Brain Tumor Segmentation Methods using UNet Architectures

Koushik Sridhar [*1]

Abstract

This paper explores UNET and Swin-UNET architectures for the task of multi-region segmentation of brain tumors using stacked modalities and combined MRI scans. UNET and Swin-UNET models were developed using MONAI and PyTorch, and transformer blocks were implemented for Swin-UNET architecture. The models were trained and evaluated on the BRaTS 2021 Task 1 Dataset. Models were evaluated based on standard metrics of Dice Score and Loss quantification. Predicted segmentation outputs were generated. The findings from this work lay a baseline for future iterations of models, and establish that further model finetuning is needed and possible with expanded training.

1. Introduction

A brain tumor is a form of abnormal cell growth in the brain or central spinal tract, and can be a sign of one of the most life-threatening forms of cancer. The National Brain Tumor Society estimates that every year 13,000 patients die, and 29,000 patients suffer from primary brain tumors (Saouli et al., 2018). With this number consistently growing, the experts are consistently outnumbered by new cases.

With developments in imaging technology, imaging technology has been applied to tumor detection. Initially, computed tomography technology was used for detection, but with further developments in magnetic resonance technology and theories surrounding digital image reconstruction, magnetic resonance imaging is popular among experts as it does not cause ionizing radiation damage to the body (Wadhwa et al., 2019). Regardless of imaging type, brain tumor diagnoses are based on clinician experience. When a clinician analyzes an MRI, they undergo a process of manually segmenting, diagnosing and annotating the tumor. This process is highly inefficient and a demanding task for image analysts, leading to missed treatment windows. Therefore, this work seeks to aid in increasing the efficiency and reducing diagnostic errors involved in these manual processes.

Interdisciplinary works in medicine and machine learning have explored the integration of deep learning methods for tumor segmentation in the past. Autonomous brain tumor segmentation aims to computationally identify the size and location of brain tumors from MRI scans. Convolutional Neural Networks (CNN) have been explored by researchers, with results displaying good segmentation performance and convenient feature extraction (Hao et al., 2021). However, CNNs face issues when it comes to processing and analyzing a large number of dense images (Yang et al., 2020). Alternative models often require large amounts of annotated data or depend on heavily augmented data (Liu et al., 2021a). Therefore, lightweight deep learning architectures such as UNet have been proposed.

1.1. U-Net Architecture

UNets were first developed by Olaf Ronnenberg et al. (Ronneberger et al., 2015) for the purpose of biomedical image segmentation, and remained popular due to its accurate results and performance based on a smaller amount of training data. The model takes a form similar to the architecture of an auto-encoder, with a contracting path serving as an encoder, and an expanding path serving as a decoder. The contracting path is based on the structure of a CNN, and down-samples the input image. The expanding path is built with deconvolutional and convolutional layers, recovering the input image resolution using some optimized techniques such as concatenated skip connections. Through the expansion path, the network learns spatial classification information by generating predictions in a higher resolution. The output resolution is continuously increased and finally passed to a final convolutional layer which creates the segmented image in the same shape as the input image. Through this process, the network processes an input image with a specific shape (h, w, d) to generate an output image (h, w, d), while highlighting the segmented region of interest.

Within the contracting path, the model follows a typical CNN Network, consisting of successive convolutions followed by activations and pooling layers. This is repeated until reaching the bottleneck. The input is successively downsampled due to strided convolutions, but the number of channels is successively increased. Within the expansion path, each stage features up-convolutional and normal convolutional layers. With each upsampling, the number of channels is halved, and the up-convolution increases the width and height of the image. In addition, dimensions are

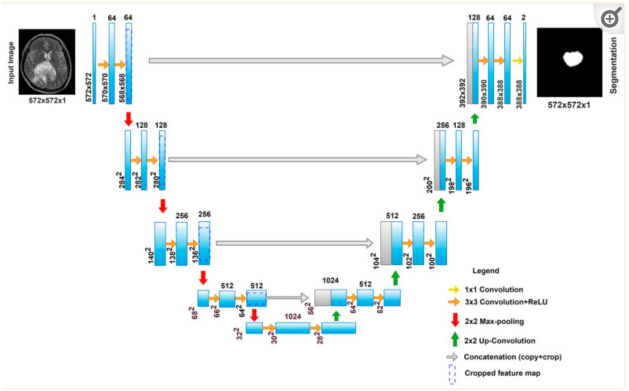

Figure 1. Basic UNet Structure for brain tumor segmentation (Yousef et al., 2023)

increased by the concatenation of feature maps from the down-sample block in the same layer within the contracting path. Named skip-connections, this method increases the efficiency of the model and recovers data that was lost with the down-sampling process. The final layer in the UNet is a final convolution that reduces the feature map to fit the sizing of the input.

1.2. Swin UNETR

Swin **UNE**t **TR**ansformers, alternatively known as Swin UNETR is proposed as a specific model for 3D brain tumor segmentation where the semantic segmentation problem is restructured into a sequence-to-sequence prediction problem. Specifically, the Swin UNETR methodology takes multi-modal input data and projects it into a 1D sequence of embedding, which is then passed to a Swin Transformer as an encoder (Hatamizadeh et al., 2021). The Swin Transformer encoder extracts feature maps at differing resolutions (classically at five different resolutions as a result of four depth layers) via shifted windows for computing self-attention, and is connected to a CCN-based decoder at each resolution via skip connections.

1.2.1. Swin Transformer

The Swin Transformer is a hierarchical transformer whose representation is computed with **S**hifted **Win**dows. The positive of this is that the shifted window scheme brings higher efficiency by limiting self-attention computation to non-overlapping local windows while allowing for cross-window connection. This allows for greater flexibility at higher-order scales, offering linear computational complexity increase as image size grows. (Liu et al., 2021b)

The transformer itself splits an input image into non-overlapping patches via a patch-splitting module. Then each patch (henceforth referred to as a "token") and its features are set as a concatenation of raw RGB values. A linear embedding layer is applied to project it to an arbitrary dimension. The set of tokens is modified with the application of several transformer blocks featuring modified self-attention computation. As the transformer is applied over various depths, the number of tokens is reduced by merging layers, reducing the count by a downsampling factor. This continues in the Swin UNETR architecture until the bottleneck structure of the network. The transformer block itself consists of a shifted window-based MSA (multi-headed self-attention) module which alternates partitioning configurations in consecutive Swin Transformer blocks, solving modeling power limitations found in base window-based self-attention modules.

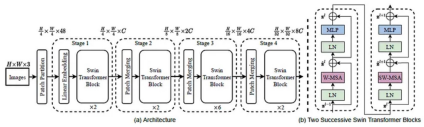

Figure 2. Architecture of Swin Transformer and Transformer Block (Liu et al., 2021b)

2. Methodology

2.1. Dataset Analysis and Preprocessing

This project utilizes the Brain Tumor Segmentation (BraTS) 2021 challenge dataset. The dataset is comprised of 1251 training and 219 validation samples of brain MRIs. The dataset consisted of mpMRI images in NIfTI format and included native (T1), post-contrast T1-weighted (T1CE), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes, along with manually annotated GD-enhancing tumor, peritumoral edematous/invaded tissue, necrotic tumor core, and normal tissue. (Baid et al., 2021).

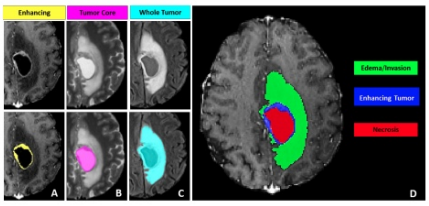

Figure 3. Typical tumor segments based on BRaTS dataset (Baid et al., 2021)

Each series of files associated with one MRI was stacked and standardized in preparation for training processing. By stacking the MRIs, volumes of the brain were merged to form a 4D array of the 3 modalities x length x width x number of slices, allowing for richer segmentation capabilities. The merged scans were saved, and mask feature labels were converted to labels of 0, 1, 2, 3. This data was divided into a training and validation split of 80:20.

2.2. UNet Model Development

In developing the UNet Model, several hyperparameters were explored, including batch size, dropout rate, optimizer functions, learning rate, and overlap. To prevent overfitting, dropout approach and batch normalization were applied for model regularization. Dropout rate was set to 0.2 for all encoder and decoder modules of the UNet. The batch size was set to 1 due to computational restrictions related to compute capacity. The small batch size enabled for regularizing effect, which typically leads to lower generalization error. (He et al., 2019) Adam optimizer was used as the optimizer function with a variable learning rate via a Pytorch learning scheduler based on Cosine Annealing. Finally, overlap was set to 0.5 similar to previous works (Noori et al., 2019).

In terms of model structure, the model was built with a high-level structure of 4 blocks down and 3 blocks up. On the encoder side, the model featured four layers of depth described as follows. The first layer of depth featured one convolution and a downsampling block. The second depth layer features two convolutions and a downsampling block. The third depth layer features two convolutions and a final downsample leading to the bottleneck. Each of the downsampling blocks downsample the feature set by a factor of two. In the bottleneck, the model features 4 convolution steps. After completion, the model moves to the decoder side, featuring three layers of depth, each with one upsample block followed by a convolution. The uneven structure between the encoder and decoder blocks fall in line with standardized UNet architectures as evidenced by previous works in the space.

2.3. Swin UNetr Model Development

In developing the SwinUNetr Model, the hyperparameters of batch size, fold, overlap, dropout rate, optimizer functions, and learning rate were explored. For this model, batch size was set to 2, and fold was set to 1 as a result of compute resource restrictions similar to those mentioned while developing the base UNet Model. Overlap was set to 0.5 as it is the standard in this field, and the dropout was set to 0. Adam optimizer was used as the optimizer function with a variable learning rate via a Pytorch learning scheduler based on Cosine Annealing.

In terms of model structure, the model was built with encoder and decoder halves, similar to the UNet mentioned previously, but follows the protocol used in previous standard works. The input to the Swin UNetr model is passed through as a token with a patch resolution of 2x2x2x4, with the 4th dimension in the patch courtesy of the four input channels in the MRI images from the BRaTS dataset. Then self-attention was computed into the non-overlapping windows that were created in the partitioning stage for efficient token interaction modeling. The encoder stage has 4 depth layers, where with each subsequent depth layer, the feature sizes of the representations are downsampled by a factor of 2.

The decoder half is built starting from the bottleneck at the end of the encoder step. Bottleneck output feature representations are reshaped to the dimensions of its encoder counterpart at the same depth layer and fed into a residual block comprised of two convolutional layers which are normalized by pooling layers. The resolution of the feature representations is upsampled by a factor of 2. Skip connections are added at each depth layer connecting encoder components "skipping" the bottleneck. The skip connection features are concatenated with the newly created feature representations and fed into the next residual block a depth layer above the current layer. The final segmented outputs are passed through a final convolutional block of size 1x1x1 and a sigmoid activation function.

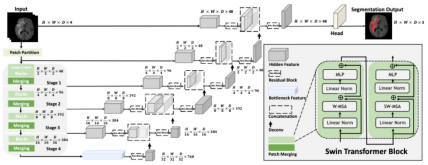

Figure 4. Swin UNETR architecture followed and created based on standard work. (Hatamizadeh et al., 2021)

2.4. Implementation Details

The Dice Loss (Equation 1) was utilized as an evaluation metric for both the training and testing phases. It calculates the ratio between the intersection and the union of the segmented and ground truth regions, focusing only on the segmentation classes and not the background class. In the equation, y is the true segmentation while $\overline{p}$ is the predicted segmentation.

$$DiceLoss\,(y, \overline{p}) = 1 - \frac{(2y\overline{p} + 1)}{(y + \overline{p} + 1)} \tag{1}$$

Both models were built using PyTorch and additional MONAI libraries. The models were trained on NVIDIA A100 Tensor Core GPUs with a 6-day cap on continuous training and 6GB of usable GPU memory via the University of North Carolina Longleaf Computing Cluster. As such, the UNET model completed 300 epochs of training, while the Swin UNETR model completed 100 epochs of training.

3. Results & Discussion

Best Dice Score metric results are presented in Table 1. The scores indicate that the developed models both present promising results in brain tumor segmentation. For the UNET model, the best model was achieved in epoch 223.

For the Swin-UNETR, the best model was achieved in epoch 100. Additionally, Figure 5 and 6 display the loss graphs and dice score curves for both UNET and Swin-UNETR models. These curves suggest that with further training Dice Scores and Loss can be improved upon with greater epochs of training.

Table 1. Best Dice Score among UNET and Swin-UNETR models

MODEL	DICE SCORE	LITERATURE VALUE
UNET	0.7506	0.899
SWIN-UNETR	0.7769	0.90

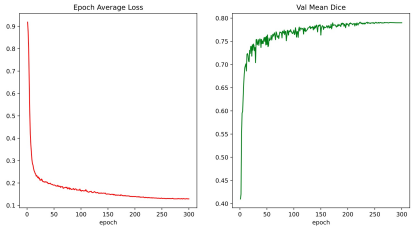

Figure 5. UNET Loss and Dice Score curves.

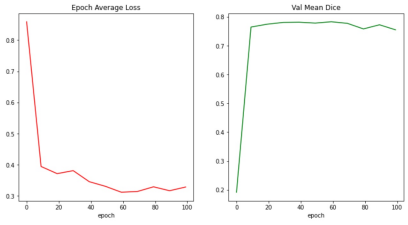

Figure 6. SWIN-UNETR Loss and Dice Score curves.

The results were further visualized by running segmentation tasks on sample MRIs. Figures 7 and 8 are the segmented MRI images from the UNET and Swin-UNETR models respectively. Overall, both models performed effective segmentations. However, the common trend between both models was that certain edge definitions or smaller edge variations would not maintain the specificity depicted in the label images. It is believed that with further training epochs, these issues may be resolved over time. Regardless, the segmented outputs appear to have generalizable positive impacts for segmenting brain MRI images.

Both methods have significant implications for early brain tumor detection, which is crucial for effective treatment and ultimately saving lives. With tumors being one of the leading causes of mortality worldwide, the model outputs are critical in detecting and forecasting tumor expansion. Early detection provides patients with the best chance for survival and successful treatment. The developed models help facilitate accurate and effective medical diagnostics and provide a basis for future research.

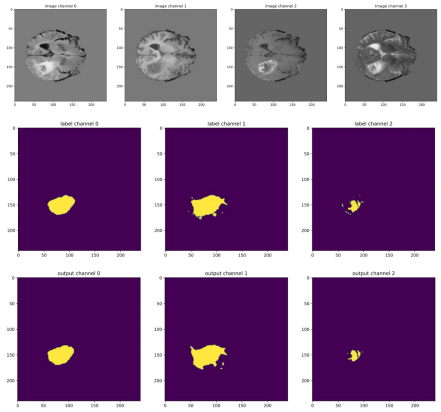

Figure 7. UNET Model Segmented MRI images. The first row features raw inputs of all MRI types (T1, T1CE, T2, FLAIR). The second row features labels among tumor channels (Enhancing, Tumor Core, Whole Tumor). The third row features the model's predicted segmented outputs.

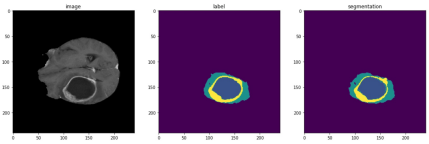

Figure 8. SWIN-UNETR Segmented MRI images. Moving left to right, the first image is the raw FLAIR input, the second image is the expected segmented output, and the third image is the model's predicted segmented output.

3.1. Limitations

The development of such models is dependent on the computing hardware that was used. During the development process, training was initially done on Google Colab TPUs. However, training times, efficiency and memory usage on TPUs were capped at 3 hours a session such that a maximum of 10 epochs of training would be completed within that timespan. Training shifted to the UNC Longleaf platform, with a max training session length of 6 days. Given the size of the UNET and Swin-UNETR models, this allowed for a limited number of epochs (300 and 100 respectively) to be trained upon. The number of epochs falls short of other standard works for this model architecture as well. Future work will be done looking to optimize training methods and extend epochs trained.

4. Conclusion

In this paper, UNET and Swin-UNETR GAN Models were implemented for the task of segmenting brain MRI images. The data was preprocessed by stacking modalities of the MRI scan in the NIfTI format to achieve a richer feature representation. The work exhibits the benefits possible using GAN methodologies, and the improvements found when using Transformer architecture for segmentation tasks. This

stacking of modalities also facilitated the one-time segmentation tasks across both models. The solutions developed in this project contribute further to the future development of accurate segmentation tools for brain tumors, allowing physicians to develop effective treatment plans for patients based on tumor regions observed from the accurate segmentations obtained.

Acknowledgements

The author would like to thank Jorge Silva for sponsoring this independent project, providing resources to explore MLOps and alternative model structures, and providing introductory knowledge on Machine Learning via previous courses. The author would also like to thank Martin Styner for his medical-based expertise and for providing introductory knowledge related to GAN implementations in neurological contexts.

References

Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F. C., Pati, S., Prevedello, L. M., Rudie, J. D., Sako, C., Shinohara, R. T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Moawad, A. W., Coelho, L. O., McDonnell, O., Miller, E., Moron, F. E., Oswood, M. C., Shih, R. Y., Siakallis, L., Bronstein, Y., Mason, J. R., Miller, A. F., Choudhary, G., Agarwal, A., Besada, C. H., Derakhshan, J. J., Diogo, M. C., Do-Dai, D. D., Farage, L., Go, J. L., Hadi, M., Hill, V. B., Iv, M., Joyner, D., Lincoln, C., Lotan, E., Miyakoshi, A., Sanchez-Montano, M., Nath, J., Nguyen, X. V., Nicolas-Jilwan, M., Jimenez, J. O., Ozturk, K., Petrovic, B. D., Shah, C., Shah, L. M., Sharma, M., Simsek, O., Singh, A. K., Soman, S., Statsevych, V., Weinberg, B. D., Young, R. J., Ikuta, I., Agarwal, A. K., Cambron, S. C., Silbergleit, R., Dusoi, A., Postma, A. A., Letourneau-Guillon, L., Perez-Carrillo, G. J. G., Saha, A., Soni, N., Zaharchuk, G., Zohrabian, V. M., Chen, Y., Cekic, M. M., Rahman, A., Small, J. E., Sethi, V., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Freymann, J. B., Kirby, J. S., Wiestler, B., Crivellaro, P., Colen, R. R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.-A., Mahajan, A., Menze, B., Flanders, A. E., and Bakas, S. The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification, 2021.

Hao, K., Lin, S., Qiao, J., and Tu, Y. A generalized pooling for brain tumor segmentation. *IEEE Access*, 9:159283–159290, 2021.

Hatamizadeh, A., Nath, V., Tang, Y., Yang, D., Roth, H. R., and Xu, D. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pp. 272–284. Springer, 2021.

He, F., Liu, T., and Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. *Advances in neural information processing systems*, 32, 2019.

Liu, X., Song, L., Liu, S., and Zhang, Y. A review of deep-learning-based medical image segmentation methods. *Sustainability*, 13(3):1224, 2021a.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021b.

Noori, M., Bahri, A., and Mohammadi, K. Attention-guided version of 2d unet for automatic brain tumor segmentation. In *2019 9th international conference on computer and knowledge engineering (ICCKE)*, pp. 269–275. IEEE, 2019.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pp. 234–241. Springer, 2015.

Saouli, R., Akil, M., Kachouri, R., et al. Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in mri images. *Computer methods and programs in biomedicine*, 166:39–49, 2018.

Wadhwa, A., Bhardwaj, A., and Verma, V. S. A review on brain tumor segmentation of mri images. *Magnetic resonance imaging*, 61:247–259, 2019.

Yang, T., Song, J., Li, L., and Tang, Q. Improving brain tumor segmentation on mri based on the deep u-net and residual units. *Journal of X-ray Science and Technology*, 28(1):95–110, 2020.

Yousef, R., Khan, S., Gupta, G., Siddiqui, T., Albahlal, B. M., Alajlan, S. A., and Haq, M. A. U-net-based models towards optimal mr brain image segmentation. *Diagnostics*, 13(9):1624, 2023.