POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Brain MRI Tumor Segmentation with Vision Transformers

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Massimo Maitan

Advisor: Prof. Daniele Loiacono

Co-advisor: Leonardo Crespi

Academic year: 2022-2023

1. Introduction

This thesis aims to explore recently proposed Transformer-based architectures to perform brain tumor segmentation on MRI scans. The investigation will specifically focus on two approaches: the analysis of individual images obtained by slicing magnetic resonances on the axial plane (referred to as "slices") and the analysis of entire three-dimensional MRI scans. The primary objective of this research is to train models to identify three nested tumor subregions as defined the BraTS challenge: Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). In the context of brain tumor segmentation, WT identifies the complete extent of the disease. On the other hand, ET delineates the region that becomes more visible in the T1 modality when a contrast agent, typically Gadolinium, is administered to the patient. This region is often the most actively growing part of the tumor. TC encompasses both the necrotic region (representing the inner core of the tumor composed of dead cells) and the ET. Both classes of models in this study are trained on a subset of the BraTS2019 challenge. In the upcoming section, we will delve into fundamental concepts related to MRI and the BraTS Challenge. Subsequently, we'll outline the experimental setup for our study. The results obtained on the test set will serve as a basis for comparing the effectiveness of the two approaches. Additionally, these outcomes will play a pivotal role in assessing the potential of the Transformer architecture in comparison to other state-of-the-art architectures that are conventionally based on Convolutional Neural Networks (CNNs). Specifically, our observations indicate that, in the context of brain tumor segmentation, Transformer-based approaches utilizing full three-dimensional scans generally exhibit superior performance compared to CNN-based architectures for predicting the Whole Tumor and Tumor Core subregions.

2. Background and State of the art

Brain tumor segmentation is a crucial aspect of disease diagnosis and treatment. Over the years, Convolutional Neural Networks (CNNs), like those proposed by Myronenko and Hatamizadeh [3] and the 3D U-Net introduced by Wang et al. [6], have been widely used for this purpose. Recently, Transformers [5] have also shown success in image analysis. In 2021, the Google Brain team introduced the Vision Transformer (ViT, [1]), utilizing Transformer capabilities to capture long-range dependencies in image analysis. ViT has demonstrated superior performance and computational efficiency compared to Convolutional Neural Networks, especially with extensive datasets. In scenarios with limited data, such as in this thesis, a combination of data augmentation, regularization, and transfer learning can address these constraints. Based on this, some intriguing Transformer-based architectures [4] have been proposed for image segmentation, in some cases specifically tailored for glioma segmentation [2].

3. MRI and BraTS dataset

Magnetic Resonance Imaging (MRI) is a medical imaging technique that exploits the magnetic properties of the human body, particularly hydrogen atoms. By applying magnetic fields and measuring two distinct response times, T1 and T2, this method generates detailed body scans. There are four fundamental measurements, or "modalities": T1, T1ce (T1 with the administration of a contrast agent), FLAIR (Fluid Attenuated Inversion Recovery), and T2. Annually, the MICCAI (Medical Image Computing and Computer Assisted Intervention) society organizes the **BraTS** (Brain Tumor Segmentation) challenge. This competition focuses on segmenting brain tumors in provided MRI scans, encompassing T1, T1ce, T2, and FLAIR modalities. It aims to identify sub-regions like the Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (ET). Participants develop and assess algorithms to enhance the accuracy of automated brain tumor segmentation, contributing to advancements in medical image analysis. The dataset for the BraTS2019 challenge consists of MRI scans from 335 patients for the training set, 125 cases for the validation set, and 166 cases for the test set. For our experiments, we have utilized a subset of this dataset, comprising 227 patients for training, 49 patients for validation, and 59 patients for the final test.

4. Experimental design

The models were trained using a NVIDIA TITAN V, generously provided by Politecnico di Milano, and on A100 and V100 GPUs rented through Google Colab Pro.

4.1. Segmenter - 2D slice analysis

The first architecture chosen for this experiment draws inspiration from the work of Robin Strudel et al. [4], who introduced the Segmenter — a Vision Transformer extension designed for semantic segmentation tasks. Employing a ViT as the encoder, this model is intended to extract features from RGB input images. Subsequently, it utilizes these features to generate full-resolution segmentation maps through a dedicated decoder known as the Mask Transformer. Since our dataset is not composed of RGB images but rather a set of four grayscale images (one for each modality) for each slice, we have chosen to preprocess the dataset and obtain RGB images through two distinct paths:

- **Consequent-slice merging:** three consecutive grayscale images representing three subsequent slices of a given modality are merged into one RGB image.
- **Modality merging:** we chose three of the four modalities and use them to construct a single RGB image. For this experiment, we selected T1, T1ce and FLAIR.

Each of the 227 training MRI scans, as well as the 49 and 59 patients for validation and test, has been partitioned into 128 slices, cut on the axial plane. Specifically, the training set underwent a cleaning phase, during which slices containing only background information were removed.

The trained models vary in terms of the number of decoder and encoder layers, the number of heads, embedding space dimension and preprocessing technique applied to the dataset. In total, we have trained ten models:

- *cons-t1-256-4-4-2*
- *cons-t1-256-8-8-2*
- *cons-flair-256-8-8-2*
- *mod256-8-8-2*
- *mod256-8-8-4*
- *mod256-16-16-8*
- *pret192-12-3-8*
- *pret768-12-12-8*
- *mod-wt256-2-2-8*
- *mod-wt256-8-8-4*

The prefix of each model indicates the preprocessing strategy applied:

- **cons:** Consequent slice merging
- **mod:** Modality merging technique
- **pret:** Fine-tuning on a pre-trained back-

bone

- **mod-wt:** Modality merging on a binary dataset designed specifically for the whole tumor problem

For the *cons* models, an additional label indicates the chosen modality. The second number in each model name indicates the embedding space dimension, while the subsequent three numbers represent the number of encoder layers, the number of attention heads, and the number of decoder layers, respectively. Each model employs a patch size of 6×6 pixels and a batch size of 10, and has been trained for 3,000 epochs.
There are some exceptions:

- *pret192-12-3-8* and *pret768-12-12-8* models have patch sizes of 16×16 and 32×32, respectively.
- *mod256-16-16-8*, *mod-wt256-2-2-8*, and *mod-wt256-8-8-4* were trained with a batch size of 1.
- *mod256-16-16-8* has been trained for 1,000 epochs due to its high computational requirements.

For each trained model, we employed the same combination of dropout and stochastic depth techniques for regularization. Additionally, we incorporated various data augmentation techniques, including resizing, random flipping, and photometric distortion.

4.2. SwinUNETR - 3D scan analysis

For the second experiment, we opted for the SwinUNETR architecture [2], which analyzes the four T1, T1ce, T2, and FLAIR entire scans of each single patient (processed as Nifti files) to generate the segmentation map. Notably, this architecture harnesses the Swin Transformer, an algorithm that calculates attention on input patches by utilizing shifting windows at various patch resolutions. Following this, a CNN-based decoder processes the contextualized intermediate representations of the input to generate the final 3D segmentation map.
In particular, we have trained from scratch three models:

- *SWIN96-48*
- *SWIN64-48*
- *SWIN96-60*

In this case, the first number in the model's name denotes the size of the Region of Interest used for random cropping data augmentation and sliding window inference, while the second number represents the embedding space dimension of the model. The models underwent training for 2000 epochs using the AdamW optimizer and incorporated various data augmentation techniques, including foreground cropping, random cropping, random flipping, and intensity scale/shifting. However, each of the three models experienced overfitting issues before the completion of the training. For this reason we have chosen, for each model, the best checkpoint for assessing the test set score.

5. Results

The results on the test set of the experiments are condensed in table 1, which reports, for each subregion, the average of the DICE score computed for each patient.

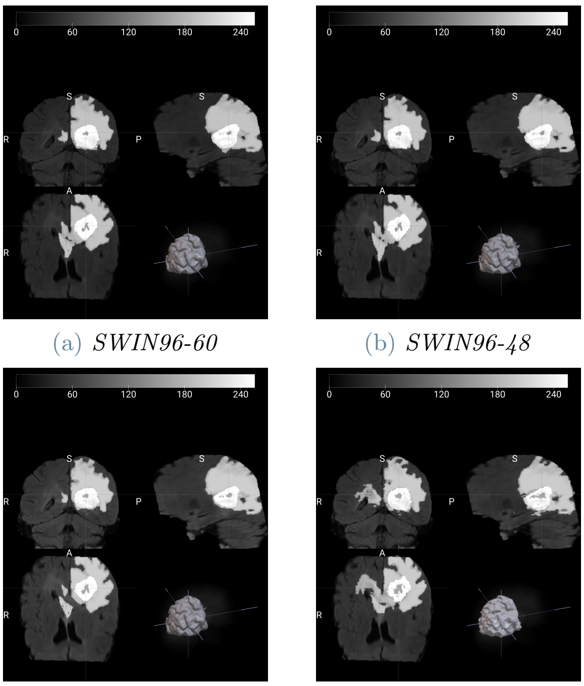
Brats2019 DICE results comparison

	WT	**TC**	**ET**
cons-t1-256-4-4-2	67.47	43.11	16.85
cons-t1-256-8-8-2	69.21	42.63	17.96
cons-flair-256-8-8-2	81.26	46.23	22.89
mod256-8-8-2	84.75	75.08	62.28
mod256-8-8-4	84.81	74.61	62.70
mod256-16-16-8	83.11	72.06	61.18
pret192-12-3-8	**86.56**	**80.14**	**64.47**
pret768-12-12-8	83.91	75.45	55.47
mod-wt256-2-2-8	85.94	-	-
mod-wt256-8-8-4	85.47	-	-
SWIN96-48	89.07	**85.24**	75.01
SWIN64-48	87.77	84.00	73.77
SWIN96-60	**89.32**	85.04	**76.16**
CNN-Myronenko [3]	88.20	**83.70**	**82.60**
CNN-3D U-Net [6]	85.20	79.80	77.80

Table 1: Brats2019 challenge comparison

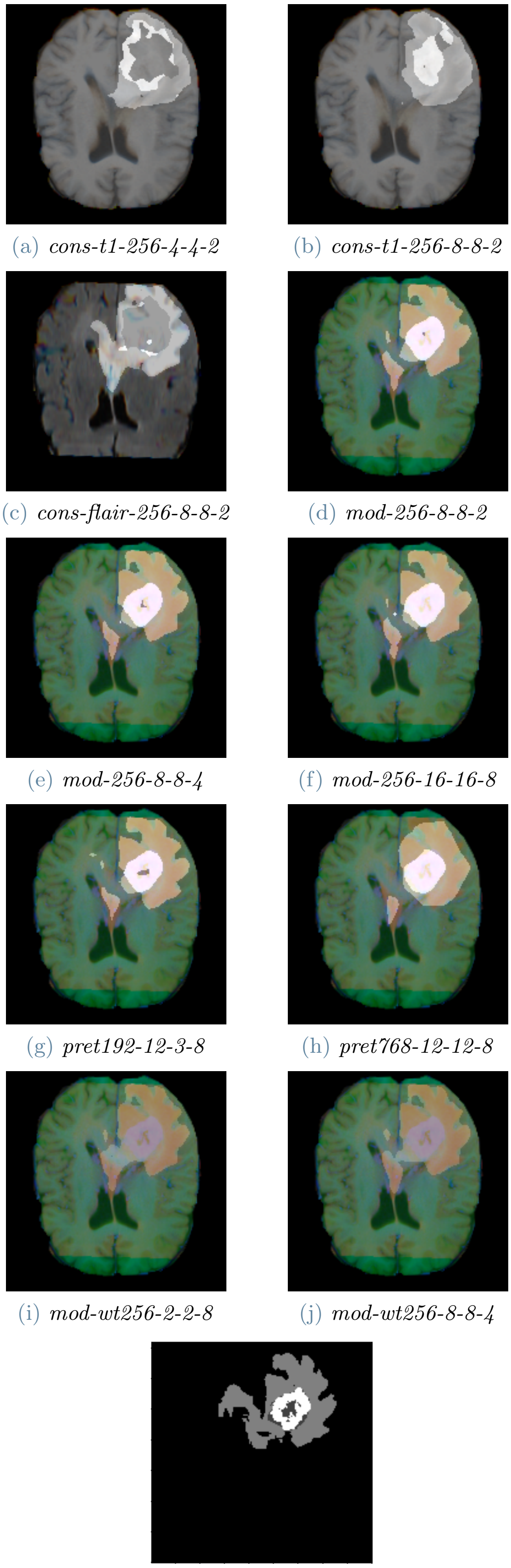
To provide a comprehensive overview of our model's overall performance, we have included the test set results reported by other CNN-based works ([3, 6]) in the table.
From this comparison, it is evident that the Swi-

nUNETR models consistently outperform the Segmenter models across all subregions. They excel in predicting Whole Tumor and Tumor Core, surpassing the performance of CNN-based architectures. However, there is still a challenge in achieving comparable results in the Enhancing Tumor region.
Conversely, Segmenter models encounter difficulties in matching the results of CNN-based approaches, particularly in the Enhancing Tumor region. This discrepancy might be attributed to the fact that the power of Transformer models lies in the ability to capture long-range dependencies. This strength is not fully leveraged when utilizing 2D input images. Notably, both CNN architectures considered in this comparison analyze three-dimensional scans, emphasizing the importance of exploiting the third dimension for improved performance.
The superior capability of Transformers in leveraging 3D inputs becomes more apparent in the following visual examples, where the predictions of SwinUNETR are compared with those of Segmenter models.



(a) *SWIN96-60* (b) *SWIN96-48* (c) *SWIN64-48* (d) Ground truth

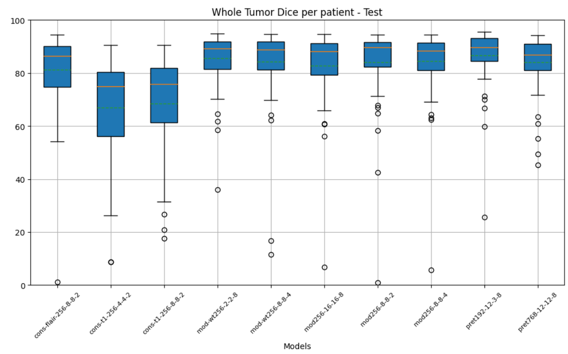
Figure 1: SwinUNETR models prediction on slice $67 \times 126 \times 73$



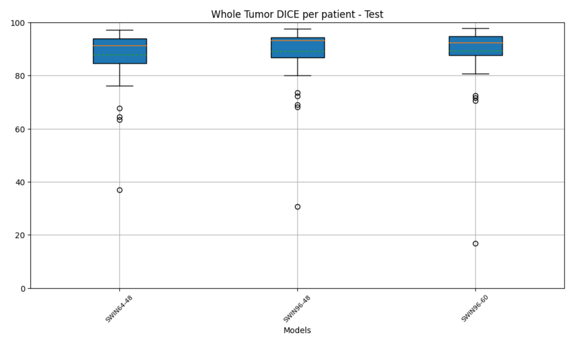
(a) *cons-t1-256-4-4-2* (b) *cons-t1-256-8-8-2* (c) *cons-flair-256-8-8-2* (d) *mod-256-8-8-2* (e) *mod-256-8-8-4* (f) *mod-256-16-16-8* (g) *pret192-12-3-8* (h) *pret768-12-12-8* (i) *mod-wt256-2-2-8* (j) *mod-wt256-8-8-4* (k) *Ground truth*

Figure 2: Segmenter models prediction on slice 73

Furthermore, both quantitative and qualitative assessments unequivocally demonstrate the limited effectiveness of Segmenter models trained with consequent-slice merging techniques, in contrast to the superior performance observed in models employing modality-merging techniques. This underscores the critical importance of cross-modal analysis for accurate identification of various tumoral tissues. Moreover, the disparity in performance between Segmenter and SwinUNETR appears to diminish in pretrained models, highlighting the effectiveness of transfer learning in the realm of medical image analysis.
Further insights into the per-patient DICE score distributions of the Segmenter and SwinUNETR models can be gained by visualizing them through box-plots. Notably, these diagrams reveal that SwinUNETR models not only deliver consistent accuracy but also exhibit reduced variance, indicating enhanced robustness in their predictions.

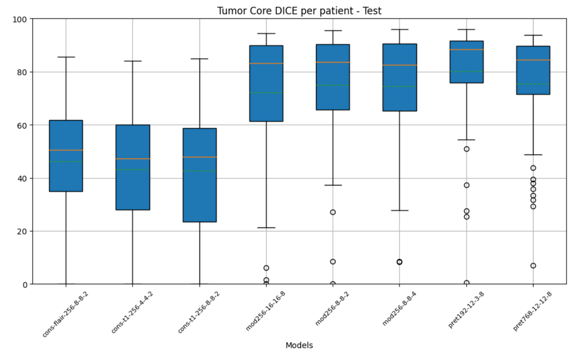


(a) Segmenter models

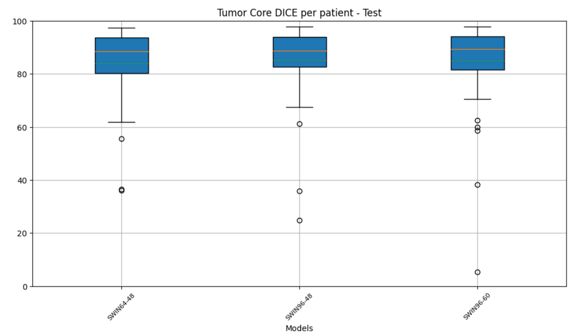


(b) SwinUNETR models

Figure 3: Box plot - Whole Tumor DICE metric on test set

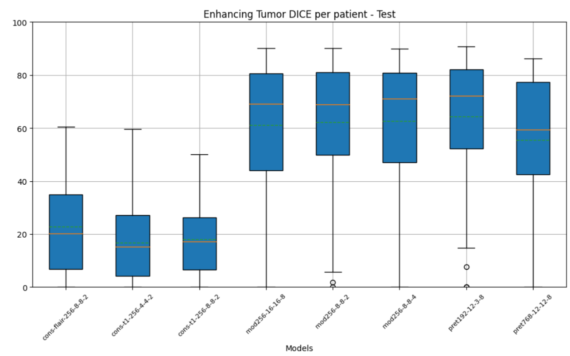


(a) Segmenter models

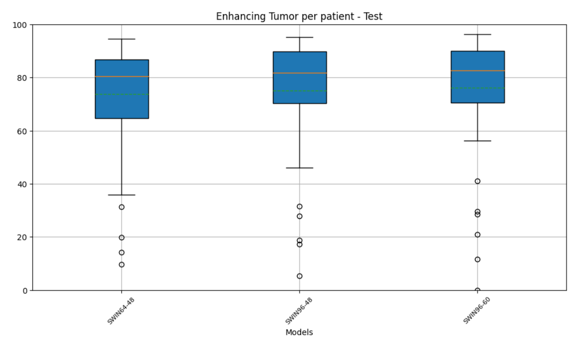


(b) SwinUNETR models

Figure 4: Box plot - Tumor Core DICE metric on test set



(a) Segmenter models



(b) SwinUNETR models

Figure 5: Box plot - Enhancing Tumor DICE metric on test set

6. Conclusions

From these observations, we can distill the findings of this thesis into the following conclusions:

- The simultaneous analysis of different MRI modalities serves as the foundational element for achieving favorable results in all subregions.
- Models employing three-dimensional input generally outperform those using single slices. On the other hand, the use of transfer learning on bi-dimensional models can help to bridge the gap between these approaches.
- In the medical image field, which is characterized by a limited data availability, we observe that, when using the same three-dimensional approach, Transformer models challenge CNNs, outperforming them in the prediction of Whole Tumor and Tumor Core subregions.

References

[1] 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. *arXiv preprint*, 2021.

[2] Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger R Roth, and Daguang Xu. Swin unetr: Swin transformers for semantic segmentation of brain tumors in mri images. In *International MICCAI Brainlesion Workshop*, pages 272–284. Springer, 2021.

[3] Andriy Myronenko and Ali Hatamizadeh. Robust semantic segmentation of brain tumor regions from 3d mris. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5*, pages 82–89. Springer, 2020.

[4] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[6] Feifan Wang, Runzhou Jiang, Liqin Zheng, Chun Meng, and Bharat Biswal. 3d u-net based brain tumor segmentation and survival days prediction. In *International MICCAI Brainlesion Workshop*, pages 131–141. Springer, 2019.