Combination of U-Net and Transformer for Segmentation of Medical Images

CS766 project proposal

Introduction and Motivation

The medical image segmentation is an important part of current clinical workflows because the segmented parts can be used for accurate disease diagnosis and treatment planning. However, manual segmentation is time-consuming and labor-intensive considering the fact that the number of scan slices for each patient is more than one hundred. Therefore, it will be helpful if such data processing can be achieved automatically.

In this work, we will focus on the image segmentation tasks, especially for medical image segmentation. Given scans for each patient, we want to separate different organs in the scan. For example, the left lung, right lung, eso, heart, spinal cord parts are expected to be separated from each other in the CT scan of the abdomen area.

Method and Related Works

Recently, a number of studies have used deep learning and convolutional neural networks to do image segmentation. Among the many kinds of convolutional neural networks, the U-Net [1] has shown outstanding performance in medical image segmentation and synthesis. However, CNN-based approaches are limited by the intrinsic locality of convolution operations, which generally yields weak performances. Luckily, this problem can be solved by the Transformer-based model which can capture global features. Transformer [2] is a self-attention model, and its self-attention module allows for modeling long-range information by pairwise interaction between token embeddings and hence leading to more effective local and global contextual representations. Therefore, many researchers naturally think of combining U-Net with Transformer to improve the segmentation accuracy by incorporating the capture of global context into local features [3, 4, 5].

We are going to re-implement a state-of-art model called UNETR proposed recently for 3D medical image segmentation[5].

Evaluation and Comparison

Dice similarity coefficient (DSC) and Hausdorff Distance (HD) will be employed to evaluate the effectiveness of the implemented model[5]. DSC computes the overlap between the ground truth and prediction by:

$$\text{Dice}(G, P) = \frac{2\sum_{i=1}^{I} G_i P_i}{\sum_{i=1}^{I} G_i + \sum_{i=1}^{I} P_i},$$

where G represents the ground truth while P represents the prediction, and i denotes the indices of pixels.

HD metric is defined as:

$$\mathrm{HD}(G', P') = \max\left\{ \max_{g' \in G' p' \in P'} \left\| g' - p' \right\| \right.$$
$$\left. \max_{p' \in P' g' \in G'} \left\| p' - g' \right\| \right\}$$

Where G' and P' denote the surface points sets of ground truth and prediction regions, respectively. Thus the HD metric indicates the distance between the surface of ground truth and prediction.

The performance of the implemented model(combination of U-Net and transformer) will be compared to the conventional U-Net model, and the combined model is expected to outperform the simple U-Net model.

Milestones

02/24	Submit proposal
03/17	Finish the training of UNet model
03/31	Submit midterm report
04/10	Finish the training of UNet with Transformer model
04/26	Project presentation
05/01	Finish project website

References

[1] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015.
[2] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).
[3] Hatamizadeh, Ali, et al. "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images." arXiv preprint arXiv:2201.01266 (2022).
[4] Yan, Xiangyi, et al. "After-unet: Axial fusion transformer unet for medical image segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.
[5] Hatamizadeh, Ali, et al. "Unetr: Transformers for 3d medical image segmentation." Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022.