MoTr: Efficiently Bridging MobileNet and Transformer for Abdominal Organ Segmentation

Tianqi Zhang
Laboratory of Beijing Engineering Research Center of Mixed Reality and Advanced Display
School of Optics and Photonics, Beijing Institute of Technology,Beijing,China
3120200599@bit.edu.cn

Yonglin Bian
Laboratory of Beijing Engineering Research Center of Mixed Reality and Advanced Display
School of Optics and Photonics, Beijing Institute of Technology,Beijing,China
byl1112138@163.com

Ruirui An
Laboratory of Beijing Engineering Research Center of Mixed Reality and Advanced Display
School of Optics and Photonics, Beijing Institute of Technology,Beijing,China
18811370128@163.com

Abstract

Convolutional neural networks (CNNs) have been the de facto standard for nowadays 3D medical image segmentation. The con- volutional operations used in these networks, however, inevitably have limitations in modeling the long-range dependency due to their induc- tive bias of locality and weight sharing. Although Transformer was born to address this issue, it suffers from extreme computational and spatial complexities in processing high-resolution 3D feature maps. In this work, we propose a novel framework that efficiently bridges a MobileNet and a Transformer (MoTr) for accurate abdominal organ segmentation. In order to achieve a balance between inference accuracy and speed, we use MobileNet as the CNN module.Under this framework, the CNN is constructed to extract feature representations and an efficient deformable Transformer (DeTrans) is built to model the long-range dependency on the extracted feature maps. Different from the vanilla Transformer which treats all im- age positions equally, our DeTrans pays attention only to a small set of key positions by introducing the deformable self-attention mechanism. Thus, the computational and spatial complexities of DeTrans have been greatly reduced, making it possible to process the multi-scale and high- resolution feature maps, which are usually of paramount importance for image segmentation. We conduct an extensive evaluation on the Fast and Low GPU memory Abdominal Organ Segmentation (FLARE) dataset that covers 4 major human organs. The results indicate that our MoTr has achieved excellent performance in inference speed and accuracy on the 3D multi-organ segmentation task.

1. Introduction

Image segmentation is a longstanding challenge in medical image analysis. Since the introduction of U-Net [1], fully convolutional neural networks (CNNs) have become the predominant approach to addressing this task. Despite their prevalence, CNNs still suffer from the limited receptive field and fail to capture the long-range dependency, due to the inductive bias of locality andweight sharing [2].

Transformer, a sequence-to-sequence prediction framework, has a proven track record in machine translation and nature language processing [3], due to its strong ability to long-range modeling. The self-attention mechanism in Transformer can dynamically adjust the receptive field according to the input content, and hence is superior to convolutional operations in modeling the long- range dependency.

In this work, we propose a hybrid framework that efficiently bridges MobileNet and Transformer (MoTr) for 3D medical image segmentation. MoTr has an encoder-decoder structure. In the encoder, a conciseCNN structure is adopted to extract feature maps and a Transformer is used to capture the long-range dependency (see Figure 1). Inspired by [4], we introduce the deformable self-attention mechanism to the Transformer. This attention mechanism casts attentions only to a small set of key sampling points, and thus dramatically reduces the computational and spatial complexity of Transformer. As a result, it is possible for the Transformer to process the multi-scale feature maps produced by the CNN and keep abundant high resolution information for segmentation. The main contributions of this paper are three-fold: (1) we are the first to explore Transformer for 3D medical image segmentation, particularly in a computationally and spatially efficient way; (2) we introduce the deformable self-attention mechanism to reduce the complexity of vanilla Transformer, and thus enable our MoTr to model the long-range dependency using multi-scale fea- tures; (3) we introduced the MobileNet [5] structure in the CNN coding part. The results show that our MoTr has achieved excellent performance in inference speed and accuracy.

2. Method

MoTr aims to learn more effective representations for medical image segmentation via bridging MobileNet and Transformer. As shown in Figure 1, it consists of a CNN encoder (MobileNet) for feature extraction, a deformable Transformer-encoder (DeTrans- encoder) for long-range dependency modeling, and a decoder for segmentation.

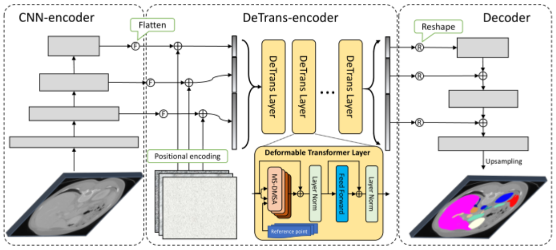

Figure 1. Diagram of MoTr

2.1. Preprocessing

The baseline method includes the following preprocessing steps:

- Cropping strategy: We randomly cropped sub-volumes of size 80×192×160 from CT scans as the input.
- Online data argumentation method:
  To alleviate the over-fitting of limited training data, we employed the online data argumentation, including the random rotation, scaling, flipping, adding white Gaussian noise, Gaussian blurring, adjusting rightness and contrast, simulation of low resolution, and Gamma transformation, to diversify the training set.
- Intensity normalization method:
  We first truncated the HU values of each scan using the range of [-958,327] to filter irrelevant regions, and then normalized truncated voxel values by subtracting 82.92 and dividing by 136.97.

2.2. Proposed Method

Full description of the proposed method. ***Pre-trained models are not allowed to use in this challenge.***

- Network architecture details:

  The CNN-encoder contains a Conv-IN-ReLU block and three stages of MobileNet blocks. The Conv-IN-ReLU block contains a 3D convolutional layer followed by an instance normalization (IN) [6] and Rectified Linear Unit (ReLU) activation. The numbers of MobileNet blocks in three stages are three, three, and two, respectively.

  Considering that Transformer processes the information in a sequence-to-sequence manner, we first flatten the feature maps produced by the CNN-encoder into a 1D sequence. Unfortunately, the operation of flattening the features leads to losing the spatial information that is critical for image segmentation. To address this issue, we supplement the 3D positional encoding sequence to the flattened 1D sequence.

  Due to the intrinsic locality of convolution operations, the CNN-encoder can- not capture the long-range dependency of pixels effectively. To this end, we propose the DeTrans-encoder that introduces the multi-scale deformable self- attention (MS-DMSA) mechanism for efficient long-range contextual modeling. The DeTrans-encoder is a composition of an input-to-sequence layer and stacked deformable Transformer (DeTrans) layers.

  The output sequence of DeTrans-encoder is reshaped into feature maps according to the size at each scale. The decoder, a pure CNN architecture, progressively upsamples the feature maps to the input resolution using the transpose convolution, and then refines the upsampled feature maps using a 3D residual block. Besides, the skip connections between encoder and decoder are also added to keep more low-level details for better segmentation. We also use the deep supervision strategy by adding auxiliary losses to the decoder outputs with different scales.

- Loss function: The loss function of our model is the sum of the Dice loss and cross-entropy loss [7].
- Number of model parameters: 12.66M
- Number of flops: 590861472000 (can be computed via such as fvcore library for Pytorch)

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

3. Dataset and Evaluation Metrics

3.1. Dataset

- A short description of the dataset used:
  The dataset used of FLARE2021 is adapted from MSD [8] (Liver [9], Spleen, Pancreas), NIH Pancreas [10, 11, 12], KiTS [13, 14], and Nanjing University under the license permission. For more detail information of the dataset, please refer to the challenge website and [15].

- 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. The proposed network was implemented by using the deep learning library of PyTorch 1.7.1. All experiments were conducted on Ubuntu 18.04.5 system with Nvidia GeForce RTX 3090, which has 24GB RAM.

4.2. Training protocols

The training protocols of the baseline method is shown in Table 3. In the training stage, we randomly cropped

Table 2. Environments and requirements.

Windows/Ubuntu version	Ubuntu 18.04.5 LTS
CPU	Intel(R) Xeon(R) Gold 5118 CPU@2.30GHz
RAM	48×4GB; 2.7MT/s
GPU	Nvidia GeForce RTX 3090
CUDA version	11.3
Programming language	Python3.6
Deep learning framework	Pytorch (Torch 1.7.1, torchvision 0.8.2)
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 [16].
Batch size	2
Patch size	80×192×160
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	168 hours
CO_2eq[1]	

sub-volumes of size 80×192×160 from CT scans as the input. To alleviate the over-fitting of limited training data, we employed the online data argumentation, including the random rotation, scaling, flipping, adding white Gaussian noise, Gaussian blurring, adjusting rightness and contrast, simulation of low resolution, and Gamma transformation, to diversify the training set. Due to the benefits of instance normalization, we adopted the micro-batch training strategy with a small batch size of 2. To weigh the balance between training time cost and performance reward, MoTr was trained for 1000 epochs and each epoch contains 250 iterations. We adopted the stochastic gradient descent algorithm with a momentum of 0.99 and an ini- tial learning rate of 0.01 as the optimizer.

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.
- Patch aggregation method:
  We employed the sliding window strategy, where the window size equals to the training patch size. The softmax prediction values of the overlapping regions are averaged when aggregating predictions across patches.Besides, Gaussian importance weighting and test time augmentation by flipping along all axes were also utilized to improve the robustness of segmentation.

5. Results

5.1. Quantitative results on validation set.

Table 4 illustrates the results on validation cases. It is worth pointing out that for liver segmentation, the DSC scores are 90.6% , indicating great segmentation performance in terms of region overlap between the ground truth and the segmented region. NSD values are 66.6% demonstrating that the boundary regions contain more segmentation errors, which need further improvements [15].

Table 4. Quantitative results on validation set.

Organ	DSC (%)	NSD (%)
Liver	90.6±10.2	66.6±17.1
Kidney	74.5±15.8	57.8±16.4
Spleen	78.5±28	68±25
Pancreas	56.8±24.3	40.6±19.6

5.2. 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.

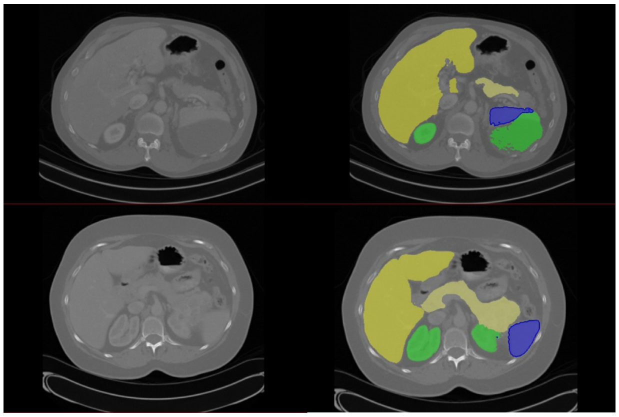

Figure 2. Challenging examples.First column is the image, second column is the predicted results by our proposed method.

6. Discussion and Conclusion

The proposed method can work well on cases which is in the same data center as the training set. 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.

The processing of multi-center data is very important. We have introduced federated learning in our work. However, the effect achieved is not very obvious, and further analysis is needed.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] 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

[2] N. Cohen and A. Shashua, "Inductive bias of deep convolutional networks through pooling geometry," *arXiv preprint arXiv:1605.06743*, 2016. 1

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017. 1

[4] J. Dai, H. Qi, Y. Xiong, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," *Proceedings of the IEEE international conference on computer vision*, pp. 764–773, 2017. 1

[5] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, and W. Wang, "Mobilenets: efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017. 2

[6] D. Ulyanov, A. Vedaldi, V. Lempitsky, D. Kalenichenko, and W. Wang, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016. 2

[7] F. Isensee, S. Kohl, J. Petersen, and K. Maier-Hein, "Automated design of deep learning methods for biomedical image segmentation," *arXiv preprint arXiv:1904.08128*, 2019. 2

[8] 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. 3

[9] 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. 3

[10] H. Roth, A. Farag, E. Turkbey, L. Lu, J. Liu, and R. Summers, "Data from pancreas-ct. the cancer imaging archive (2016)." 3

[11] 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. 3

[12] 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. 3

[13] 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. 3

[14] 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. 3

[15] 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. 3, 4

[16] 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. 3