University of Science and Technology of Ha Noi

Information and Communication Technology Department

INTRODUCTION TO DEEP LEARNING

REPORT

Topic: Vision Transformer for medical image segmentation

Major: Data Science

Student Name	Student ID
1. Nguyen Minh Khoi	22BI13220
2. Nguyen Khac Cong	BA12-033
3. Dinh Tuan Kiet	22BI13229
4. Chu Hoang Viet	22BI13462
5. Bui Dang Quang	22BI13378
6. Nguyen Minh Tuan	22BI13447

Submission Date : 2/10/2024

Contents

1 Introduction

1.1 Motivation

Medical image segmentation is a fundamental task in medical imaging analysis, aimed at delineating anatomical structures and pathological regions from images such as MRI, CT scans, and ultrasound. Traditional approaches based on Convolutional Neural Networks (CNNs) have successfully addressed this challenge; however, these methods often need help capturing long-range dependencies and global contextual information, particularly in complex medical images with heterogeneous structures.

The Vision Transformer (ViT), initially developed for image classification, has gained attention for its remarkable ability to capture global features through the self-attention mechanism. Unlike CNNs, which focus on local receptive fields, ViT processes an image as a sequence of patches, allowing it to model relationships between distant regions within the image.

1.2 Goal

This report will concentrate on technical explanation and provide a concrete case study of the application of Vision Transformer in medical image segmentation.

2 Method

2.1 Transformer

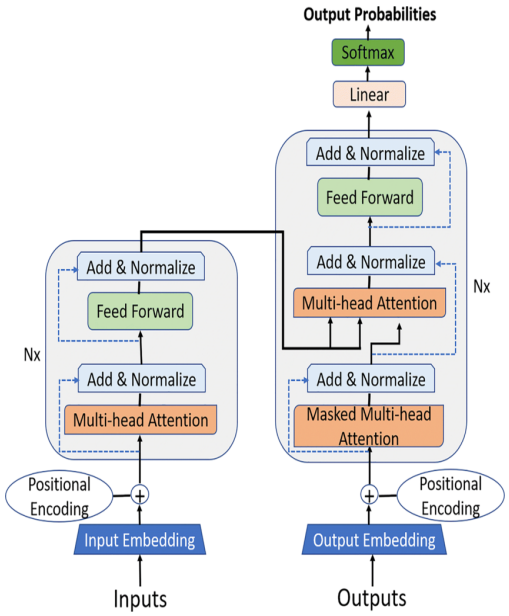

Figure 1: Transformer Architecture

The Transformer architecture consists of two main components: encoders and decoders.

2.1.1 Encoder

The Encoder is made by stacking N identical layers, each layer has 2 main components: a Multi-head Attention and position-wise fully connected feed-forward network. Each of layer is applied residual connection and normalized.

2.1.2 Decoder

The structure of Decoder is nearly similar to Encoder, it has additive layer to receive features from the Encoder. The first layer in Decoder has been modified to ensure that the predictions for position i can depend only on the previous position less than i.

2.1.3 Attention

An attention function maps a query along with a set of key-value pairs to an output, with the query, keys, values, and output all represented as vectors. The output is determined by a weighted sum of the values, where the weights are calculated based on a compatibility function that compares the query to each corresponding key.

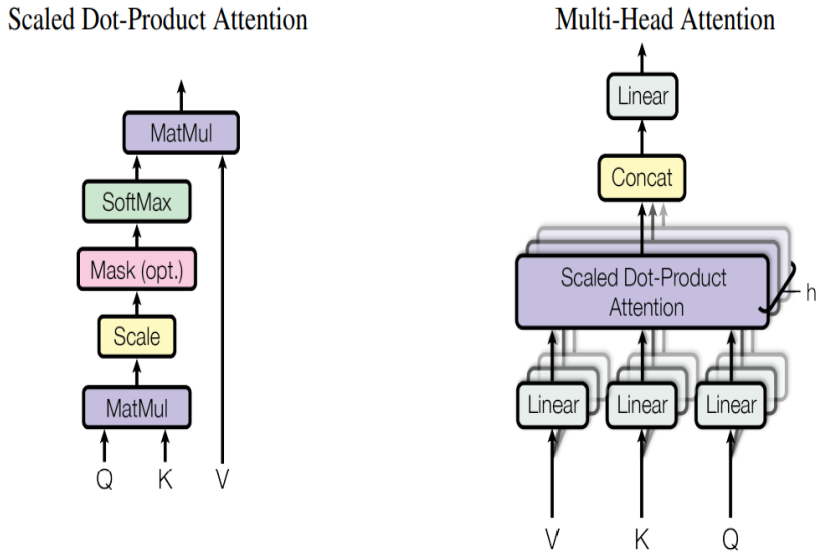

Figure 2: Scaled Dot-Product Attention & Multi-Head Attention

The formula for 2 functions in Figure 2 sequentially are:

- $Attention(Q, K, V) = Softmax(QK^T)V$
- $MultiHead(Q, K, V) = Concat(head_1, head_2, ..., head_n)$
  where $Head_i = Attention(Q_i, K_i, V_i)$

2.2 Vision Transformer (ViT)

2.2.1 Structure

Several proposed vision transformer models can be found in the literature. The general architecture of a vision transformer involves the following steps:

1. Split image into fixed-size patches and flatten these patches.
2. Generate linear embeddings of lower dimensions from the flattened patches.
3. Prepend [class] token to embedded patches.
4. Add positional information to embedded patches.

5. Input the resulting sequence into Transformer Encoder.
6. Extract the [class] token section of the Transformer Encoder output and feed into MLP Head.

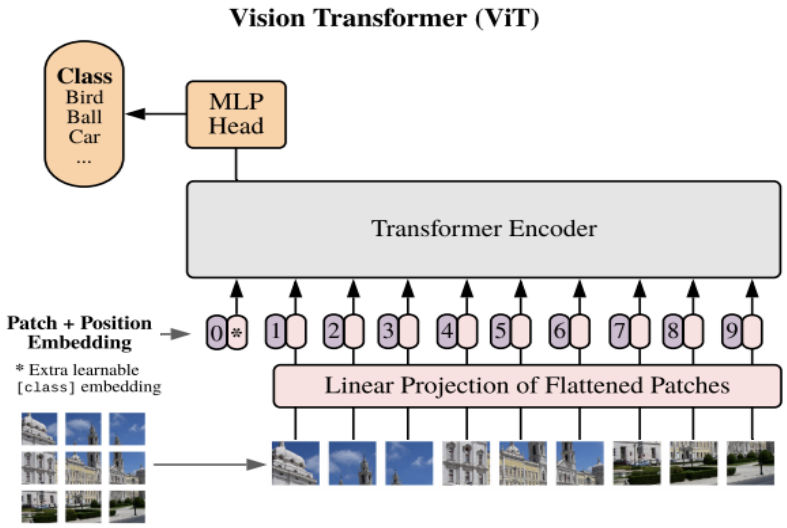

Figure 3: Structure of Vision Transformer

2.2.2 Vision Transformer (ViT) for segmentation

The Vision Transformer (ViT) model for segmentation tasks use the same patch embedding method and Encoder as classified tasks. In segmentation, the goal is to predict a label for each pixel, requiring spatially detailed output, while classification focuses on predicting a single label for the entire image. Therefore, the modification to adapt segmentation emerges in the Decoder section.

In the Decoder, we use upsampling layer to generate a new image has the same size as original and apply normalization and ReLU as activation function.

3 Case Study

3.1 Dataset

Brain Tumor Segmentation The dataset consists of 3064 paired MRI brain images with a fixed size of 512x512x3, where each image is matched with a corresponding binary mask. The MRI images represent various brain sections, and the binary masks highlight areas where tumors are present. Each mask is a pixel-wise annotation, where pixel values are either 0 (representing healthy tissue) or 1 (indicating tumor presence). The dataset is intended for tasks such as medical image segmentation, which aims to accurately identify and separate tumor regions from the surrounding brain tissue in the MRI scans.

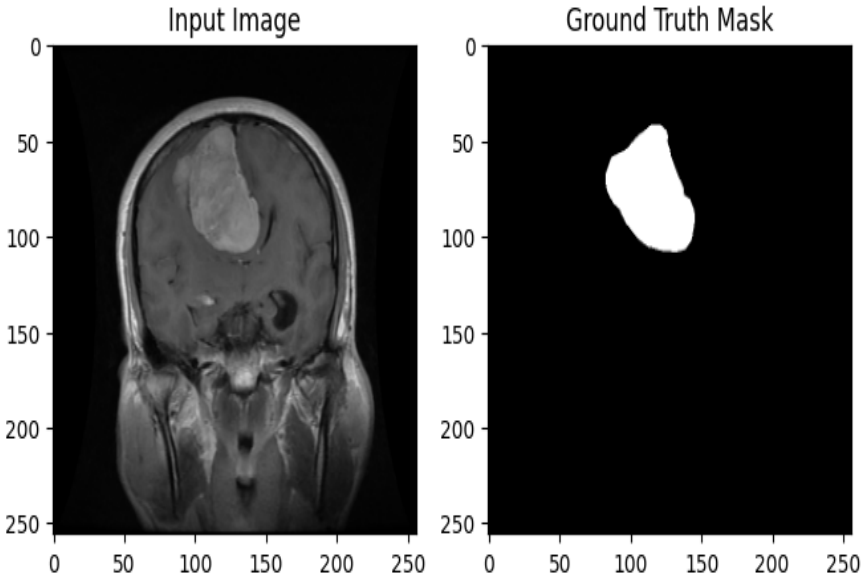

Figure 4: Input image and Mask

3.2 Model

In this section, We use a Vision Transformer for segment medical image, and we also propose Swin-Unet model, which is the combination of Unet and Swin Transformer architecture.

3.2.1 Vision Transformer

We have explained this model in the previous section of this report and connect them together to have a completed ViT model for medical images segmentation. Firstly, the input image is splitted into patches with positional embedding. Then we use linear projection and flatten those patches to 1D vector before passed through Encoder block. The Decoder block increases the dimension of the output from previous Encoder. Since our dataset has 2 class to segment for each pixel (tumor exist or not), we also use sigmoid function and threshold = 0.5 during inference.

3.2.2 Swin-Unet

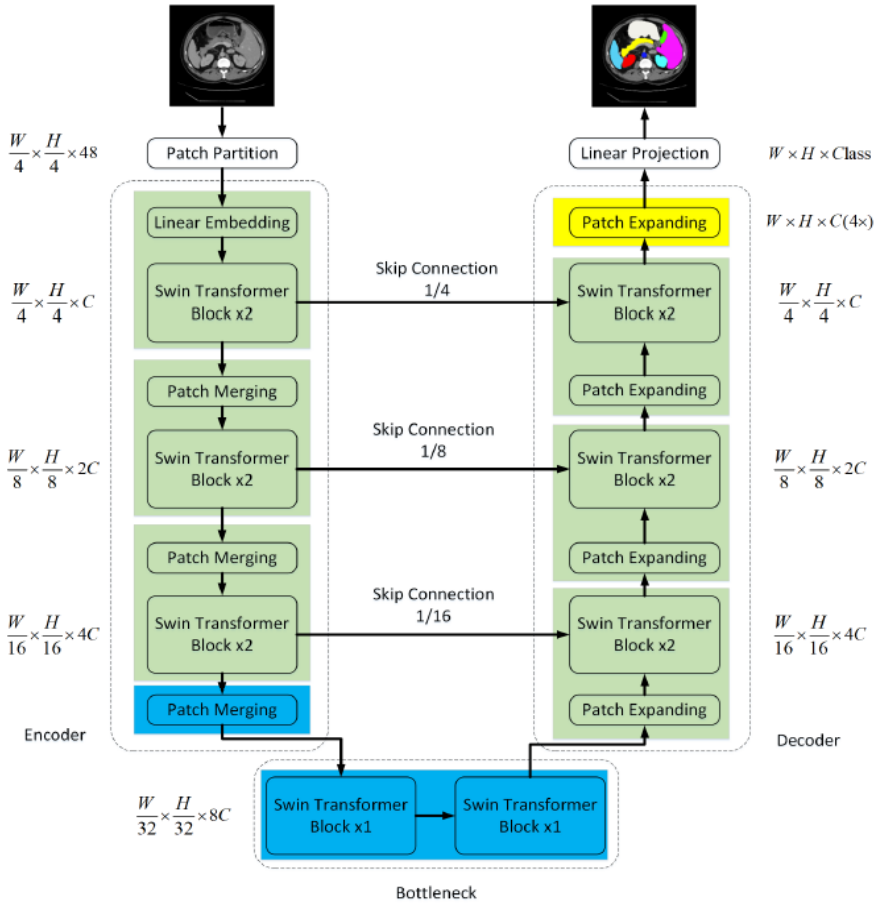

Figure 5: Swin-Unet architecture

Figure 4 will show the overall architecture of the proposed Swin-Unet. Swin-Unet consists of 3 components: Encoder, bottleneck, and decoder. The Swin-Unet architecture is fundamentally built upon the Swin Transformer block, which serves as its key component, driving its performance and functionality.

Encoder: The input brain image is divided into 4x4 patches, with each patch containing 48 pixels (4x4 spatial dimensions and 3 color channels). A linear embedding layer projects these patches into a feature size C. The features then pass through several Swin Transformer Blocks and Patch Merging layers, which downsample and double the feature dimensionality, refining the representation for further processing.

Decoder: Inspired by U-Net, is a symmetric transformer-based structure using Swin Transformer blocks and Patch Expanding layers. Skip connections from the encoder fuse context features with multiscale encoder details to recover spatial information lost during downsampling. Unlike merging layers, Patch Expanding is specifically designed for 2x up-sampling. The final layer performs 4x up-sampling to restore the feature map resolution to (W × H) x C(4x). Finally, a linear projection layer generates pixel-level segmentation predictions.

Swin Transformer block The Swin Transformer was developed as an alternative version to the conventional multi-head self-attention (MSA) mechanism used in the Vision Transformer, offering improved performance and efficiency for medical segmentation. This block is constructed based on shifted windows. Each Swin Transformer consists of 8 layer components:4 Normalize Layer (LN), 2 MLP Layer (Multilayer Perceptron), a module W-MSA (Window-Based multi-head self attention), a module SW-MSA (shifted window-based multi-head self

attention). Figure 2 presents two sequential Swin Transformer blocks.

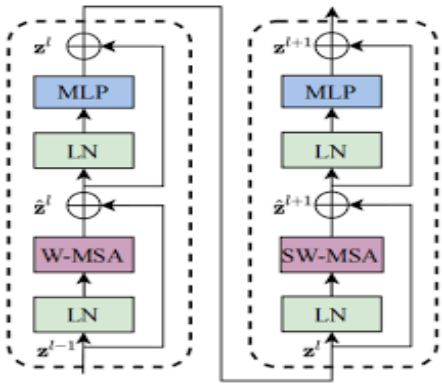

Figure 6: Swin transformer block.

3.2.3 Evaluation

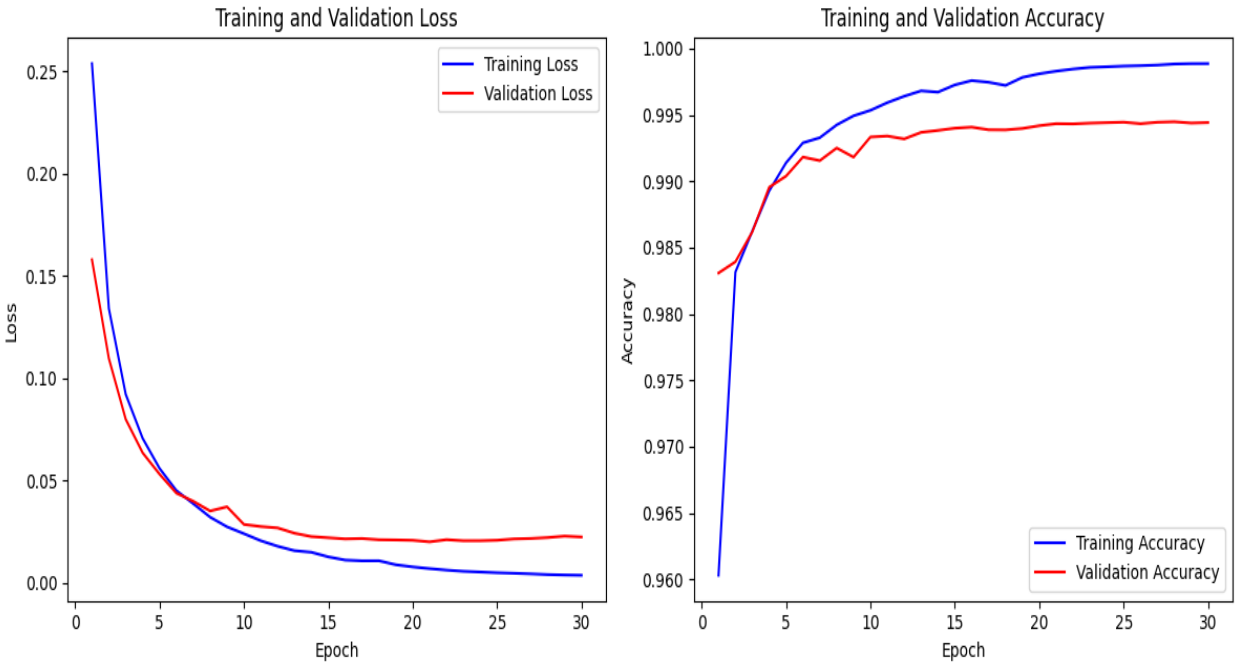

Figure 7: Training and Validation Loss and Accuracy Curves of ViT

- Training Accuracy of ViT: 0.9989
- Validation Accuracy of ViT: 0.9944

The accuracy on both training set and validation set is pretty high due to the cleaned data on prestige website.

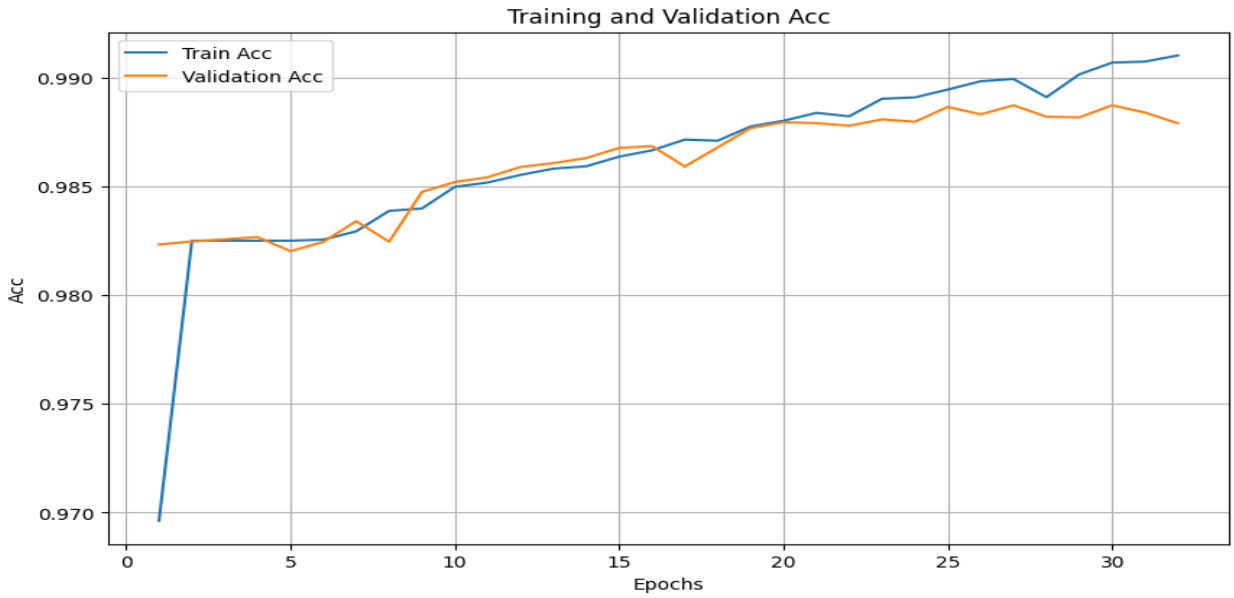

Figure 8: Result accuracy of Training set and Validation set of Swin-Unet

- Training Accuracy of Swin-Unet: 0.9855
- Validation Accuracy of Swin-Unet: 0.9846

We can conclude that the accuracy of ViT model is a little higher than Swin-Unet. The diference of accuracy between 2 model maybe depends on the architecture of Swin-Unet and ViT. While ViT model captures global dependencies of image, Swin-Unet focuses more on hierarchical feature extraction.

References

[1] James Bernhard. Alternatives to the scaled dot product for attention in the transformer neural network architecture, 2023.

[2] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton. Imagenet classification with deep convolutional neural networks. *Neural Information Processing Systems*, 25, 01 2012.

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation, 2015.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.