Mamba-Based Tumor Segmentation

Clay Crews[†]
Department of Computer Science
University of South Carolina
Columbia, SC, United States
jccrews@email.sc.edu

Lexington Whalen[†]
Department of Computer Science
University of South Carolina
Columbia, SC, United States
LAWHALEN@email.sc.edu

Abstract—U-net and transformer-based architectures have dominated medical image segmentation, particularly for tumor segmentation from brain MRI scans. However, Structured State Space Sequence Models (SSMs), as highlighted by Gu, Goel, and Ré, present a promising alternative with comparable modeling power to transformers but linear scalability in sequence length, making them well-suited for large medical images.

We propose exploring the application of the Mamba block, a variant of SSMs, for tumor segmentation in brain MRI scans. By leveraging selective state spaces, Mamba blocks can potentially enhance the efficiency and effectiveness of tumor segmentation while maintaining linear scalability.

Our methodology involves implementing and fine-tuning a Mamba-based architecture on a dataset of annotated brain MRI scans. We will evaluate the model's performance using standard segmentation metrics and compare the results with U-net and transformer-based approaches.

Index Terms—Mamba, Segmentation, Selective State Spaces

I. INTRODUCTION

MRI imaging to identify brain tumors has naturally been a target for accurate image segmentation through the use of ML architecture. The use of computer-aided diagnosis of tumors can be crucial to quickly identify tumors and determine a course of action for the patient. Gliomas are the most common primary brain malignancy with varying degrees of aggressiveness in the brain. The protocol for MRI image annotation of these tumors consists of the following labels: the whole tumor extent, the tumor core, the non-enhancing/necrotic tumor region, and regions of low grade gliomas [5]. These annotations of an MRI are used to determine the size and severity of a tumor.

To accurately segment these images, picking out the small features of the image becomes the focus. The use of the U-Net architecture, presented by Ronneberger, Fischer, and Brox [1], has had success in MRI image segmentation. U-Net builds on the approach of fully convolutional networks by improving the limitation on the large amount of data needed to accurately train the model. Very few training images are needed due to the contracting and expansive paths in this architecture. In the contracting path, images are downsampled and pooled along with increasing the number of feature channels. Each downsampled resolution produces a multi-channel feature map. The expansive path up samples and combines with each segmentation feature map from the contracting layer

to localize features in the image. A 1x1 convolution layer is applied to the final image to map the large feature vector to a selected number of classes. All of this essentially creates a feature map where each pixel's relevance is taken into account and evaluated in the end result. An approach using the U-Net architecture would provide the detailed analysis needed for tumor annotation.

An arising issue in the complex recurrent or convolutional neural networks used for image segmentation is the length of the sequences given as input. Maintaining relevance of the current area of an image in relation to the rest of the image is key for more accurate segmentation. An approach with the Transformer architecture has a large emphasis on learning specific features in an image, outlined by Vaswani et al. 2017 [2]. This architecture follows an encoder-decoder pattern, connected through an attention mechanism, a transformer block. Transformer blocks show capability of learning long-distance dependency throughout an input image. In the self-attention mechanism for this block, a single element in a given sequence is compared to all of the elements in the sequence. However, this mechanism becomes computationally very expensive when it comes to long sequences.

Selective State Spaces (SSMs) pose an improvement in computational efficiency over Transformers for this long dependency range. SSMs represent the relevance and context of different parts of the input sequence giving selective attention. This model is widely used in control theory for time variant systems and in fields such as computational neuroscience. Selective attention is very efficient in handling long sequences over time and presents a more localized context of the data. Sections of sequences that deserve attention are represented by the model and will dynamically update these values to reflect the contextual relevance.

Making use of Structured State Spaces, presented by Gu, Goel, and Ré [3], for their superiority in long range dependency tasks, a specific type of SSM, the Mamba architecture, by Gu and Dao [4], proposed a foundational model to operate on arbitrary sequences from a variety of inputs in the domain of sequence modeling. This model achieves the power of Transformers while scaling linearly in sequencing length to be computationally efficient. Additionally, the selection mechanism of Mamba improves on an SSMs ability to focus on or ignore sections in an input sequence by parameterizing the parameters of an SSM to reflect the current context of the

---

[†] Equal contributions

input. The Mamba model and its application to brain tumor MRI imaging segmentation will be evaluated in this project.

II. COMPARED MODELS

Prior to an analysis of how the Pyramidal U-Mamba compares, we shall explain what models we chose for comparison and why.

For this our analysis, we compare four models against our's. The models are U-Net [1],ResNet18, ResNet50, U-Mamba [6], and SegViT [11]. Below we explain our choices.

A. U-Net

The core idea behind U-Net is to complement a traditional contracting network with successive layers that replace pooling operations with upsampling operators. These upsampling layers aim to increase the resolution of the output. Subsequent convolutional layers can then learn to assemble a precise output based on this high-resolution information [1]. A key modification in the U-Net architecture is the inclusion of a large number of feature channels in the upsampling part. This allows the network to effectively propagate contextual information to higher resolution layers. As a result, the expansive path of the network becomes more or less symmetric to the contracting part, leading to a distinctive U-shaped architecture. Unlike traditional convolutional networks, U-Net does not employ any fully connected layers and only utilizes the valid part of each convolution [2]. To predict pixels at the border regions of the image, the missing context is extrapolated by mirroring the input image. This tiling strategy is crucial for applying the network to large images, as it circumvents resolution limitations imposed by GPU memory constraints [1]. Due to its encoder-decoder architecture, skip connections, multi-scale feature extraction, and its efficiency, U-Net has been used in many denoising and diffusion models. For instance, DDPMs (Denoising Diffusion Probabilistic Models) use an architecture similar to U-Net for denoising and sample generation [8] while the popular Stable Diffusion architecture uses a U-Net based architecture for the diffusion process [7]. We have chosen to compare our novel segmentation architecture against the U-Net architecture due to its well-established reputation and widespread adoption in the field of medical image segmentation.

B. ResNet

We have also selected ResNet18 and ResNet50 as additional benchmarks. These architectures, introduced by He et al. in their seminal work "Deep Residual Learning for Image Recognition" [9], have revolutionized the field of deep learning by addressing the problem of vanishing gradients in deep neural networks. The key innovation in ResNets is the introduction of residual connections, which allow the network to learn residual functions with reference to the input layer, thereby facilitating the training of much deeper networks. ResNet18 and ResNet50, with 18 and 50 layers respectively, have been widely adopted in various computer vision tasks, including image classification, object detection, and segmentation. These

models have demonstrated exceptional performance and generalization ability across diverse datasets. By comparing our proposed architecture against ResNet18 and ResNet50, we aim to assess its effectiveness in relation to these well-established and highly influential architectures. This comparison will provide valuable insights into the capabilities of our model and its potential to advance the state-of-the-art in image segmentation tasks.

C. U-Mamba

As our model takes much inspiration from the recently developed U-Mamba design [6], we also choose to incorporate it in our comparison. U-Mamba addresses these limitations by introducing a novel hybrid CNN-SSM block that leverages the strengths of both architectures. The convolutional layers in the block are responsible for local feature extraction, while the State Space Sequence Models (SSMs) (Gu et al., 2022), a new family of deep sequence models, are known for their strong capability in handling long sequences and capturing long-range dependencies. By integrating these two components, U-Mamba achieves a balance between local and global information processing, enabling it to effectively handle long-range dependencies in biomedical image segmentation tasks. Moreover, U-Mamba incorporates a self-configuring mechanism that allows it to automatically adapt to various datasets without manual intervention, enhancing its versatility and usability.

D. SegViT

SegViT [10] explores the capability of plain Vision Transformers (ViTs) [11] for semantic segmentation, leveraging the attention mechanism to generate masks for semantic segmentation. Unlike previous ViT-based segmentation networks that learn pixel-level representations from the output of the ViT, SegViT introduces the Attention-to-Mask (ATM) module. In the ATM module, the similarity maps between a set of learnable class tokens and the spatial feature maps are transferred to the segmentation masks, enabling effective semantic segmentation. SegViT has demonstrated impressive performance on several benchmark datasets, including ADE20K, COCO-Stuff-10K, and PASCAL-Context, achieving state-of-the-art results. Moreover, to address the computational cost associated with the ViT backbone, SegViT proposes query-based down-sampling (QD) and query-based up-sampling (QU) to build a Shrunk structure. By comparing U-Mamba against SegViT, we aim to evaluate the effectiveness of our hybrid CNN-SSM approach in relation to the attention-based segmentation mechanism employed by SegViT. This comparison will provide insights into the strengths and weaknesses of both architectures and their ability to handle complex biomedical image segmentation tasks. Furthermore, it will help us understand the potential of attention mechanisms and hybrid approaches in advancing the state-of-the-art in biomedical image segmentation.

III. Results

We now compare our model against three other common segmentation models, U-Mamba [], U-Net [], and SegViT []. And it was good.

Model	Param Count	Size (MB)	Avg. IOU	Avg. Dice
Pyramidal Mamba (Ours)	10M	109.15	0.8004	0.8867
U-Mamba	9.8M	108.10	0.7908	0.8799
U-Net	7.8M	31.09	0.8071	0.8929
SegViT	Y	Z	A	B
ResNet18	13.9M	55.76	0.7623	0.8648
ResNet50	42.9M	171.85	0.7661	0.8672
Ultra LightM-UNet	0.049M	0.20	0.6725	0.7955

IV. Loss Curves

We also display the loss curves for each model. Each model was trained for 100 epochs under [OTHER CONDITIONS].

Acknowledgment

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv:1505.04597 [cs.CV], 2015.

[2] A. Vaswani et al., "Attention Is All You Need," arXiv:1706.03762 [cs.CL], 2023.

[3] A. Gu, K. Goel, and C. Ré, "Efficiently Modeling Long Sequences with Structured State Spaces," arXiv:2111.00396 [cs.LG], 2022.

[4] A. Gu and T. Dao, "Mamba: Linear-Time Sequence Modeling with Selective State Spaces," arXiv:2312.00752 [cs.LG], 2023.

[5] S. Bakas et al., "Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge," arXiv:1811.02629 [cs.CV], 2019.

[6] J. Ma, F. Li, and B. Wang, "U-Mamba: Enhancing Long-range Dependency for Biomedical Image Segmentation," arXiv:2401.04722 [cs.CV], 2024.

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," arXiv:2112.10752 [cs.CV], 2021.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," CoRR, vol. abs/2006.11239, 2020. [Online]. Available: https://arxiv.org/abs/2006.11239

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," CoRR, vol. abs/1512.03385, 2015. [Online]. Available: http://arxiv.org/abs/1512.03385

[10] B. Zhang et al., "SegViT: Semantic Segmentation with Plain Vision Transformers," arXiv:2210.05844 [cs.CV], 2022.

[11] A. Dosovitskiy et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," arXiv:2010.11929 [cs.CV], 2021.