Efficient abdominal multi-organ segmentation based on the fusion of convolution and Transformer

Zongyu Li
Institute of Engineering Medicine,BIT
Beijing,China
3120201952@bit.edu.cn

Abstract

Deep learning is becoming more and more popular in medical image segmentation, and has achieved very good results.However, the existing methods have some defects, such as poor generalization ability and high computing resource consumption.In this paper, we propose a lightweight multi-organ segmentation network for abdominal multi-organs called Efficient multi-organ segmentation network(EMS-Net).The network is a combination of convolution network and Transformer.Transformer can better extract global features and alleviate the limitation of receptive field of CNN.In addition, in order to reduce the number of parameters in the Transformer network, the Deformer Transformer Block is only used for self-attention calculation of several points, which avoids the high calculation cost for global calculation.Meanwhile, Shuffle Unit is adopted in the encoder part to reduce the number of parameters.The final segmentation results show that our method can achieve good segmentation results with low computational resource consumption.

1. Introduction

In recent years, deep learning has achieved rapid development in the field of medical image processing, among which many segmentation methods have achieved remarkable results.Driven by the successful application of deep learning technology in many fields, image segmentation technology based on deep learning has been widely applied in medical teaching, surgical planning, surgical simulation and various medical research.

Abdominal multi-organs segmentation is an important research method in medical image segmentation, and many excellent algorithms have emerged.However, most methods show poor generalization when faced with multi-centers data. In addition, the number of network parameters is increasing, but the prediction speed is slow, and the computing resources are extremely expensive. It is due to the existence of many defects that the deep learning-based segmentation method cannot be well applied in clinical environment.

2. Method

We use CoTr[1] as our backbone, which uses a structure that combines CNN and Transformer. First, feature extraction is performed through a 3D Unet encoder, and the obtained multi-scale feature maps are concatenate and then extracted using Deformable Transformer Block[2] for global feature extraction. Finally, the extracted features are passed through a decoder to obtain the final segmentation result. Figure 1 illustrates the network structure.

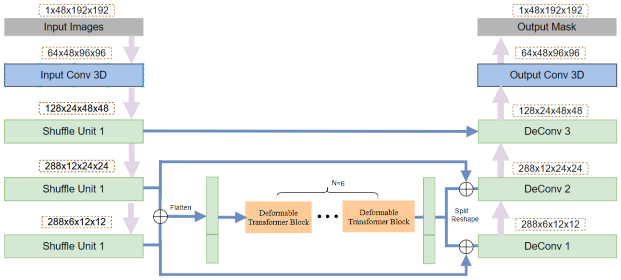

Figure 1. Network architecture

2.1. Preprocessing

The method includes the following preprocessing steps:

- Resampling method for anisotropic data:
  Resampling across the out of plane axis is done with nearest neighbor for both data and one-hot encoded segmentation maps.

- 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 global z-scorea normalization scheme is determined

based on the intensities found in foreground voxels across all training cases.

2.2. Proposed Method

We use a network structure that combines CNN and Transformer.Compared with CNN, Transformer can extract long-range information better and provide richer global information for the network. At the same time, multi-scale features can alleviate the network performance loss caused by the difference in organ size. However, the traditional Transformer network, such as ViT[3], has a large amount of parameters and consumes huge hardware resources. Therefore, the Deformable Transformer Block is adopted, which uses the idea of deformable convolution, and only considers a few points when calculating self-attention, instead of considering the full image, which greatly reduces the amount of parameters of the Transformer module.

Although the Transformer module has been improved, the overall network parameters have maintained a high level. In order to reduce the amount of network parameters, we used the ShuffleNet unit in ShuffleNetv2[4] to replace the 3*3 convolution block in the encoder, which significantly reduced the amount of model parameters and the hardware requirements. In addition, we appropriately adjusted the number of feature maps at each stage of encoder and decoder to reduces the amount of network parameters and memory usage.The number of model parameters is 6.62M.

In order to accelerate the training and inference speed of the model, we have adopted a multi-threaded processing and mixed precision method to train the model.

We jointly use the Dice loss and cross-entropy loss as the finally loss function.And the deep supervision is used

2.3. Post-processing

In order to reduce the inference time, we do not use any post-processing.

3. Dataset and Evaluation Metrics

3.1. Dataset

- A short description of the dataset used:
  The dataset used of FLARE2021 is adapted from MSD [5] (Liver [6], Spleen, Pancreas), NIH Pancreas [7, 8, 9], KiTS [10, 11], and Nanjing University under the license permission. For more detail information of the dataset, please refer to the challenge website and [12].

- 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)

4. Implementation Details

4.1. Environments and requirements

The environments and requirements of the baseline 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:
  No post-processing step is used.

- If using patch-based strategy, describing the patch aggregation method:
  The same patch-based strategy is applied as nnU-Net [13]. Voxels close to the center are weighted higher than those close to the border, when aggregating predictions across patches.

5. Discussion and Conclusion

In this work, we propose a lightweight abdominal multi-organ segmentation network that can achieve fast segmentation with low computational resource consumption.It can be found from the observation of the results that the method proposed by us can achieve a good segmentation effect for organs such as liver, which may be because liver occupies most of the area in the image and is significantly different from the background.However,due to the defects such as unclear pancreas boundary and large shape change, the segmentation effect of pancreas is generally normal, and there is still a lot of room for improvement.

The following work will focus on improving the segmentation effect of pancreas, further reducing the complexity of the model, and improving the infernce speed.

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

Table 2. Environments and requirements.

Windows/Ubuntu version	Ubuntu 18.04.5
CPU	Intel(R) Core(TM) Xeon Gold 5118
RAM	192GB; 2666MT/s
GPU	Nvidia RTX3090
CUDA version	11.1
Programming language	Python3.7
Deep learning framework	Pytorch (Torch 1.9, torchvision 0.10.0)
Specification of dependencies	nnUNet

Table 3. 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 [13].
Batch size	2
Patch size	48×192×192
Total epochs	500
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 (500).

Table 4. Quantitative results of DSC.

Training Dataset	Liver	Kidney	Spleen	Pancreas
	97.5	95.1	96.5	77.3

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] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," *arXiv preprint arXiv:2103.03024*, 2021. 1

[2] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020. 1

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020. 2

[4] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131. 2

[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] 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

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

[8] 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

[9] 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

[10] 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

[11] 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

[12] 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

[13] 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. 2, 3