UNet and Transformers: Deep Learning Based Methods for Medical Image Segmentation

Zhirui Ren
Department of Mechanical Engineering, Rensselaer Polytechnic Institute, Troy, NY, United States of America

Keywords: Image Segmentation, Deep Learning, Transformer, Convolutional Neural Network, Medical Image.

Abstract: As a vital sub-part of medical image analysis and processing, image segmentation is a time-consuming and heavily experienced task when performed manually. With the revolutionary development of artificial intelligence (AI), intended to utilize the high efficiency and reliability of computerized information processing to address the problem of increasingly large quantities of medical images waiting to be processed, deep learning-based methods for segmentation tasks have become popular. Convolutional Neural Networks (CNNs) are an old leader in the computer vision community, but as transformer models have obtained excellent results in the field of natural language processing (NLP), increasing researchers have begun to explore whether they can also bring significant breakthroughs for image processing. In this review, some evaluation metrics are first to be introduced. Subsequently, the introduction of the core ATTENTION mechanism of transformers and three selected models with their performance follows. Through the survey, using the mature UNet method alone, good accuracy can be achieved, and if combined with the superiority of transformers' global context capture ability, even better results can be obtained. Dedicated to promoting the birth of a generalized image model with high accuracy, this article is provided for researchers' reference.

1 INTRODUCTION

Coming from a variety of sources including magnetic resonance imaging (MRI), computed tomography (CT), ultrasound imaging (UI), positron emission tomography (PET), and X-ray imaging, medical images are crucial for medical care (Liu et al. 2021) these images can assist health care and provide access to disease diagnosis and surgical guidance, they are mainly studied by experts with visual interpretation and these procedures consume time and the outcomes are subjectively generated depending on personal experience (Patil & Deore 2013). As the first step of image analysis, segmentation of the region of interest, which could be abnormal tissues, tumors, and organs, often leads the way for upcoming applications such as the study of anatomical structure, localization of pathology, and treatment planning (Patil & Deore 2013). The quality of the segmentation usually influences the diagnosis, and sometimes trivial differences matter. Segmenting images is a critical step in medical analysis. With the progress achieved in computer-aided techniques, automatic image segmentation methods, including the ones of thresholding, region growth, clustering, edge detection, and model-based, have become a research hotspot (Ramesh et al. 2018). Among them, deep learning-based methods with their powerful end-to-end functionality of processing complex as well as multivarious data and providing the target results, are the most studied and developed.

In the last era of the imaging process, Convolutional Neural Networks (CNNs) has an unshakeable status. The convolution operator, which functions locally and offers translational equivariance, is the mainstay of CNNs (Shamshad et al. 2023). Even though these qualities aid in the development of effective medical imaging solutions, the local receptive field in convolution operation restricts the ability to capture long-range pixel associations, and the stationary weights of the convolutional filters are not adjusted for variable input images (Shamshad et al. 2023). Witnessing how transformers have dominated the field of natural language processing (NLP), researchers then started to push the development and implementation of transformers trying to introduce the strong point of their ability to capture context relationships in the medical image segmentation field.

Ren, Z.
UNet and Transformers: Deep Learning Based Methods for Medical Image Segmentation.
DOI: 10.5220/0012838300004547

In *Proceedings of the 1st International Conference on Data Science and Engineering (ICDSE 2024)*, pages 547-551
ISBN: 978-989-758-690-3

In this paper, UNet and transformers variants are focused on. In section 2, some well-used evaluation metrics for image segmentation tasks are first to be introduced. Then comes the reviews of several popular deep-learning models and their performance in experiments. Section 3 concludes the paper.

2 DEEP LEARNING-BASED METHODS

2.1 Segmentation Evaluation Metrics

To determine how well an algorithm performs on the task, certain evaluation metrics are to be introduced. For medical image segmentation, the outcomes produced by the algorithm are compared with the doctors' manually annotated masks (known as Ground Truth, GT) by implementing mainly two methods. Both methods concern the intersection of the predicted and ground truth segmentations.

Dice coefficient (Dice similarity coefficient): DC (or DSC) calculated the ratio of twice the intersection of the two segmentations divided by the sum of their areas. Given two sets A and B, the metric is defined as: Dice (A, B) = $\frac{2*|A\cap B|}{|A|+|B|}$ (Dosovitskiy et al. 2020).

DSC is also calculated as: DSC = $\frac{2TP}{FP+2TP+FN}$, where True Positive (TP), False Positive (FP), and False Negative (FN) denote true positive, false positive and false negative respectively.

Jaccard index (Intersection over Union-IOU): The Jaccard index is similar to the Dice coefficient. It calculates the ratio of the intersection to the union of the two segmentations.

Given two sets A and B, the metric is defined as: Jaccard (A, B) = $\frac{|A\cap B|}{|A\cup B|}$ (Dosovitskiy et al. 2020)

Both metrics range from 0 to 1, with 1 indicates a perfect match and 0 stands for poor segmentation.

Hausdorff Distance (HD): HD is a measure of dissimilarity between two sets. It calculates the maximum distance between corresponding points in two sets considering both the precision (how well the segmented outputs match the ground truth) and the recall (how well the ground truth is covered by the outputs). A smaller HD indicates a higher quality and accuracy of the algorithm.

2.2 Deep Learning

As a branch of machine learning (ML) and artificial intelligence (AI), deep learning, along with its learning capabilities from data, has become a hot topic in various areas (Sarker 2021). In a neural network, the neuron is the smallest unit which can be viewed as a small information processor. In certain ways, the combination of several neurons forms a layer of neural network. With the employment of transformations and graph technologies, multi-layer learning models are built, and hence deep learning emerges (Alzubaidi et al. 2021). The emergence of neural networks opens the door for end-to-end methods dealing with problems in many fields, including medical image segmentation (Liu et al. 2021).

2.3 UNet

UNet has walked into a relatively popular spot in medical image segmentation ever since it was introduced by Ronnenberger et al in 2015 (Ronneberger et al. 2015). It is an efficient Fully Convolutional Network(FCN) based DL network. Many variants of UNet have also come out showing its unshakable position in the field. As shown in Figure 1, this model comprises a symmetric encoder-decoder architecture resembling a U-shape with skip-connections in between. On the left side is a contracting path consisting of several repeated applications of 3x3 convolutions, a rectified linear unit(ReLU), and a 2x2 max pooling operation (Ronneberger et al. 2015). And for the right side, a similar expansive path exists performing the upsampling process.

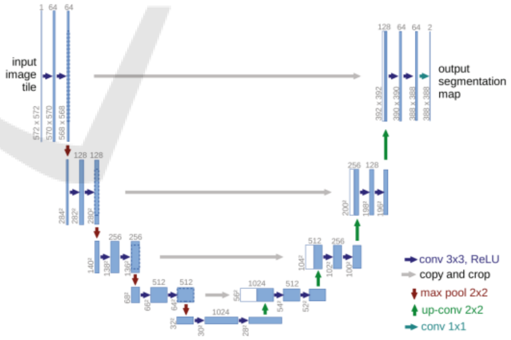

Figure 1. The architecture of UNet (Ronneberger et al. 2015).

Back in 2015, UNet achieved an average IOU of 92% on the segmentation task on the "PhC-U373" dataset which was remarkably better than the algorithm in the second place with 83% (Ronneberger et al. 2015). Enlightened by UNet, many other derivative models have also got outstanding results with metrics over 80% on various medical image segmentation data sets (Ghosh et al. 2021, Dong et al. 2017, Alali & Ali 2022).

Considering the low resolution of the expected outing images and the locality property of the FCN-based methods, researchers lay their eyes on transformers seeking the ability to manage long-range dependencies of the feature maps (Thisanke et al. 2023).

2.4 Attention Mechanism

As the core of transformers, the attention mechanism is to be introduced first. This idea was initially brought out by Bahdanau et al. for language translation (Bahdanau et al. 2015). Then comes the breaking-through concept of Self-attention (Vaswani et al. 2017). The basic idea of self-attention is to compute the weighted sum of the values, where the weight is related to the queries and keys. To better capture hierarchical features, multi-head self-attention is introduced, and it is realized by concatenating the outputs calculated by attention (Vaswani et al. 2017).

2.5 Vision Transformers

Besides the success of NLP, transformers have also revolutionized the field of computer vision and brought significant advancements in tasks such as image classification, object detection, and of course, image recognition (Dosovitskiy et al. 2020, Carion et al. 2020, Liu et al. 2021).

2.6 Swin Transformer

Swin Transformer is a transformer-based architecture designed to understand image sciences, including semantic segmentation. It was proposed by Liu et al. in 2021 (Liu et al. 2021). As seen in Figure 2, instead of processing the entire image at once, Swin Transformer divides the image into non-overlapping patches and employs multi-head self-attention mechanisms to capture the global context information from these patches. This architecture also uses a “shifted window” mechanism (see Figure 2(b), SW-MAS stands for shifted window-based multi-head self-attention) which allows information to flow across different stages and enhance modeling power (Liu et al. 2021).

2.7 Swin-UNet

Proposed by Cao et al. in 2021, Swin-UNet is a combination of Swin Transformer and UNet (Cao et al. 2021). It is designed for semantic segmentation tasks involving assigning pixel-level labels to different regions or objects within an image. As shown in Figure 3, this model employs a U-shape architecture with skip connections, while compared to a traditional UNet model, the convolutional layers are replaced with Swin Transformer blocks (shown in Figure 2 (b)). In this way, Swin-UNet combines the strengths of the two architectures. By adopting the patch-based processing strategy from Swin Transformer, it is allowed to handle large-scale images efficiently in global context. And leveraging the skip connections helps to refine the predictions at different resolutions.

In the experiments performed on the Synapse multi-organ segmentation dataset (Synapse), Swin-UNet achieved the highest average Dice-Similarity coefficient (DSC) of 79.13 as well as the lowest average Hausdorff Distance(HD) of 21.55, outperforming other state-of-the-art models like UNetreaching 76.85 for DSC and 39.7 for HD, Att-UNet with DSC of 77.77 and HD of 36.02 and TransUNET with DSC of 77.48 and HD of 31.69 (Ronneberger et al. 2015, (Cao et al. 2021, Oktay et al. 2018, Chen et al. 2021).

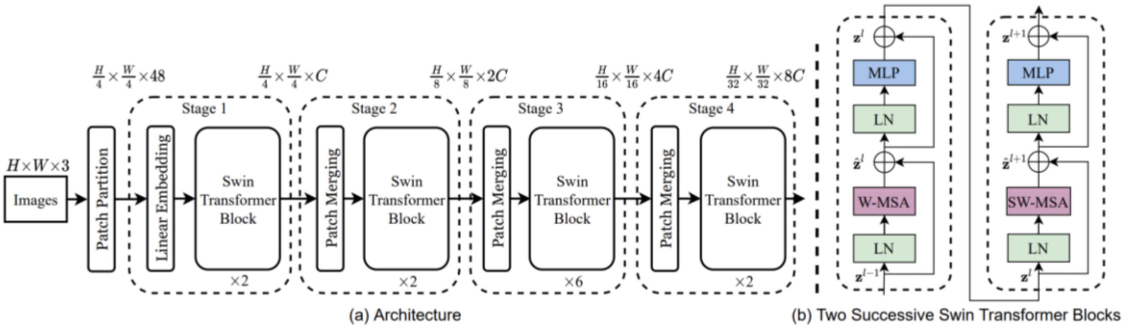

Figure 2: The framework of (a) a Swin Transformer and (b) its core blocks (Liu et al. 2021).

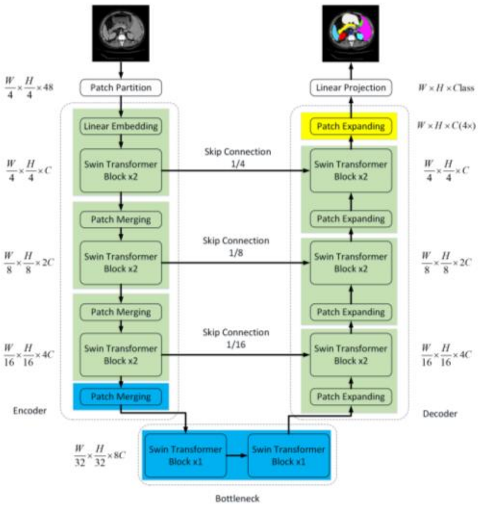

Figure 3: The Architecture of Swin-UNet (Cao et al. 2021).

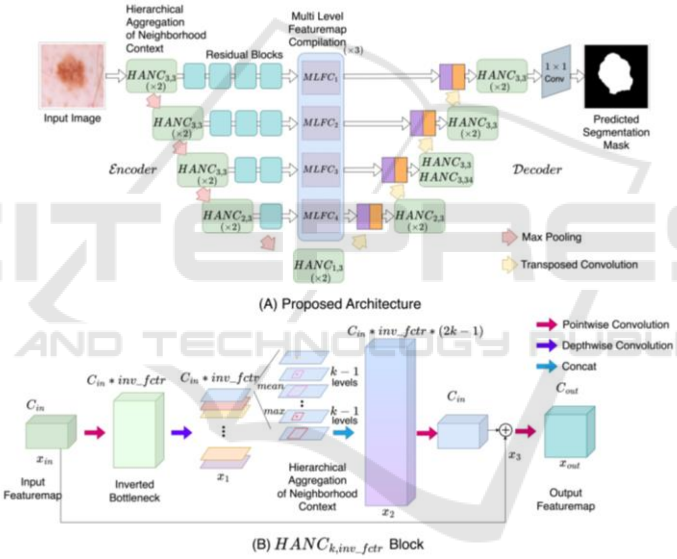

Figure 4: The Framework of (A) Acc-UNet and (B) its Core Block (Ibtehaz & Kihara 2023).

2.8 ACC-UNet

ACC-UNet is a recent extension of the UNet architecture and was proposed by Ibtehaz and Kihara in 2023 (Ibtehaz & Kihara 2023). As shown in Figure 4, it is a UNET-like model and proposed with designed convolutional blocks emulating self-attention mechanism (HANC block) and performing multi-level feature combination (MLFC block) by the authors motivated and inspired by transformer-based UNet models (Ibtehaz & Kihara 2023). Therefore, without integrating transformers into the architecture, ACC-UNet is still able to capture long-range dependencies and refine the feature representations as well as enhances the spatial resolution as completely convolutional UNet.

Performing segmentation tasks on several medical image datasets, ACC-UNet obtained all the best Dice scores and outperformed the others including UNet and Swin-UNet, though slightly. But notably, ACC-UNet uses only 59.26% number of Swin-UNet's parameters (Ronneberger et al. 2015, Cao et al. 2021).

3 CONCLUSION

Whether it is the time-honored UNet, the Swin-UNet that incorporates the popular transformer models, or the fully convolutional model that mimics the transformers, all these state-of-the-art models have obtained notable results in medical image segmentation. UNet, as an established solution, along with its variants, can obtain accurate and robust results on several tasks after training. Transformer-based models, which are good at capturing long-range dependencies and processing large-scale images with high efficiency, are more suitable for tasks requiring considering contextual information over a large spatial extent, such as organ segmentation, pathology localization, vascular segmentation, and other tasks. The more innovative models that utilize FCN to realize the advantages of transformers can also reduce the number of required parameters while considering the advantages, which is also a promising point of view. It is also a matter of choosing the right model for different tasks, taking into account data and hardware constraints. On the other hand, the experimental data presented in the papers show that the accuracy of the model is usually around 80% to 90%, which is not enough in the medical field that requires strict matching. Some special cases are difficult to overcome, such as pancreas image segmentation, which achieves relatively poor results compared to the others. Focusing on the models like Chat-GPT that are now revolutionizing the field of NLP, or even many parts of human life, it hope that one day a generalized medical image segmentation model that can be widely used in real applications will also be available and revolutionize the medical field.

REFERENCES

A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al, arXiv (Cornell University), (2020).

A. Vaswani, N. Shazeer, N. Parmar, et al, arXiv (Cornell University), 30, 5998–6008, (2017).

A. Z. Alali, K. H. Ali, Diyala Journal of Engineering Science, 17–29, (2022).

D. Bahdanau, K. Cho, Y. Bengio, arXiv (Cornell University), (2015).

D. Patil, S. G. Deore, IJCSMC, (2013).

F. Shamshad, S. Khan, S. W. Zamir, et al, Medical Image Analysis, 88, 102802, (2023).

H. Cao, Y. Wang, J. Chen, et al, arXiv (Cornell University), (2021).

H. Dong, G. Yang, F. Liu, et al, arXiv (Cornell University), (2017).

H. Thisanke, C. Deshan, K. Chamith, et al, Engineering Applications of Artificial Intelligence, 126, 106669, (2023).

I. H. Sarker, SN Computer Science, 2(6), (2021).

J. Chen, Y. Lu, Q. Yu, et al, arXiv (Cornell University), (2021).

K. Ramesh, G. Kumar, K. Swapna, et al, EAI Endorsed Transactions on Pervasive Health and Technology, (2018).

L. Alzubaidi, J. Zhang, A. J. Humaidi, et al, Journal of Big Data, 8(1), (2021).

N. Carion, F. Massa, G. Synnaeve, et al, arXiv (Cornell University), (2020).

N. Ibtehaz, D. Kihara, arXiv (Cornell University), (2023).

O. Oktay, J. Schlemper, L. L. Folgoc, et al, arXiv (Cornell University), (2018).

O. Ronneberger, P. Fischer, T. Brox, In Lecture Notes in Computer Science (pp. 234–241), (2015).

S. Ghosh, A. Chaki, K. C. Santosh, Physical and Engineering Sciences in Medicine, 44(3), 703–712, (2021).

X. Liu, L. Song, S. Liu, et al, Sustainability, 13(3), 1224, (2021).

Z. Liu, Y. Lin, Y. Cao, et al, 2021 IEEE/CVF International Conference on Computer Vision (ICCV), (2021).