Effectiveness of transformer on U-net segmentation for human ultrasound images

Jiang Zhou[1], Chikayoshi Sumi[1†] (Info. Sci., [1]Grad. School of Sophia Univ.)

1. Introduction

In recent years, Convolutional Neural Networks (CNNs) have made breakthroughs in the field of various image analysis. Especially in segmentation tasks, U-Net[1)] becomes a deep learning method widely used in the field of medical imaging. However, CNNs struggle to capture deep and extensive semantic information due to limited receptive fields. Transformer,[2)] renowned for its excellence in natural language processing, have recently made significant strides in computer vision tasks as well. It leverages Multi-head Self-Attention (MSA) to excel at capturing global context and long-range dependencies. In this study, we integrated U-net with Transformer and dealt with human in vivo breast and cardiac ultrasound echo images to investigate the effectiveness of Transformer on U-net segmentation.

2. Experiment

2.1 Experimental datasets

For the breast ultrasound, an open access BUSI[3)] dataset was used. BUSI contained 437 benign tumors and 210 malignant tumors. For the cardiac ultrasound, CAMUS[4)] dataset comprising 2D apical four-chamber view sequences of 500 patients were used.

2.2 New model

To integrate Transformer into U-net, we developed a new U-shaped structure (**Fig. 1a**). Similarly to the original U-net, the new U-shaped model undergoes down-sampling and up-sampling (four times) to utilize the high-level semantic feature map obtained by the encoder to restore up to the resolution of the original image by the decoder. Besides, to supplement the loss of information caused by the change of feature scale during decoding, the skip connections from the encoder are also used. And, in the new model, we introduced self-attention mechanism;[2)] added the transformer blocks (**Fig. 1b**) to both the encoder and decoder; and replaced some skip connections by those with the transformer decoders. Thus, the new model applies convolutional layers to extract local intensity features with avoiding large-scale pre-training of transformers; and uses self-attention to capture global features for improving segmentation results.

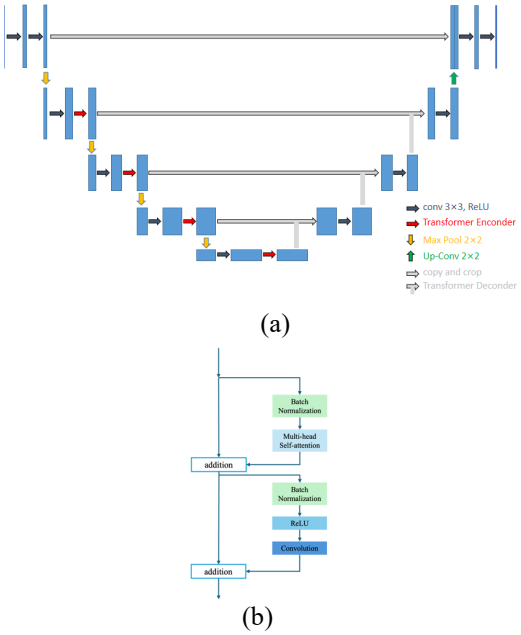

Fig. 1 (a) Overview of proposed U-net based model, and (b) transformer blocks.

In the self-attention module, inter-pixel relative position encoding is performed to compensate for the missing position information.

In order to comprehensively consider the binary classification and similarity problems at the pixel level, we used a hybrid segmentation loss that combines binary cross entropy (BCE) and Dice loss. Besides, we used the Adam parameter optimizer for training and reduced the learning rate if the Dice Loss does not increase after two iterations.

2.3 Experimental procedures

A series of comparative experiments were conducted on the new and original U-Net models. At first, for the breast ultrasound dataset, 3 experiments were performed, i.e., for the benign data only, the malignant data only, and all the mixed data. Next, left ventricular endocardial segmentation was performed at the ends of diastole (ED) and systole (ES). The training dataset, and evaluation and test dataset were divided in a ratio of 8 : 2.

Performance was evaluated using Dice coefficient (Dice). In addition, the intersection over union (IoU) value, accuracy, sensitivity, and specificity were also evaluated.

3. Results

3.1 Breast tumors

E-mail: †c-sumi@sophia.ac.jp

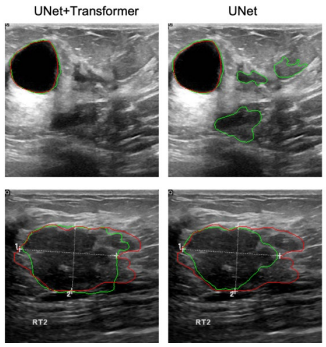

Fig. 2 Examples of segmentation results on BUSI dataset (upper) for benign and (lower) malignant tumors with (left) new and (right) original U-net models.

Table I For BUSI dataset, metric values obtained with two methods: (upper) benign only, (middle) malignant only, and (lower) mixed.

BUSI_benign	**Dice**	IoU	Accuracy	Sensitivity	Specificity
U-net	0.7945	0.6592	0.9765	0.8474	0.9839
U-net+Transformer	**0.8409**	**0.7197**	**0.9810**	**0.8672**	**0.9875**

BUSI_malignant	**Dice**	IoU	Accuracy	Sensitivity	Specificity
U-net	0.7615	0.6057	0.9256	0.7812	**0.9505**
U-net+Transformer	**0.7718**	**0.6212**	**0.9272**	**0.8245**	0.9451

BUSI_all	**Dice**	IoU	Accuracy	Sensitivity	Specificity
U-net	0.7720	0.6224	0.9611	0.8330	0.9759
U-net+Transformer	**0.8401**	**0.7146**	**0.9677**	**0.8494**	**0.9821**

Fig. 2 shows examples of segmentation results for (upper) benign and (lower) malignant tumors (left) with the new and (right) original models. The red line indicates the ground truth (GT), and the green line the predicted result. The metric values of two models for the 3 experiments are summarized in Table I. As shown, for almost the metrics, the new model was superior to the original model particularly for the benign (e.g., Dice difference, 0.0464) and all the mixed data (0.0681). The original model sometimes segmented others from the tumors simultaneously as shown for the benign tumor in Fig. 2, whereas the new model succeeded in decreasing the failure. It would have been nice if the new model succeeded in learning more features about malignant tumors more efficiently, but the new model achieved slightly better results than the original model (Dice difference, 0.0103).

As far as we have experienced for BUSI dataset,[5] the original model yields better results for the benign than for the malignant tumors as also shown in this paper. Generally, an invasive-type malignant tumor has a considerably irregular shape, which was sufficiently included in BUSI dataset. Moreover, there are in the dataset the 2-fold benign data than the malignant data. These are the reason why the original model has shown such a performance. These were same for the new model. With both the models for all the mixed data, the values of accuracy, sensitivity and specificity were

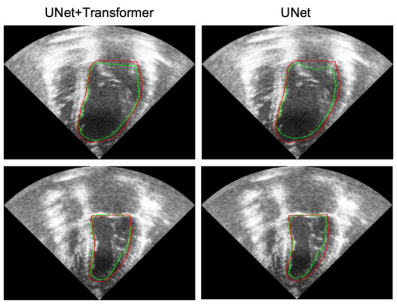

Fig. 3 Examples of segmentation results on CAMUS of the left ventricular endocardium for (upper) ED and (lower) ES datasets.

Table II For CAMUS dataset, metric values obtained with 2 methods: (upper) ED and (lower) ES.

CAMUS_ED	**Dice**	IoU	Accuracy	Sensitivity	Specificity
U-net	0.9257	0.8367	0.9812	0.8973	0.9914
U-net+Transformer	**0.9305**	**0.8482**	**0.9825**	**0.9165**	**0.9916**

CAMUS_ES	**Dice**	IoU	Accuracy	Sensitivity	Specificity
U-net	0.8978	0.7833	0.9813	0.8508	**0.9927**
U-net+Transformer	**0.9091**	**0.8101**	**0.9833**	**0.8975**	0.9910

closer to those for benign than for malignant tumors.

3.2 Cardiac

Fig. 3 and **Table II** show, for segmentation on CAMUS of the left ventricular endocardium (upper) at the end-diastole (ED) and (lower) end-systole (ES), the corresponding results as those of the breast tumors. As shown, the new model yielded the better results than the original model. Because of the more regular shape of left ventricle endocardium than that of breast tumor, both the models achieved the much more accuracy.

4. Conclusion

Integration of U-net and Transformer with self-attention mechanism achieved considerably high accuracy segmentation for the human in vivo breast tumors and left ventricular endocardium. For the additional performance evaluation, augmentation will be performed. Combination with YOLO will also be performed.

References

1) O. Ronneberger, P. Fischer, and T. Brox, Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (Springer, Berlin, 2015) p. 234.
2) J. A. Vaswani, N. Shazeer, N. Parmar, et al., Adv. Neural Inf. Proces. Syst. **30**, 5998 (2017).
3) W. Al-Dhabyani, M. Gomaa, H. Khaled, A. Fahmy, Data in Brief. **28**, 104863 (2020).
4) S. Leclerc, E. Smistad, J. Pedrosa, A. Ostvik, et al., IEEE Trans on Medical Imaging **38**, 2198 (2019).
5) J. Xiao, S. Tamatani, Y. Hirano, C. Sumi, Proc 44th symp USE (2023) 2P5-4.