Integration Self-attention with UNet for Tumor Segmentation in Breast Ultrasound

Chii-Jen Chen (陳啓禎), *Yu-Jie Chiou* (邱昱傑), *Shao-Hua Hsu* (許少華) *and Yu-Cheng Chang* (張祐誠)

Department of Computer Science and Information Engineering,
Tamkang University, New Taipei City, Taiwan,
E-mail: kevin456hope@tku.edu.tw

ABSTRACT

UNet has achieved remarkable results and made significant contributions in the semantic segmentation of medical images. Recently, the rapid development of large language models has brought new milestones to the field of artificial intelligence, inspiring us to apply their successful experiences to neural networks in computer vision. This study incorporates the self-attention mechanism into the UNet architecture, enabling each pixel to better understand global information, thereby enhancing the relationships between features. We conducted experiments on medical image datasets, and the results indicate that the enhanced model significantly improves segmentation accuracy and robustness. Our research showcases the potential of the self-attention mechanism in enhancing the performance of medical image segmentation.

Keywords: *UNet, Self-attention, Segmentation, Breast ultrasound*

1. INTRODUCTION

In the field of medical image analysis, semantic segmentation of images is a crucial technology that plays a significant role, particularly in diagnosis and treatment. Since the introduction of the UNet [1] model, it has demonstrated excellent performance in various medical image segmentation tasks and has inspired many variations, thus becoming a fundamental method in this field. However, the traditional UNet architecture primarily relies on convolution operations to extract local features, which may lead to insufficient capture of global information, thereby limiting the model's segmentation performance.

In recent years, the rapid development of large language models, especially the introduction of the Transformer architecture and the self-attention mechanism, has brought revolutionary advancements to natural language processing. Compared to RNN and LSTM [2] , these mechanisms can better integrate global information and perform parallel computation. Inspired by this, we believe that the self-attention [3] mechanism can also provide advantages in semantic segmentation.

In this study, we utilized UNet as our base model and combined it with the self-attention mechanism as a feature extraction method to better integrate multi-scale feature maps. We conducted experiments on breast cancer ultrasound image datasets, and the results indicate that the enhanced model significantly improves segmentation accuracy and robustness. Our research demonstrates the potential of the self-attention mechanism in improving the performance of medical image segmentation.

2. RELATED WORK

There has been significant research on applying self-attention mechanisms in image recognition models, such as Vision Transformer (ViT) [4] and Swin Transformer [5]. ViT abandons the traditional CNN [6] method of using convolutional kernels to scan the entire image. Instead, it converts the image into a sequence of patches, adds positional embeddings, and then processes these patches using a Transformer encoder with self-attention and multi-head attention mechanisms. This allows the model to capture long-range dependencies and global context more effectively than CNNs.

Swin Transformer addresses vision downstream tasks by structuring the model hierarchically, similar to CNNs. It divides the image into non-overlapping windows and applies self-attention within each window. Through a shifting window scheme, it facilitates cross-window connections, capturing global information while preserving computational efficiency.

From the experiences of these two approaches, we understand that achieving good results in vision downstream tasks requires the integration of multi-scale feature maps. However, solely using self-attention for feature extraction demands an extremely large training dataset and substantial computational resources. Therefore, this study aims to combine the strengths of both approaches. We apply self-attention to feature maps to capture global features and integrate CNN's convolutional operations to extract local features effectively. This hybrid model, termed self-attention UNet, aims to leverage the advantages of both methodologies to enhance performance in medical image segmentation.

3. METHODOLOGY

3.1. Self-Attention Mechanism

By applying convolutional layers to the input image, we generate the query and key, with their channel dimensions reduced by a factor of 8 to decrease computational load. Next, we calculate the similarity between the query and key using a dot product. This similarity is then transformed into corresponding attention weights through the softmax function. Finally, these attention weights are used to perform a weighted summation on the value, producing a feature map(Fig. 1) that integrates global contextual features.

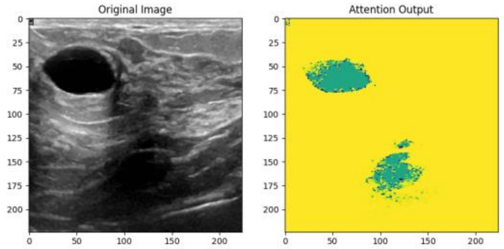


Fig. 1 Feature map generated by self-attention

3.2. Integration with UNet Architecture

This study refers to the skip-connection structure of UNet, which integrates features of different scales by concatenation. We performed experiments using UNet as the prototype at three specific locations in the model, as indicated by the labels in Fig. 2.

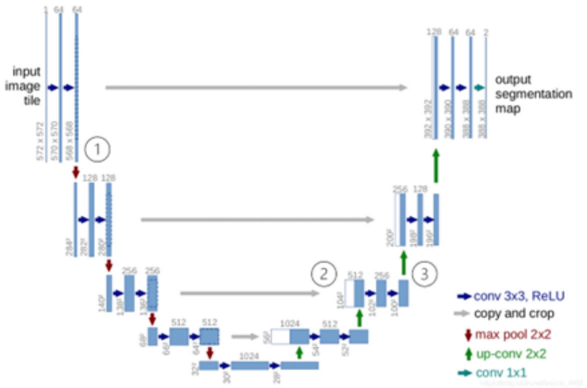


Fig. 2 Three experimental locations at Unet

3.2.1. Encoder

In the encoder, the feature maps obtained after two convolutions are processed through self-attention and then concatenated with the original feature maps, as depicted in Fig. 3.

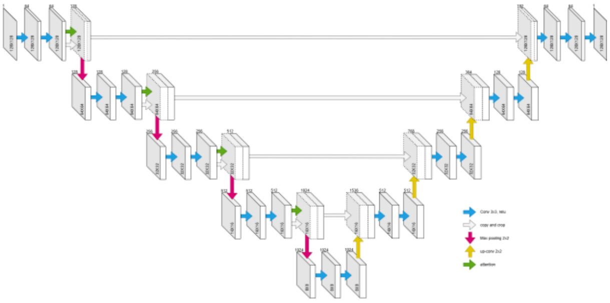


Fig. 3 Unet Encoder with Integrated Self-Attention

3.2.2. Skip connection

The skip connection feature maps are processed through self-attention and then concatenated with the original feature maps and the transposed convolution feature maps, as depicted in Fig. 4.

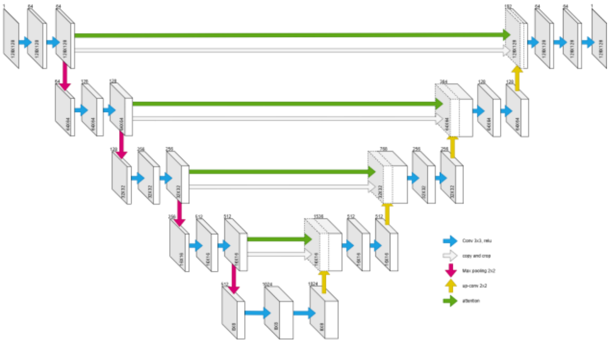


Fig. 4 Unet Skip-connection with Integrated Self-Attention

3.2.3. Decoder

In the decoder, the feature maps obtained after two convolutions are processed through self-attention and then concatenated with the original feature maps, as depicted in Fig. 5.

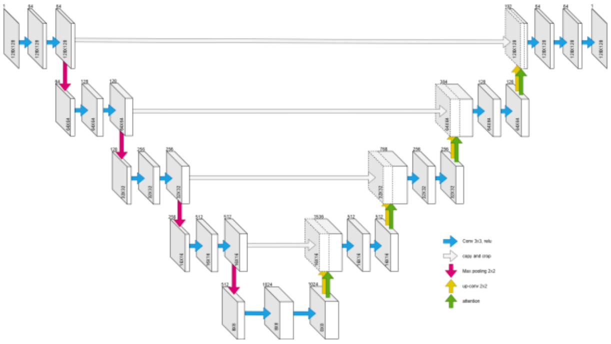


Fig. 5 Unet Decoder with Integrated Self-Attention

4. EXPERIMENTS AND RESULTS

4.1. Experimental Setup

4.1.1. Dataset

Through the Kaggle website, Al-Dhabyani W, Gomaa M, Khaled H, Fahmy A. Dataset of breast ultrasound images. Data in Brief. 2020 Feb;28:104863. DOI: 10.1016/j.dib.2019.104863.

This dataset provides breast ultrasound images of women aged 25 to 75. The data were collected in 2018, involving 600 female patients. The dataset consists of 780 images with an average size of 500x500 pixels. The images are in PNG format. Both the original and annotated images are included. The images are categorized into three classes: normal, benign, and malignant.

4.1.2. Training Details

We trained the network using the Adam optimizer with an initial learning rate of 0.001. The batch size was set to 16, and the network was trained for 100 epochs. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU with 55.9 GB of memory.

4.2. Evaluation Metrics

We chose the Dice coefficient as our evaluation metric due to the characteristics of medical images, where the proportion between lesion areas and normal areas can vary significantly. The Dice coefficient offers a more reliable assessment, as it can handle these imbalances. Our goal is for models to accurately detect disease areas without erroneously identifying normal tissues as diseased.

4.3. Experimental Results

According to Fig. 2, incremental experiments were conducted on three specific parts of the UNet model to incorporate self-attention. The experimental results are as follows

4.3.1. Phase 1 Experiment- Encoder

Through the encoder, the feature map obtained by applying self-attention to the feature map after two convolutions is concatenated with the original feature map. The experimental results, as shown in the "yellow bounding box" in Fig. 6, indicate that due to the inclusion of self-attention during feature extraction in the encoder, the feature extraction process is incomplete.

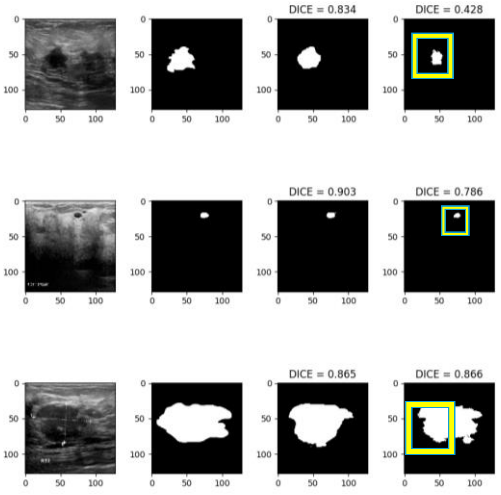
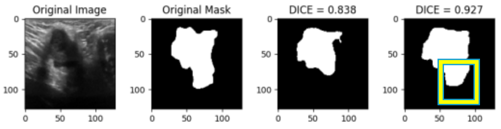


Fig. 6 Phase 1 Experimental Comparison Results

4.3.2. Phase 2 Experiment- Skip connection

Through skip connections, the feature map obtained by applying self-attention to the feature map that will be used for skip connection, along with the original feature map and the feature map obtained through transposed convolution, are concatenated. The experimental results, as shown in the "yellow bounding box" in Fig. 7, indicate an improvement in performance compared to the previous comparisons, as the features can be extracted more comprehensively.



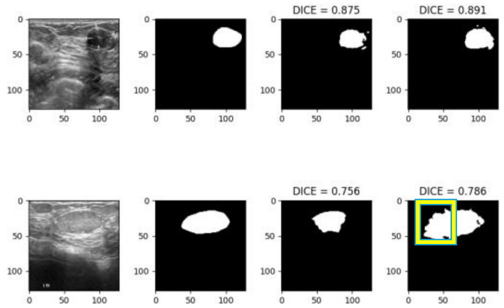


Fig. 7 Phase 2 Experimental Comparison Results

4.3.3. Phase 3 Experiment- Decoder

Through the decoder, the feature map obtained by applying self-attention to the feature map after two convolutions is concatenated with the original feature map. The experimental results, as shown in the "yellow bounding box" in Fig. 8, indicate an improvement in performance, as the features can be extracted more comprehensively.

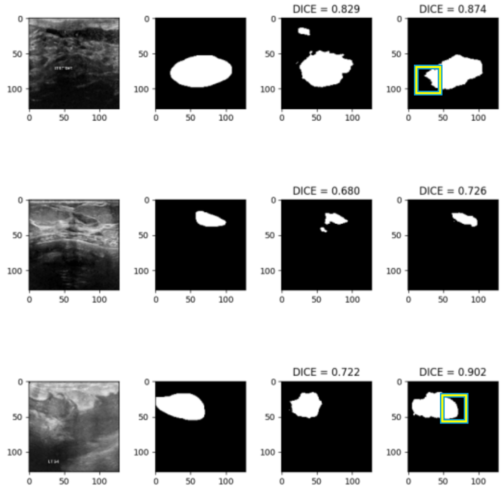


Fig. 8 Phase 3 Experimental Comparison Results

According to Table 1, it is evident that integrating the self-attention mechanism, especially in the skip connection and decoder stages, can significantly enhance the segmentation performance across all indicators.

Table 1 Performance metrics (DICE, IOU, Sensitivity) for different self-attention integrations of the UNet model.

	DICE	IOU	Sensitivity
Unet	73.38	76.57	67.81
Encoder	73.43	76.78	62.07
Skip connection	76.76	78.85	69.13
Decoder	79.23	80.57	77.44

5. DISCUSSION

The addition of the self-attention mechanism achieves better results on this dataset than the original Unet model under the same training cycle and batch size, with the potential to be extended to other medical imaging tasks.

We are also looking forward to extending the application of self-attention mechanisms to UNet++ [7] . In situations where medical training data is difficult to obtain, self-attention can improve accuracy without requiring large amounts of training resources, providing hope for advances in medical image analysis.

6. CONCLUSION

Based on the experimental results, we observed that integrating self-attention mechanisms with the feature maps significantly enhances semantic segmentation performance. This improvement is particularly notable when the self-attention mechanism is applied at the skip-connection and decoder stages of the UNet architecture.

7. REFERENCES

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, 2015: Springer, pp. 234-241.

[2] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[3] A. Vaswani et al., "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[4] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2020.

[5] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 10012-10022.

[6] K. O'shea and R. Nash, "An introduction to convolutional neural networks," arXiv preprint arXiv:1511.08458, 2015.

[7] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," in Deep Learning in Medical Image Analysis and Multimodal Learning for

Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4, 2018: Springer, pp. 3-11.