Comparing Conv Nets and Transformers for Volumetric MRI Segmentation

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Medical imaging has undergone significant advancements with the advent of machine learning and deep learning methodologies. Structural Magnetic Resonance Imaging (MRI) stands out as one of the most detailed non-invasive imaging techniques, offering a comprehensive perspective of intricate brain structures. Nonetheless, accurate segmentation of volumetric MRI data remains challenging due to the complex spatial arrangements, large data sizes, and the imperative for high precision in clinical settings.

Deep learning architectures have shown potential in improving MRI segmentation accuracy. Among these, the Residual Unet has garnered attention because of its capability to capture intricate details using skip connections and strided convolutions [4]. Meshnet, in contrast, predominantly leverages dilation to obtain satisfactory results [3]. Recently, the Vision Transformer (ViT) has made significant strides in image classification tasks by employing a global receptive field and reducing computational overhead with patching [2]. We have undertaken preliminary efforts to adapt a ViT for 3D medical segmentation applications.

In our research, we conduct a comparative analysis of the three architectures: Residual Unet, Meshnet, and our preliminary extension of 3D-ViT, with a specific focus on their suitability for brain MRI segmentation [3, 2]. Our evaluation is grounded in three main metrics: segmentation accuracy (measured using Dice scores), computational efficiency (time required to process a patient batch), and GPU resource consumption during volumetric and sub-volumetric training and inference. Through a comprehensive comparison, we seek to highlight the practical implications of each model, providing insights into their potential in both research and real-world clinical scenarios. This investigation serves as a reference for present methodologies and potentially sets the stage for subsequent innovations in medical imaging and neural network architectures.

2 Methodology

We investigate three distinct models: a state-of-the-art Residual Unet, Meshnet, and our preliminary Vision Transformer [4, 3, 2]. We adopted the Residual Unet from MONAI [1] and Meshnet from Catalyst's minimal example [5].

To evaluate the effectiveness of these models, we partition the HCP dataset into training (80%), validation (10%), and test (10%) sets. Our exploration encompasses training and inference on both 128x128x128 subvolumes and full 256x256x256 volumes. Dice scores are reported for both training and validation phases. Additionally, we provide metrics on GPU memory usage and batch processing time for each model during training and inference, considering a batch size of 1. For the subvolume approach, we segment the non-overlapping subvolumes and their respective labels, implementing segmentation on these during subvolume training and inference.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3 Results

3.1 Dice Scores

3.1.1 Best Models

The Dice scores obtained during subvolume training for both the Residual Unet and Meshnet approached perfection. A considerable portion of our efforts was dedicated to refining 3D-ViT; however, our results remain preliminary. Further improvements are necessary before 3D-ViT can be compared favorably with either Meshnet or Residual Unet. Dice scores for these models are presented in Figure 1.

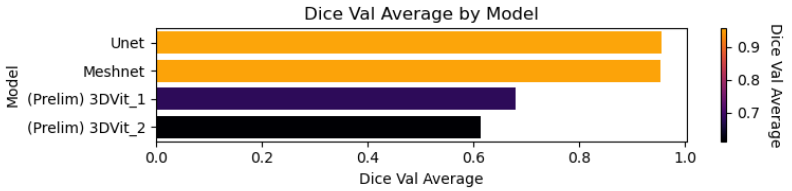

Figure 1: Dice scores of selected models on the validation set.

3.1.2 Dice Score Training Behavior

The training behavior, in terms of Dice score per epoch, is remarkably similar between Meshnet and Residual Unet as visualized in Figure 2. However, we noted that the Residual Unet can exhibit instability under certain learning rate schedules. The most prominent distinction between the two models lies in their memory usage and batch processing times, which we will discuss in subsequent sections.

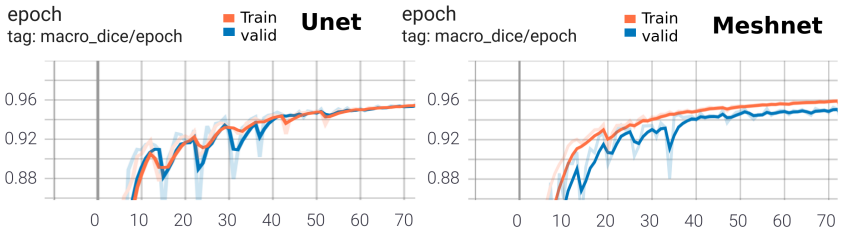

Figure 2: Dice score training behavior for both training and validation sets.

3.2 GPU Consumption and Inference Times

The efficiency with which these models are trained plays a pivotal role in assessing their suitability for practical use. Our metrics on GPU consumption and inference times are illustrated in Figure 3. The Residual Unet outperforms Meshnet both in terms of speed and memory footprint. However, when training for 104-class segmentation on full volumes, both models become exceedingly resource-intensive, to the point that they exceed the capacity of our available GPUs and cause crashes.

For 3-class segmentation, the Residual Unet stands out as the only model capable of handling full-volume training within the confines of a 32GB V100 GPU. Both Meshnet and 3D-ViT demonstrate untenable memory demands during full-volume training. Interestingly, Meshnet boasts superior inference speeds compared to the Residual Unet, despite trailing during training. Similarly, 3D-ViT outpaces Residual Unet in inference times when handling full volumes.

It's worth mentioning that leaner variants of Meshnet might align more closely with the Residual Unet's memory and time metrics. Likewise, shallow ViTs exhibit commendable speed. Given the near-instantaneous growth rate of its receptive field, one could speculate that a more concise ViT might be feasible for such tasks. It should also be noted that the Meshnet model referenced in our study is

sourced from Catalyst [5] owing to its public availability. However, our internal experiments have yielded more efficient Meshnet variants, which we anticipate releasing in forthcoming publications.

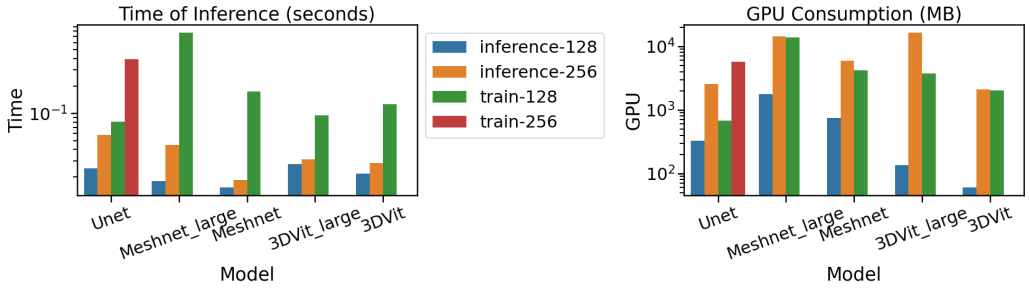

Figure 3: Inference time and GPU consumption during training and inference for both full volumes (256x256x256) and subvolumes (128x128x128) in a 3-class segmentation context.

3.3 Example Segmentation Cross Sections

Inspecting Figure 4, it becomes evident that the Residual Unet and Meshnet exhibit remarkably congruent behavior, a conclusion that aligns with their Dice score evaluations.

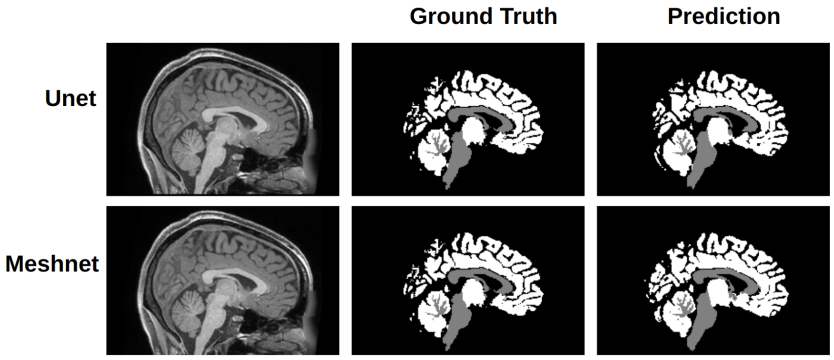

Figure 4: A 2-dimensional slice representation of the 3-dimensional, 3-class segmentation predictions made by the Residual Unet and Meshnet.

4 Discussion

Both Meshnet and Residual Unet are promising architectures that can provide insights into refining the 3D-ViT. The strategies employed by Meshnet and Residual Unet focus on efficiently expanding the receptive field—Meshnet achieves this through dilation while the Residual Unet utilizes strides. On the other hand, while the 3D-ViT boasts a global receptive field, this comes at a cost. However, this expense is mitigated to some extent through patching.

There are potential avenues for further enhancing these models. Residual connections could potentially be integrated into both Meshnet and 3D-ViT to facilitate training. For the 3D-ViT, introducing patching strategies that incorporate strides or dilation might be beneficial. A significant challenge in our application of the 3D-ViT has been the positional encoding; in our implementation, it hasn't been aptly redefined for a three-dimensional context. We are in the process of revising this component of the model and anticipate enhancements in the near term. Furthermore, we are gearing up to release advanced Meshnet architectures that demonstrate greater training efficiency.

References

[1] M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, Andriy Myronenko, Wentao Zhu, Yun Liu, Mingxin Zheng, Yucheng Tang, Isaac Yang, Michael Zephyr, Behrooz Hashemian, Sachidanand Alle, Mohammad Zalbagi Darestani, Charlie Budd, Marc Modat, Tom Vercauteren, Guotai Wang, Yiwen Li, Yipeng Hu, Yunguan Fu, Benjamin Gorman, Hans Johnson, Brad Genereaux, Barbaros S. Erdal, Vikash Gupta, Andres Diaz-Pinto, Andre Dourson, Lena Maier-Hein, Paul F. Jaeger, Michael Baumgartner, Jayashree Kalpathy-Cramer, Mona Flores, Justin Kirby, Lee A. D. Cooper, Holger R. Roth, Daguang Xu, David Bericat, Ralf Floca, S. Kevin Zhou, Haris Shuaib, Keyvan Farahani, Klaus H. Maier-Hein, Stephen Aylward, Prerna Dogra, Sebastien Ourselin, and Andrew Feng. Monai: An open-source framework for deep learning in healthcare, 2022.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Alex Fedorov, Jeremy Johnson, Eswar Damaraju, Alexei Ozerin, Vince Calhoun, and Sergey Plis. End-to-end learning of brain tissue segmentation from imperfect labeling. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3785–3792. IEEE, 2017.

[4] Eric Kerfoot, James Clough, Ilkay Oksuz, Jack Lee, Andrew P King, and Julia A Schnabel. Left-ventricle quantification using residual u-net. In *Statistical Atlases and Computational Models of the Heart. Atrial Segmentation and LV Quantification Challenges: 9th International Workshop, STACOM 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers 9*, pages 371–380. Springer, 2019.

[5] Sergey Kolesnikov. Catalyst - accelerated deep learning rd. `https://github.com/catalyst-team/catalyst`, 2018.