ViTResUNet: A Hybrid CNN-Transformer Architecture for Medical Image Segmentation

Radeen Mostafa, Syed Ashir Abrar

Department of Statistics, Shahjalal University of Science and Technology, Sylhet

"Towards a future of imagination and invention"
1st Jamal Nazrul Islam NATIONAL CONFERENCE-2022 for Young Researchers

Abstract

The use of transformers in computer vision tasks is still minimal. Most researchers still use convolution-based architecture (CNN) in this field. But thanks to Alexey et al. (2021) for implementing a pure transformer model, without the need for convolutional blocks, on image sequences to classify images and showcases how a ViT can attain better results than most state-of-the-art CNN networks on various image recognition datasets while using considerably fewer computational resources. Disease diagnosis and treatment planning are becoming more and more accurate thanks to Medical Image Segmentation playing a crucial role. Currently, U-Net is widely used in medical image synthesis and segmentation. This research uses a hybrid CNN-Transformer architecture-able to leverage both detailed high-resolution spatial information from CNN features and the global context encoded by Transformers.

Method

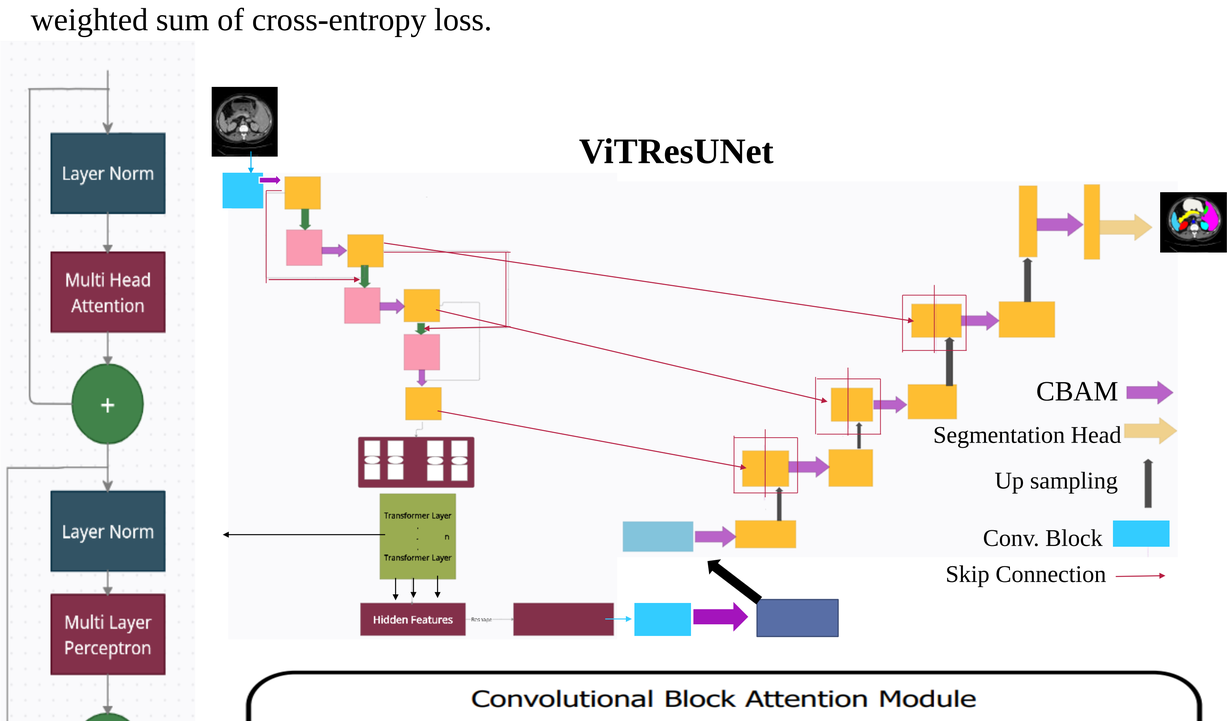
We propose a new encoder-decoder architecture called ViTResUNet, which equips the classical TransUNet with an extra key element: Convolutional Block Attention Modules (CBAM) [2]. We integrate CBAM, which consists of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM), to both the encoder and decoder path of the [1] TransUNet model, where CAM is calculated as:

$$\begin{aligned}\mathbf{M_c}(\mathbf{F}) &= \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))) \\ &= \sigma(\mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{avg}})) + \mathbf{W_1}(\mathbf{W_0}(\mathbf{F^c_{max}}))),\end{aligned}$$

and SAM is computed as:

$$\begin{aligned}\mathbf{M_s}(\mathbf{F}) &= \sigma(f^{7\times7}([AvgPool(\mathbf{F}); MaxPool(\mathbf{F})])) \\ &= \sigma(f^{7\times7}([\mathbf{F^s_{avg}}; \mathbf{F^s_{max}}])),\end{aligned}$$

For feature extraction, we introduce skip connection for feature reusability instead of using convolution for only down sampling. In particular, in the encoder path, CBAM is incorporated into the CNN part of the hybrid CNN-Transformer layer, and in the decoder path, it is placed after all convolution layers. This allows the model to perform both channel-wise and spatial-wise attention and, therefore, can better explore the inter-channel and inter-spatial relationship of the features. The channel attention essentially provides a weight for each channel and thus enhances those particular channels that contribute most to learning and thus boost the overall model performance. The loss function here is the weighted sum of cross-entropy loss.



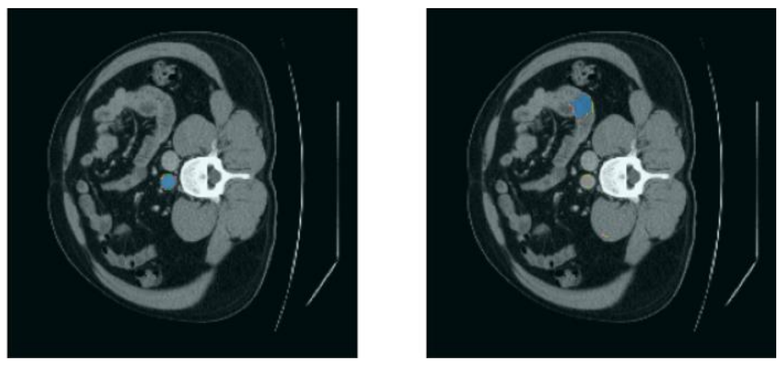
We have used Synapse Dataset with 512*512 images reduced to 224×224 using spline interpolation of order 3 and preprocessed it using NumPy format. We also used the pre-trained weight of ImageNet21k, and also R50_ViT was being used as an encoder.

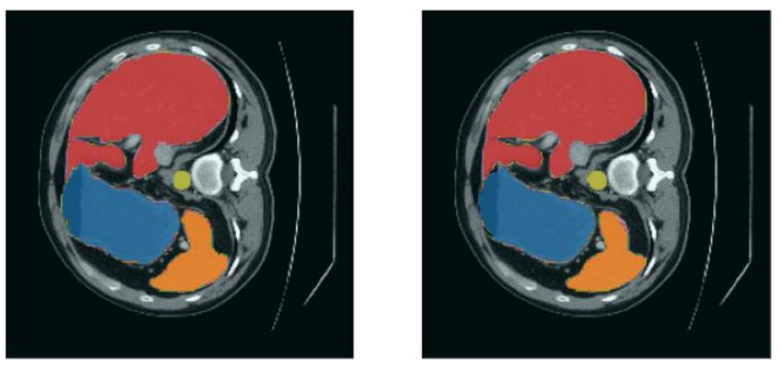
Introduction

With the development and increasing use of medical imaging modalities (X-ray, CT, MRI, Ultrasound, Microscopy, PET, Endoscopy, OCT, and many more) the tools for automating the information extraction from these images becomes as important as the modalities itself. Nowadays, most of the frequently used methods are based on Deep Learning. The U-Net is the basis for the most common architectures in medical image segmentation. The main disadvantage of CNNs is that if images contain structural information with large variations in shape and texture per image, CNNs tend to perform poorly. To overcome this shortcoming, in a Visions Transformer Network (ViT) was proposed for encoding. In this Project, we are going to reproduce the experiments done on the Synapse multi-organ segmentation dataset.

Result

Following are the obtained results of different metrics on the test set. The best results are presented in bold. From the table, we can see that our model achieves the lowest MSE value, also a 5% improvement compared to classical TransUNet. Though Recall is better in TransUNet and F1 score is better in UNet.





Original Label ViTUResNet

Model	MSE	Accuracy	Precision	Recall	F1
TransUNet	0.0212	0.867	0.637	**0.806**	0.705
ViTResUNet	**0.0203**	**0.872**	**0.671**	0.784	0.711
UNet	0.0233	0.854	0.666	0.792	**0.719**

Conclusion

Overall, we can see that the use of Vision Transformers for encoding can achieve state of the art results in biomedical image segmentation. The very strong global modeling from Vision Transformers can increase performance in segmentation, but lacks detailed information in the embedding. This is why it is necessary to use additional embeddings from CNNs to get performances of over 70% However, we could not observe that behavior for the other models, as the decoding might be not powerful enough to extract all necessary information from the embeddings of not trained models if skip connections are not used. Hence, we would suggest running experiments on additional datasets and using a more powerful decoder.

Reference

1. Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation, 2021.

2. S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 3–19.

GitHub Repository