SegOne: Reinventing Text-Based Image Semantic Segmentation with Only 1D Channel Convolutions

Sanghyun Byun byuns@usc.edu
Kayvan Shah kpshah@usc.edu
Ayushi Gang agang@usc.edu
Christopher Apton apton@usc.edu

Abstract

Many modern computer vision architectures leverage diffusion for its ease of feature extraction. However, these models are often heavy due to their gradual noise generation and removal processes. We propose to lighten the model by generating an open-vocabulary segmentation mask with diffusion using only 1D channel convolutions. The pixel-unshuffle operation is utilized to capture spatial relationships in the channel dimension so that 2D information is distributed to channels. The architecture comprises a 1D diffusion and a CLIP encoder, with text embedding augmented to latent space and/or decoder skip-connection layers. To further enhance performance, output image segments are compared to text prompts with a teacher model (SAM2) to ensure consistency during training. For the sake of simplicity, we focus on MRI image segmentation task.

1 What is the problem?

Problem Statement
Edge devices, such as mobile phones and embedded systems, often lack the powerful GPUs needed to run the heavy diffusion models that are common in current computer vision research. Most models in this field are designed with the assumption that the hardware is capable of performing basic tensor operations, typically requiring significant computational resources. To make these models more accessible on low-power (edge) devices, there is a need to explore decomposition methods that reduce computational load without sacrificing performance.

Research Hypothesis
Analogous to how using 2D convolutions lightens the computational burden of 3D pipelines, we hypothesize that 1D convolutions could similarly reduce the load of 2D pipelines. The goal of this research is to demonstrate that a 1D convolution approach can be effectively applied to complex 2D diffusion models, offering a path toward significant reduction in computational intensity while maintaining model efficacy.

1.1 Challenges

- **Feature Capture Limitations**: 1D convolutions inherently capture fewer spatial features compared to 2D convolutions. A key challenge is to build a pipeline that enables 1D convolutions to capture enough relevant features to make them viable in 2D applications.
- **Architectural Overhaul**: Since 1D convolutions operate along the channel dimension only, the diffusion architecture must be redesigned to accommodate this, requiring new techniques to ensure performance is not compromised.

Preprint. Under review.

1.2 Benefits

- **Edge Device Applications**: This approach would make computer vision technologies more accessible to companies/groups developing edge-device applications, allowing them to use cheaper processors without losing performance.
- **Mobile and Embedded Systems**: By creating a less computationally intensive pipeline, we enable more efficient deployment of diffusion models on mobile applications, older devices, and in concurrent processing environments.

2 How is it currently approached?

This section provides an overview of the key existing methods used in the field to address this problem.

- Gupta et al. [2] for brain tumor segmentation from MRI images.
  - **Method Overview and relevance** This paper uses deep learning models like U-Net, Attention U-Net, and Recurrent Residual U-Net for brain tumor segmentation from MRI images. These architectures focus on pixel-level segmentation through techniques like residual blocks and optimization algorithms like Adam. Our approach, however, employs 1D channel convolutions for segmentation rather than 2D convolutions and incorporates Pixel Unshuffle and Laplacian pyramids to handle spatial attention.
  - **Strengths** The paper utilizes established architectures like U-Net and its variations, known for their robust performance in medical image segmentation. The Recurrent Residual U-Net achieves a high MIoU of 0.8665, making it reliable for clinical applications.
  - **Weaknesses** The paper only handles MRI-based segmentation and does not involve multimodal tasks. Our approach, which uses CLIP to ensure textual consistency with segmented outputs, adds flexibility to tasks beyond medical imaging.
  - **Evaluation** The paper evaluates its models on a public MRI dataset of 3064 images from 233 patients with brain tumors, using metrics like Mean IoU and Dice Coefficient. The Recurrent Residual U-Net model outperformed others with an MIoU of 0.8665, demonstrating its strong segmentation capabilities on MRI data.
- Fan et al. [5], which is designed for efficient and accurate semantic segmentation specifically optimized for mobile devices.
  - **Method Overview and relevance** The method presented in PP-MobileSeg focuses on fast and efficient semantic segmentation optimized for mobile devices. It aims to provide accurate segmentation results with a lightweight model. While PP-MobileSeg is aimed at lightweight models for mobile devices, our method uses a novel approach to handling spatial attention and multimodal integration through CLIP, adding flexibility for tasks involving text and image consistency.
  - **Strengths** The model is specifically optimized for mobile devices, which means it's lightweight and fast, making it practical for real-time applications. Despite being designed for low-resource environments, it maintains high segmentation accuracy on mobile platforms, which is critical in practical deployments.
  - **Weaknesses** Due to its optimization for mobile platforms, PP-MobileSeg may not perform as well on large-scale datasets or tasks that demand high spatial attention. Our approach overcomes this limitation by employing a novel technique that simplifies spatial attention without sacrificing efficiency.
  - **Evaluation** PP-MobileSeg is evaluated using common benchmarks for semantic segmentation, such as Cityscapes and ADE20K datasets, which are widely accepted in the computer vision community. These datasets test the model's ability to perform accurate segmentation in real-world scenarios while being resource-efficient. The results demonstrate that PP-MobileSeg balances accuracy and efficiency, making it suitable for real-time mobile applications. Its performance is convincing for its target task: semantic segmentation on mobile platforms.
- Tang et al. [1] for fusing features from CT and MRI images to enhance segmentation performance.

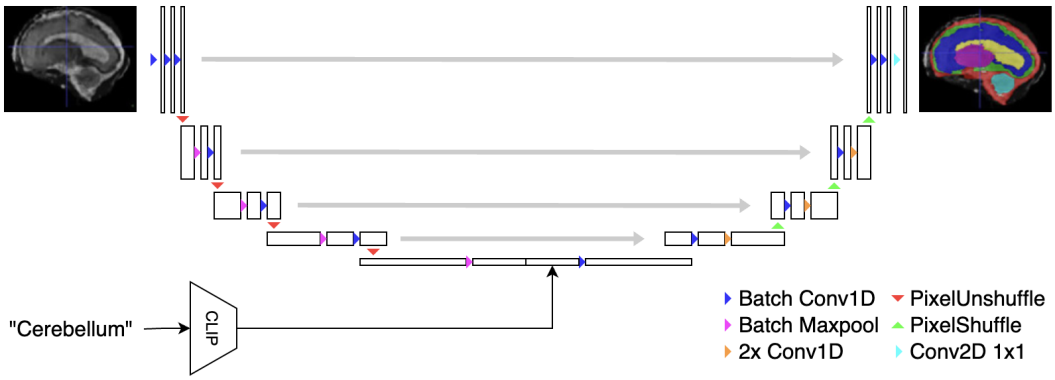

Figure 1: **Model Architecture of SegOne.** The 1D diffusion pipeline follows the architecture of a standard U-Net, with layers altered to fit the 1D modeling as discussed in 3.1. Text is encoded with CLIP (T5) and embedded to latent space and possibly decoder layers for open-vocabulary segmentation.

- **Method Overview and relevance** The MicFormer model leverages a dual-stream architecture incorporating a Cross Transformer to fuse multimodal features—primarily CT and MRI images. The Cross Attention mechanism identifies correlations between modalities, enhancing segmentation by utilizing both strengths. This is especially useful when CT data, for instance, lacks clear boundaries, and MRI data helps clarify them. MicFormer emphasizes cross-modal interaction (image-image), while our approach involves text-image consistency and optimized spatial handling.
- **Strengths** By fusing CT and MRI features, it improves segmentation accuracy through better delineation of difficult-to-identify regions (e.g., tissue edges in CT scans).
- **Weaknesses** The Cross Transformer architecture and deformable attention mechanisms require significant computation, which may not be optimal for real-time or mobile applications. Our model tackles the computational cost by utilizing 1D channel convolutions to simplify spatial attention handling, which can be more efficient while maintaining good performance.
- **Evaluation** MicFormer was tested on the MM-WHS dataset. The evaluation metrics used were the Dice coefficient, MIoU, and HD95. It achieved superior performance compared to models like VT-Unet and Swin-Unet, making the results highly convincing in multimodal medical segmentation.

3 How do you plan to approach it?

The proposed model consists of two parts — a 1D diffusion model with PixelUnshuffle [4] and a CLIP encoder. A diagram of the proposed model is shown in Fig. 1.

3.1 Model Architecture

To take into account channel convolutions, each operation must be altered to fit the model similarly to a U-Net with skip connections. In our approach, we utilize channel-wise 1D convolution in the space of a 2D convolution layer. To simulate the changing channel dimension of 2D convolutions, we propose to perform channel-wise maxpool as well as upscaling 1D convolutions. As pixel-unshuffling reduces the image scale while increasing the channel, we use it in place of the maxpool operation. Finally, instead of upscaling 2D convolutions, we utilize pixel-shuffling to increase resolution. We generate the final segmentation mask through a 2D convolution head with a kernel size of 1, which is effectively a 1D convolution.

To incorporate the open-vocabulary aspect into the model, we propose further enhancing the model by providing the associated word through CLIP encoding. The exact plans to incorporate this information remain undecided.

3.2 Training and Evaluation

In training the model, we will use $l1$ loss as shown by SAM2 [3]. Similarly, for evaluation, the mean absolute error (MAE) will be used for more calculations as well. We train and evaluate the model on multiple MRI datasets, reporting in- and out-of-domain accuracies.

As the proposal aims to achieve results similar to state-of-the-art segmentation models with only 1D convolutions, we seek to not lighten the model but replicate the functions with similar model size and accuracy. To determine the success of the model, we will mainly use the mIOU metric, with the highest goal of reaching roughly 63.

3.3 Project Timeline

Week 5: Dataset Collection and Literature Review

Week 6: Mid-term Report and Model Initialization

Week 7: Model Pre-training and Testing 203 Week 8: Model Testing and Debugging

Week 9: Model Finetuning on the MRI Data Set

Week 10: CLIP-Encoder Experiments and Testing

Week 11: Further CLIP-Encoder Testing

Week 12: mIOU Calculations and Evaluation

Week 13: Model Finalization and Reporting

Week 14: Final Report Work

References

[1] Xinxin Fan, Lin Liu, and Haoran Zhang. Multimodal information interaction for medical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

[2] Ayan Gupta, Mayank Dixit, Vipul Kumar Mishra, Attulya Singh, and Atul Dayal. Brain tumor segmentation from mri images using deep learning techniques. *arXiv preprint arXiv:2305.00257*, 2023.

[3] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[4] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *The IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2016.

[5] Shiyu Tang, Haoran Zhang, Lin Linu, and Xinxin Fan. Pp-mobileseg: Fast and accurate semantic segmentation model for mobile devices. In *Proceedings of the International Conference on Mobile Computing and Applications*. Mobile Computing Society, 2024.