Image Segmentation Using Transformer like U-Net with Attention

Moaaz Ur Rehman Azhar Khokhar - 0081002
Amir Mohamad Akhlaghi Gharelar - 0075302

Abstract

Image segmentation is an important and challenging task which deals with pixel wise classification to mark specific regions in an image. In this project, we propose to segment images using U-Net architecture which embeds Transformer like architecture leveraging attention (self and cross attention, both) mechanisms to focus on critical pixels. This network takes an image and gives the output image with pixels marked/segmented region. We initially propose to perform semantic segmentation on brain tumor MRIs and later we might use datasets from different domains (satellite imagery segmentation, pancreas segmentation, etc.) to test our model.

Introduction

Organ segmentation is of crucial importance in medical imaging and computed aided diagnosis, e.g. for radiologists to assess physical changes in response to a treatment or for computer assisted interventions.

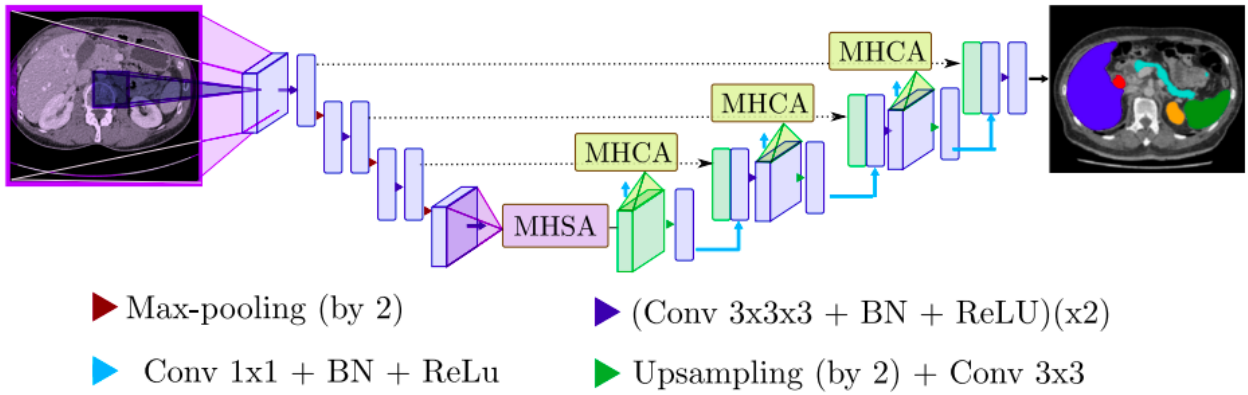
Fully convolutional networks are data hungry and they perform better when a huge amount of data is available. However, in the field of medical images we do not have as much data, comparatively. Moreover, as we can apply different transformations for data augmentation for real world images, not all of these augmentation techniques can be applied on medical images. For example, a lung MRI cannot be transformed horizontally or vertically as it would mean something else for the network.

U-Net (Ronneberger et al., 2015) was specifically designed to cater to this problem which can work on small training examples and specifically in the domain of medical image segmentation. It is a type of encode-decoder architecture where each decoder layer gets skip connection from its parallel encoder layer in the U shape, hence it is called u-net. It supplements a usual contracting network by successive layers, where pooling operators are replaced by upsampling operators. Hence, these layers increase the resolution of the output.

There is a problem in this architecture that when low-level layers are connected with high level layers in the decoder through skip connections, these skip connections bring low-level features. As we proceed along the decoder, later layers connect with earlier layers in the encoder part. To overcome this, we can add an Attention mechanism, proposed by (Vaswani et al., 2017), in our architecture to focus on special areas. Transformer implementation with Attention was implemented by (Dosovitskiy et al., 2020), commonly known as Vision Transformer (ViT). However, as we are using U-Net, we need to modify it according to the U-Net. One of the implementations was proposed by (Petit et al., 2021), where they add Transformer like self and cross attention blocks at skip connections. Hence, named U-Transformer.

Architecture and Methodology

The architecture which we intend to use for our project for U-Transformer is given in the following figure.



U-Transformer (Petit et al. 2021) augments U-Nets with transformers to model long-range contextual interactions. The Multi-Head Self-Attention (MHSA) module at the end of the U-Net encoder gives access to a receptive field containing the whole image (shown in purple), in contrast to the limited U-Net receptive field (shown in blue). Multi-Head Cross-Attention (MHCA) modules are dedicated to combine the semantic richness in high level feature maps with the high resolution ones coming from the skip connections. It implements the same Self and Cross attention modules as were mentioned by (Vaswani et al., 2017). We also thought of replacing Conv-blocks with residual Conv-blocks in order to test increase in accuracy (IoU).

We will be using PyTorch to implement this architecture and there is no official implementation available for this paper.

Dataset

We searched the web for suitable datasets and we came across the Br35H dataset on Kaggle (https://www.kaggle.com/datasets/ahmedhamada0/brain-tumor-detection). This dataset contains 800 images with training, validation, and test splits (500, 200, and 100 images respectively). This will be our initial dataset and we might change this dataset later, as discussed with Yücel hoca, in order to perform testing and validation on different datasets or a different problem domain.

References

Dosovitskiy, Alexey, et al. “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.” *arXiv*, 2020. *arXiv*, https://arxiv.org/abs/2010.11929.

Petit, Olivier, et al. “U-Net Transformer: Self and Cross Attention for Medical Image Segmentation.” *arXiv*, 2021. *arXiv*, https://arxiv.org/abs/2103.06104.

Ronneberger, Olaf, et al. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *LNCS*, vol. 9351, 2015, pp. 234-241, https://doi.org/10.1007/978-3-319-24574-4_28.

Vaswani, Ashish, et al. “Attention is All you Need.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. *NeurIPS*, https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.