a U-Net [2] architecture is adopted.

2.1. Preprocessing

Our method includes the following preprocessing steps:

- Cropping strategy: None.

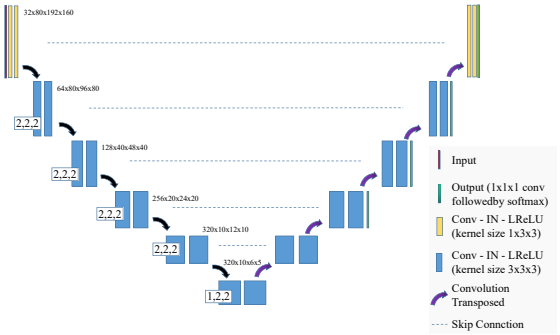

Figure 1. Network architecture

- Resampling spacing:
  We use the same way to calculate the target spacing as nnU-Net. And then we double the target spacing to reduce resolution to increase the receptive field.

- Resampling method for anisotropic data:
  In-plane with third-order spline interpolation, out-of-plane with nearest neighbor interpolation.

- Intensity normalization method:
  First, the dataset is clipped to the [0.5, 99.5] percentiles of the intensity values of the training dataset. Then a z-score normalization is applied based on the mean and standard deviation of the intensity values to each case.

2.2. Proposed Method

- Data augmentation: We use ScaleTransform, RandomRotateTransform, GammaTransform, BrightnessMultiplicativeTransform, ContrastAugmentationTransform, GaussianNoiseTransform, GaussianBlurTransform, RandomCrop.

- Network architecture details: we use the same network architecture as 3D nnU-Net [1, 2] is used as shown in Figure 1, all the hyper-parameters are set as the defaulted ones.

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

- Loss function: we use the summation between Dice loss and cross entropy loss because it has been proved to be robust [3] in medical image segmentation tasks. In addition, we use deep supervision. For each deep supervision output, a corresponding downsampled ground truth segmentation mask is used for loss computation.
- Number of model parameters: 30,787,584
- Number of flops: 590861472000

2.3. Post-processing

Restore to original resolution.

3. Dataset and Evaluation Metrics

3.1. Dataset

- A short description of the dataset used:
  The dataset used of FLARE2021 is adapted from MSD [4] (Liver [5], Spleen, Pancreas), NIH Pancreas [6, 7, 8], KiTS [9, 10], and Nanjing University under the license permission. For more detail information of the dataset, please refer to the challenge website and [11].
- Details of training / validation / testing splits:
  The total number of cases is 511. An approximate 70%/10%/20% train/validation/testing split is employed resulting in 361 training cases, 50 validation cases, and 100 testing cases. The detail information is presented in Table 1.

3.2. Evaluation Metrics

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

Table 2. Environments and requirements.

Windows/Ubuntu version	Ubuntu 16.04.1
CPU	Intel(R) Xeon(R) CPU E5-2620 v4 @ 2.10GHz
RAM	8×32GB; unknown MT/s
GPU	Nvidia TITAN XP
CUDA version	10.1
Programming language	Python3.7
Deep learning framework	Pytorch (Torch 1.7.1, torchvision 0.8.2)
Specification of dependencies	nnUNet
(Optional) code is publicly available at	None

Table 3. Training protocols.

Data augmentation methods	Rotations, scaling, Gaussian noise, Gaussian blur, brightness, contrast and gamma correction.
Initialization of the network	"he" normal initialization
Patch sampling strategy	random crop
Batch size	2
Patch size	192×160×80
Total epochs	1000
Optimizer	Adam
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 hours
CO_2eq[1]	

Table 4. Quantitative results of validation in terms of DSC and NSD.

Liver		Kidney		Spleen		Pancreas	
DSC (%)	NSD(%)	DSC (%)	NSD(%)	DSC (%)	NSD(%)	DSC (%)	NSD(%)
98.09	93.60	95.50	91.37	97.19	96.26	81.57	62.65

4. Implementation Details

4.1. Environments and requirements

The environments and requirements of our proposed method is shown in Table 2.

4.2. Training protocols

The training protocols of the baseline method is shown in Table 3.

4.3. Testing protocols

- Pre-processing steps of the network inputs:
  The same strategy is applied as trainging steps.
- Post-processing steps of the network outputs:
  Restore the original resolution through nearest neighbor interpolation.
- If using patch-based strategy, describing the patch aggregation method:
  We use normal sliding windows strategy.

5. Results

5.1. Quantitative results for validation.

The provided results analysis is based on randomly selecting 20% data from the training set as the validation set. Table 4 illustrates the results of validation.

5.2. Quantitative results on validation set.

Table 5 illustrates the results on validation cases. It is worth pointing out that for liver segmentation, the DSC scores are 95.3% , indicating great segmentation performance in terms of region overlap between the ground truth and the segmented region. NSD values are 81.5% demonstrating that the boundary regions contain more segmentation errors, which need further improvements [11].

Table 5. Quantitative results on validation set.

Organ	DSC (%)	NSD (%)
Liver	95.3±7.93	81.5±14.5
Kidney	91.1±14.2	76.8±17.9
Spleen	93.6±13.9	86.4±16.3
Pancreas	70.1±22.9	52.0±19.3

5.3. Qualitative results

Figure 2 presents some challenging examples. It can be found that the baseline method can not segment the lesion-affected organs well. The first row of Figure 2 illustrates a fatty liver case where the liver is darker than healthy ones. The baseline method fails to segment the spleen (blue) and the liver (red) in this case. Second row of Figure 2 shows an example with kidney (green) tumor which causes incorrect segmentation.

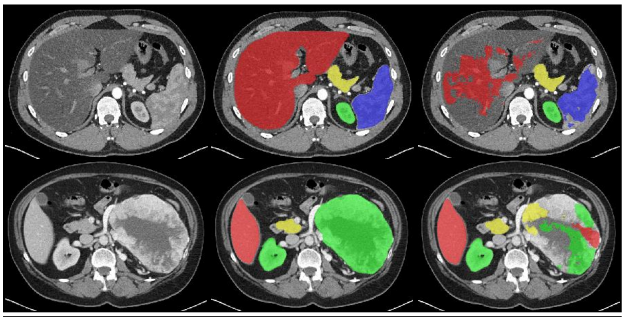

Figure 2. Challenging examples.First column is the image, second column is the ground truth, and third column is the predicted results by the baseline method [11].

6. Discussion and Conclusion

What kind of cases the proposed method works well?

The baseline method can work well on cases where no diseases exist. Besides, the DSC and NSD scores of liver segmentation is higher than the other organs, indicating liver maybe a comparable easier task as a result of its bigger size and consistent shape. Disappointing performance is obtained for pancreas segmentation as a result of the inter-patient anatomical variability of volume and shape.

What are the possible reasons for the failed cases?

The existence of leison is a critical factor for the segmentation performance. How to properly segment those cases is important. Besides, obtaining an accurate boundary segmentation need further investigate. Moreover, disappointing performance is obtained for pancreas segmentation as a result of the inter-patient anatomical variability of volume and shape.

Acknowledgment

The authors of this paper declare that the segmentation method they implemented for participation in the FLARE

challenge has not used any pre-trained models nor additional datasets other than those provided by the organizers.

References

[1] 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] 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*. Springer, 2015, pp. 234–241. 1

[3] J. Ma, J. Chen, M. Ng, R. Huang, Y. Li, C. Li, X. Yang, and A. L. Martel, "Loss odyssey in medical image segmentation," *Medical Image Analysis*, vol. 71, p. 102035, 2021. 2

[4] 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

[5] P. Bilic, P. F. Christ, E. Vorontsov, G. Chlebus, H. Chen, Q. Dou, C.-W. Fu, X. Han, P.-A. Heng, J. Hesser *et al.*, "The liver tumor segmentation benchmark (lits)," *arXiv preprint arXiv:1901.04056*, 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

[9] N. Heller, F. Isensee, K. H. Maier-Hein, X. Hou, C. Xie, F. Li, Y. Nan, G. Mu, Z. Lin, M. Han *et al.*, "The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge," *Medical Image Analysis*, vol. 67, p. 101821, 2021. 2

[10] N. Heller, S. McSweeney, M. T. Peterson, S. Peterson, J. Rickman, B. Stai, R. Tejpaul, M. Oestreich, P. Blake, J. Rosenberg *et al.*, "An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in ct imaging." *American Society of Clinical Oncology*, vol. 38, no. 6, pp. 626–626, 2020. 2

[11] J. Ma, Y. Zhang, S. Gu, C. Zhu, C. Ge, Y. Zhang, X. An, C. Wang, Q. Wang, X. Liu, S. Cao, Q. Zhang, S. Liu, Y. Wang, Y. Li, J. He, and X. Yang, "Abdomenct-1k: Is abdominal organ segmentation a solved problem?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 3