Project proposal: TransUNet-based approach for Segmentation in Medical Images

Nguyen Thien Bao
Faculty of Information Technology
VNUHCM - University of Science
22127032@student.hcmus.edu.vn

Le Phuoc Phat
Faculty of Information Technology
VNUHCM - University of Science
22127322@student.hcmus.edu.vn

Vo Hoai Viet
Faculty of Information Technology
VNUHCM - University of Science
vhviet@fit.hcmus.edu.vn

***Abstract*—Medical image segmentation is an essential condition for enhancing medical systems. Convolutional Neural Networks and recently Vision Transformers have proved their excellent performance in various medical image segmentation tasks. TransUNet, which is a combination of Transformers and a U-shaped architectural model, becomes a strong alternative for the medical image segmentation. In this proposal, we will present some technical knowledge about the topic of segmentation in medical images in general and TransUNet in particular. Based on that, we will propose an approach and a plan to improve the existing drawbacks of the model.**

***Index Terms*—Segmentation in Medical Images, Deep Convolutional Neural Network, Vision Transformer.**

I. Introduction

Medical image segmentation plays an essential role in many medical fields, especially in medical diagnostics, treatment planning, and disease monitoring. Segmenting accurately the anatomical structures from medical images such as MRI, CT, and X-ray scans is necessary for automated clinical decisions. Thresholding, region-based and many traditional segmentation methods often struggle with the complexity of medical images, such as intensity, noise, and anatomical differences across patients.

Convolutional neural networks (CNNs) [1], [2] have proved their success in various computer vision tasks. U-Net [3], which is a CNN variant, has become one of the dominant approaches in medical image segmentation. The U-Net network architecture is designed as a U-shape with 2 branches: encoder and decoder. The encoder takes responsibility for down-sampling and extracting features from input images, while the decoder, combined with skip connections, is responsible for up-sampling and restructuring the extracted features to the segmentation masks.

In recent years, Transformers [4] have sparked a revolution in the field of natural language processing (NLP). Transformers have proposed a mechanism called self-attention, allowing the model to capture long-range dependencies and contextual relationships in an effective way. With the success of the self-attention mechanism, Transformers have not stopped at NLP but have also expanded into the computer vision field. Unlike the prior CNNs, which process images in a sequential way, CNN using a Transformer can process data in parallel, and capture features of images more effectively.

Inspired by the U-shaped architecture of U-Net and the self-attention mechanism of Transformers, TransUNet [3] has been proposed. With hybrid CNN-Transformer architecture, TransUNet can use Transformer to encode self-attentive features and then upsample at the decoder.

In this project, we will explore the mechanism of TransUNet. From that, we aim to enhance the TransUNet by integrating additional components into the original architecture. Moreover, we plan to experiment and fine-tune the modified model with different types of data. With these improvements, we hope to propose a model with better accuracy and performance.

II. Related Works

Segmentation in medical imaging has experienced remarkable advancements with the emergence of various deep learning architectures, each of which presents distinct the oretical foundations and trade-offs. Some structures below have been introduced as effective approaches to tackling the challenges of medical image segmentation.

A. U-Net Architecture

Ronneberger et al. [5] first introduced U-Net, which is a CNN with a U-shaped structure. The architecture of U-Net consists of 2 parts: an encoder symmetrizing with a decoder. The encoder consists of a series of 3 convolutional layers alternating with a max pool layer. When an input image of size 32x32 passes down, the encoder extracts it into a 1024-dimensional feature vector. Then the decoder, which consists of a series of 3 convolutional layers alternating with an up-convolutional layer, takes the extracted features and generates the output. The key idea in this architecture is the skip connection between the encoder and decoder. The network copies and crops the feature map at each of the 3 convolutional layers at the encoder and concatenates it with the feature map at each up-convolutional layer at the decoder. The purpose of these skip connections is to help the decoder recover the spatial structure of the input image for the output.

B. Transformer and Vision Transformer

In recent years, Transformer has experienced an upward trend in the success of AI models. Its architecture, first proposed by Vaswani et al. in 2017 [4], has initially reshaped many natural language processing tasks by relying entirely on self-attention mechanisms to model global dependencies without using any recurrent or convolutional layers. Building on this foundation, the Transformer has since been adapted for a variety of computer vision tasks, including object detection (DETR [6]), semantic segmentation (Cross-modal self-attention networks [7]), and image classification (Vision Transformer - ViT [8]). Regarding its structural composition, this architecture is divided into 2 main parts: encoder and decoder. The Transformer Encoders are responsible for analyzing and representing the input sequence in a way the model can understand by embedding the tokens and augmenting them with positional encodings, which preserve the order of the sequence. Its architecture typically consists of multiple layers, each of which includes a multi-head self-attention mechanism (which computes the importance of different input sequence parts by calculating the embeddings' dot product) and a feed-forward network (which allows the model to extract higher-level features from the input data and more compactly and usefully represent the input). The Transformer Decoder mirrors this structure but incorporates an additional cross-attention sub-layer between the self-attention and feed-forward components, enabling it to effectively integrate contextual information from the encoder. To the best of our knowledge, we aim to use Transformer in the improved model-based medical image segmentation framework, which builds upon the highly successful ViT.

III. Proposed Method

A. TransUNet Architecture

TransUNet, first introduced by Chen et al. [3] in 2021, is among the first models to integrate Transformer into medical image analysis. This hybrid approach merits both the global context encoded by Transformers and detailed high-resolution spatial information from U-Net features, which originated from CNNs. Below is the architecture of the TransUNet model introduced by Chen et al. [3]:

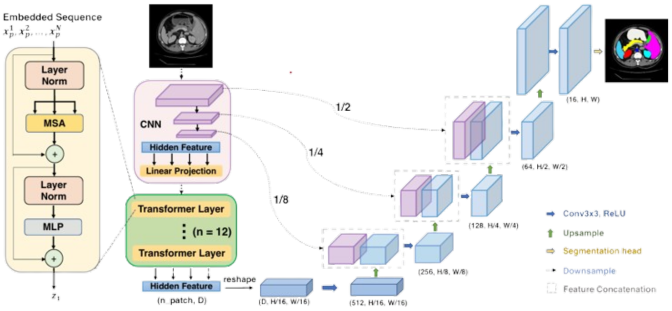

Fig. 1. TransUNet architecture [3]

In TransUNet, the encoder consists of 2 components: a series of Convolutional layers like the original U-Net encoder and a series ofTtransformer layers. When an input image passes down, the Convolutional layers first extract its hidden features. After that, the Transformer layers break up the extracted feature map into patches and use self-attention to find global dependencies. These Transformer layers can help to understand the context of the image better, complementing the local features extracted by the Convolutional layers. The features extracted by Transformer layers are then reshaped to the desired shape and continue to be restructured at the decoder with up-sampling and skip connection mechanism similar to the original U-Net.

B. Datasets

To evaluate and compare the TransUNet model with other models, we will experiment with various benchmarks that these models used for their evaluation. So far, we have learned about several benchmarks that can be used for this project.

Synapse multi-organ segmentation dataset [9]: Used by TransUNet [3] and DARR [10], the dataset is about 50 abdomen CT scans of were randomly selected from a combination of ongoing colorectal cancer chemotherapy trial, and a retrospective ventral hernia study. This data is labeled by 13 abdominal organs: spleen, right kidney, left kidney, gallbladder, esophagus, liver, stomach, aorta, inferior vena cava, portal vein and splenic vein, pancreas, right adrenal gland, and lef adrenal gland.

Automated cardiac diagnosis challenge (ADCD) [11]: Used by TransUNet [3], the dataset collects exams from different patients acquired from MRI scanners. A sequence of short-axis slices, with a slice thickness of 5 to 8 mm, span the heart from the base to the apex of the left ventricle in cine MR images that were obtained while the patient was on breath hold. The spatial resolution of the short-axis in-plane ranges from 0.83 to 1.75 mm2/pixel. Ground truth for the left ventricle (LV), right ventricle (RV), and myocardium (MYO) is manually labeled on each patient scan.

Pancreas CT dataset [12]: Used by AttnUNet [13], the dataset is about 82 abdominal contrast-enhanced 3D CT scans from 53 male and 27 female subjects. Before having a nephrectomy, 17 healthy kidney donors were scanned. A radiologist chose the remaining 65 individuals from among those without significant abdominal pathologies or pancreatic cancer lesions. Subjects' ages range from 18 to 76 years with a mean age of 46.8.

PROMISE12 dataset [14]: Used by V-Net [15], the data contains medical data acquired in different hospitals, using different equipment and different acquisition protocols. This data is about MR images of 50 patients with diverse illnesses that were acquired at different locations with several MRI vendors and scanning tools.

C. Evaluation metrics

In the image segmentation task, the objective of models is to specify whether each pixel in an image belongs to a meaningful segment. Depending on different tasks, several metrics are used to evaluate the accuracy of these segmentation models. For example, intersection over union, dice score, pixel error, and warping error.

Intersection over Union (IoU) is a commonly used metric in image segmentation. IoU measures the overlap between the intersection area of predicted segmentation A and ground truth B over the union area of them.

$$IoU = \frac{|A \cap B|}{|A \cup B|}$$

Dice score is another commonly used metric. Dice score measures the similarity between the predicted mask ant the true mask. The Dice score is calculated by doubling the area of overlap between them over the total pixels in both.

$$Dice = \frac{2|A \cap B|}{|A| + |B|}$$

Pixel error is the percentage of incorrect pixels in the predicted segmentation mask over the total number of pixels of the ground truth.

$$E_{pixel} = \frac{|A \neq B|}{|B|}$$

Warping error between two segmentations is defined as the minimum mean square error between the pixels of the ground truth and the pixels of a topology-preserving warped predicted segmentation.

D. Improvement directions

Transformer-based models, like TransUNet [3], have proved their power in modeling global context. They can outperform the CNN-based models, which have difficulty in capturing long-term relationships because of the local convolution. However, there still are some challenges to solve:

- Computational Cost: TransUNet is resource-intensive, so we need to find a way to optimize for practical deployment.
- Generalization across diverse medical images: The model cannot ensure consistent performance across many types of medical images.
- Enhancing segmentation accuracy: While the model outperforms many traditional CNN-based models, further enhancements in feature fusion and attention mechanisms can enhance its predictive power.

In this project, we will investigate the drawbacks of the TransUNet model by experimenting with diverse datasets and then evaluating and comparing it with previous ones. Once we identify areas to improve, we will attempt to enhance the model using methods such as fine-tuning new data or reconstructing the original model architecture.

IV. Planning Timeline

This is a detailed implementation plan with an estimated timeline for completing the project, starting from March 10, 2025 for 6 weeks. The plan is divided into specific stages as the following table:

Week	**Stage objectives**
1	Survey papers about the segmentation in the medical image topic and select a promising model among them.
2, 3	Experiment with the TransUNet model and its related works using diverse benchmarks to identify areas to improve.
4, 5	Modify the orignal architecture to find a new better architecture and fine-tune the model if necessary.
6	Revise and summarize the reseached results, complete the required source and documentation.

V. Challenges

Although the scale of this project is not large, our group will still encounter many difficulties in finding datasets, as well as limitations in terms of equipment such as a lack of GPUs, machine configurations that cannot withstand the heat load during model training, or a wrong understanding and direction in the process of implementing this project. Our team hopes that PhD. Vo Hoai Viet will review this proposal report and give us comments and satisfactory solutions for this topic.

References

[1] E. Shelhamer, J. Long, and T. Darrell, “Fully convolutional networks for semantic segmentation,” 2016. [Online]. Available: https://arxiv.org/abs/1605.06211

[2] K. O’Shea and R. Nash, “An introduction to convolutional neural networks,” 2015. [Online]. Available: https://arxiv.org/abs/1511.08458

[3] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” 2021. [Online]. Available: https://arxiv.org/abs/2102.04306

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[5] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-end object detection with transformers,” 2020. [Online]. Available: https://arxiv.org/abs/2005.12872

[7] L. Ye, M. Rochan, Z. Liu, and Y. Wang, “Cross-modal self-attention network for referring image segmentation,” 2019. [Online]. Available: https://arxiv.org/abs/1904.04745

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” 2021. [Online]. Available: https://arxiv.org/abs/2010.11929

[9] “Synapse multi-organ segmentation dataset,” Available at https://www.synapse.org/Synapse:syn3193805/wiki/217789, 2024.

[10] S. Fu, Y. Lu, Y. Wang, Y. Zhou, W. Shen, E. Fishman, and A. Yuille, “Domain adaptive relational reasoning for 3d multi-organ segmentation,” 2020. [Online]. Available: https://arxiv.org/abs/2005.09120

[11] “Automated cardiac diagnosis challenge dataset,” Available at https://www.creatis.insa-lyon.fr/Challenge/acdc/, 2024.

[12] “Pancreas ct dataset,” Available at https://www.cancerimagingarchive.net/collection/pancreas-ct/, 2024.

[13] J. Schlemper, O. Oktay, M. Schaap, M. Heinrich, B. Kainz, B. Glocker, and D. Rueckert, “Attention gated networks: Learning to leverage salient regions in medical images,” 2019. [Online]. Available: https://arxiv.org/abs/1808.08114

[14] “Prostate mr image segmentation 2012,” Available at https://zenodo.org/records/8026660, 2024.

[15] F. Milletari, N. Navab, and S.-A. Ahmadi, “V-net: Fully convolutional neural networks for volumetric medical image segmentation,” 2016. [Online]. Available: https://arxiv.org/abs/1606.04797