Semantic Segmentation using transformers: An application to MRI.

Evelyn Cueva, Universidad San Francisco de Quito, Ecuador

***Abstract*—Medical image segmentation plays a crucial role in healthcare by enabling precise identification and delineation of anatomical structures and pathological regions, which are essential for diagnosis, treatment planning, and disease monitoring. Traditional segmentation methods are often time-consuming, heavily reliant on expert input, and prone to human error. Recent advancements in deep learning, particularly in Transformer-based architectures, have revolutionized the field, offering robust, automated, and efficient solutions.**

This study focuses on implementing the *SegFormer* model, a Transformer-based lightweight and efficient architecture, for semantic segmentation of gastrointestinal tract magnetic resonance images (MRI). Using the UW-Madison GI Tract Segmentation dataset, we evaluate the model's performance in segmenting complex and imbalanced medical data. The methodology involves preprocessing MRI scans and masks, training SegFormer with a custom loss function combining Dice coefficient and cross-entropy.

Our results demonstrate the effectiveness of SegFormer, achieving an F1-Score of 93.38% on the validation set, highlighting its capability for precise and generalized segmentation. This research underscores the potential of Transformer-based models like SegFormer in advancing medical image segmentation tasks, paving the way for their application in diverse clinical settings.

***Index Terms*—Semantic Segmentation, Magnetic Resonance Imaging, Transformers, Gastrointestinal Tract.**

I. Introduction

MEDICAL image segmentation is a critical task in healthcare, enabling precise identification and delineation of anatomical structures and pathological regions. This process is fundamental for various applications, including disease diagnosis, treatment planning, and monitoring of clinical progress. However, traditional segmentation techniques, such as manual or semi-automated methods, are labor-intensive, prone to human error, and often inconsistent. These limitations have highlighted the need for robust, automated approaches capable of handling the complexity of medical imaging data.

Recent advancements in deep learning have significantly transformed the landscape of medical image segmentation. Convolutional Neural Networks (CNNs), such as U-Net and its variants, have been widely adopted due to their strong performance in segmenting biomedical images. Despite their success, CNN-based models face challenges in capturing global contextual relationships, which are essential for accurate segmentation in complex medical images.

Transformer-based architectures have recently emerged as a promising alternative, addressing the limitations of CNNs. By leveraging self-attention mechanisms, Transformers excel at capturing long-range dependencies and global context. Among these, SegFormer stands out as a lightweight and efficient model designed specifically for semantic segmentation. Unlike traditional Transformer-based models, SegFormer employs a hierarchical encoder for extracting multi-scale features and a simple Multi-Layer Perceptron (MLP) decoder for generating segmentation maps. This design achieves competitive accuracy while maintaining computational efficiency, making it suitable for medical applications.

In this study, we explore the use of SegFormer for semantic segmentation of gastrointestinal tract magnetic resonance imaging (MRI). The primary objective is to evaluate its effectiveness in handling the complexities of medical image data, including class imbalance and high-resolution requirements. By leveraging the UW-Madison GI Tract Segmentation dataset, this work aims to demonstrate the potential of SegFormer in improving segmentation performance and its applicability in clinical settings.

II. State of the Art

A. Semantic Segmentation in Medical Imaging

Semantic segmentation has become an essential tool in medical imaging, particularly in analyzing MRI and CT scans. Early methods for segmentation relied heavily on manual annotations or classical computer vision techniques such as edge detection, thresholding, and region-growing algorithms. However, these approaches lacked robustness, scalability, and generalizability, making them unsuitable for large-scale or complex datasets.

Machine learning marked a turning point in medical image segmentation, with early models utilizing handcrafted features combined with traditional classifiers like support vector machines (SVMs). These methods provided modest improvements but were limited by their inability to capture complex patterns and relationships in medical images.

B. The Rise of Convolutional Neural Networks

The advent of deep learning, particularly Convolutional Neural Networks (CNNs), revolutionized medical image segmentation. Architectures such as U-Net[1] and V-Net[2]

became benchmarks for biomedical segmentation tasks. U-Net, with its encoder-decoder structure and skip connections, enabled precise localization of features, achieving state-of-the-art performance in various applications. Similarly, V-Net extended this approach to volumetric data, addressing 3D medical imaging challenges. Despite their effectiveness, CNN-based models often struggled to capture global context, which is critical for understanding complex anatomical structures.

C. Transformers in Medical Imaging

In recent years, Transformers have emerged as a powerful alternative to CNNs, leveraging self-attention mechanisms to model long-range dependencies and global context. TransUNet[3] pioneered the integration of Transformers with CNNs, employing a hybrid encoder-decoder architecture. Similarly, Swin-Unet[4] utilized a hierarchical Swin Transformer to enhance computational efficiency and scalability. While these models achieved impressive results, their high computational requirements posed challenges for deployment in resource-constrained environments.

D. SegFormer: A Lightweight Transformer for Segmentation

SegFormer[5] represents a significant advancement in Transformer-based architectures for semantic segmentation. Unlike hybrid models, SegFormer relies entirely on Transformers, employing a hierarchical encoder to extract multi-scale features and a simple MLP decoder for generating segmentation maps. This design minimizes computational overhead while achieving competitive performance across various benchmarks. The model's lightweight nature and efficiency make it particularly suitable for medical imaging applications, where resources and time are often limited.

E. Contribution of This Work

Building upon these advancements, this study applies SegFormer to the segmentation of gastrointestinal tract MRI scans. Unlike previous works, which primarily focused on generic datasets or hybrid architectures, this study evaluates SegFormer's performance on highly imbalanced and clinically relevant medical datasets. By demonstrating its capability to generalize across challenging conditions, this work highlights the potential of SegFormer in advancing the field of medical image segmentation.

III. Materials and Methods

A. Dataset

The UW-Madison GI Tract Segmentation dataset was used in this study, consisting of 467 de-identified MRI scans from 107 patients. Each patient underwent 1-5 scans, resulting in a total of approximately 144 axial slices per scan, with an average pixel resolution of $1.5 \times 1.5 \times 3$ mm. The segmentation task involves identifying four classes: background (0), small bowel (1), large bowel (2), and stomach (3).

The dataset was preprocessed by resizing the images to 288×288 pixels, normalizing pixel values, and applying data augmentation techniques to enhance generalization. The dataset was split into training (80%) and validation (20%) sets, ensuring a balanced distribution of classes in each subset.

B. Pipeline Overview

The proposed segmentation pipeline follows a structured approach comprising data preprocessing, model fine-tuning, and evaluation. The process is illustrated in Figures 1 and 2. The following steps are considered:

- Data Preprocessing: MRI images and their corresponding segmentation masks are preprocessed, including resizing, normalization, and augmentation.
- Dataset Splitting: The dataset is divided into training and validation sets, ensuring a balanced distribution.
- Model Fine-Tuning: A pre-trained SegFormer model is fine-tuned using a custom loss function combining the Dice coefficient and cross-entropy.
- Inference and Evaluation: The trained model is used for segmentation, and results are compared with ground truth masks.
- Visualization of Results: Segmentation outputs are visualized to assess model performance.

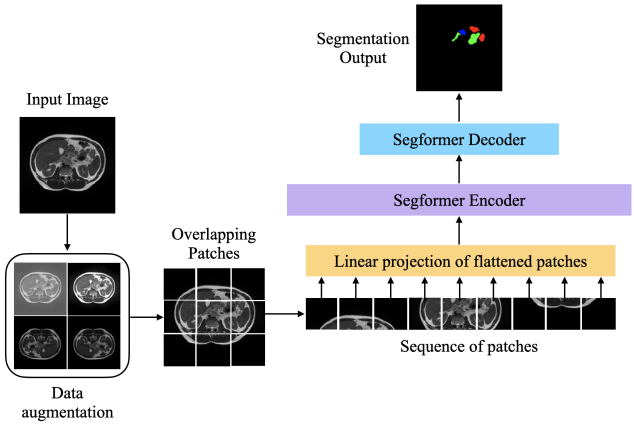

Figure 1. Overview of the segmentation pipeline. The process includes data preprocessing, overlapping patches, linear projection, Segformer encoder and decoder.

C. Data Preprocessing and Augmentation

The dataset consists of MRI scans labeled with regions of interest. Preprocessing steps include:

- Resizing images to a fixed resolution.
- Normalizing pixel values.
- Data augmentation techniques such as horizontal flipping, random cropping, and brightness adjustments.

D. Model Fine-Tuning

To adapt the SegFormer model to the MRI segmentation task, we perform fine-tuning using:

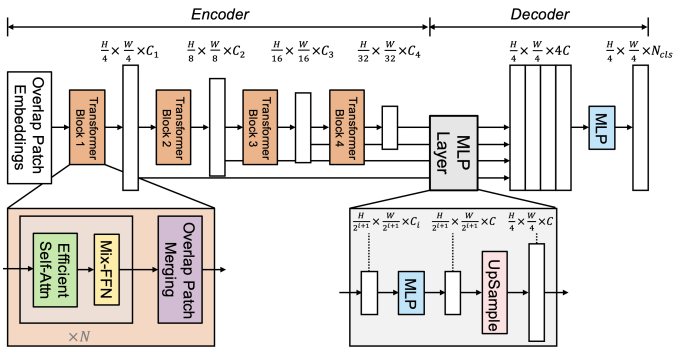

Figure 2. SegFormer architecture overview. The model consists of a hierarchical encoder for feature extraction and a simple MLP decoder for segmentation. FFN indicates feed-forward network. **Source:** [5].

1) Loss Function: A combination of Dice loss and cross-entropy loss to handle class imbalance.
2) Hyperparameters:
   - Initial learning rate: 3×10^{-3}
   - Batch size: 8
   - Optimizer: Adam
   - Training epochs: 100
3) Framework: PyTorch Lightning.
4) Hardware: Experiments were conducted on an NVIDIA DGX-2 system equipped with 16 NVIDIA V100 GPUs with Tensor Cores. These resources are part of the infrastructure provided by Universidad San Francisco de Quito, enabling efficient training and evaluation of the SegFormer model.

E. Evaluation Metrics

Model performance is evaluated using:

- Dice Coefficient: Measures segmentation overlap. The formula is given by:

$$\text{Dice} = \frac{2|A \cap B|}{|A| + |B|}$$

where,

A : Predicted segmentation area
B : Ground truth segmenatition area
$A \cap B$: Intersection of the two areas
$|A|$: area of the set A.

- F1-Score: Evaluates classification accuracy for each class. The formula is given by:

$$F1_{\text{macro}} = \frac{1}{N} \sum_{i=1}^{N} \frac{2 \cdot \text{TP}_i}{2 \cdot \text{TP}_i + \text{FP}_i + \text{FN}_i}$$

where, $i \in \{0, 1, 2, 3\}$.
TP : True possitive
FP : False possitive
FN : False negative

- Validation Loss: Tracks optimization progress.

F. Reproducibility

The full implementation, including dataset preprocessing, model training, and evaluation scripts, is available at Github. This ensures transparency and reproducibility for future research.

IV. Resultados

The model's performance was evaluated across 100 epochs using Dice Coefficient, F1-Score, and loss function values for both training and validation sets. The Table I summarizes these results:

Table I
Performance metrics for the SegFormer model.

Metric	Training	Validation
F1-Score	0.9336	0.9334
Loss	0.2064	0.3185

In Figure 3 we observe that the loss function curve demonstrates a smooth and steady decrease, with validation loss closely tracking training loss, indicating effective generalization. In Figure 4, the F1-score curve shows consistent improvement in both training and validation, stabilizing around epoch 80, suggesting strong model convergence and robustness.

During the initial training epochs, the learning rate gradually increases, ensuring stable optimization and preventing abrupt changes in weight updates. The observed fluctuations around epoch 40 reflect the transition from the warm-up phase to the decay phase of the learning rate schedule, where the model begins fine-tuning its parameters more effectively.

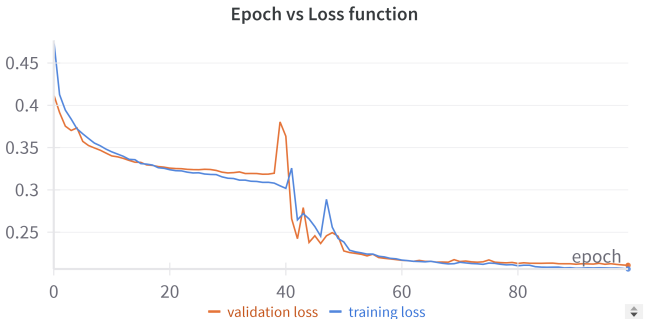

Figure 3. Training and validation loss function curves over 100 epochs.

To provide qualitative insights, Figure 5 showcases a selection of predicted segmentation masks compared with the corresponding ground truth labels. In most of the cases, the model accurately segments the gastrointestinal tract, demonstrating clear boundary delineation with minimal false positives. Slight under-segmentation is observed (row 3), where small regions of the ground truth are missed; however, the overall shape is well preserved. The segmentation is near-perfect, with an almost identical match between prediction and ground truth, highlighting the model's robustness.

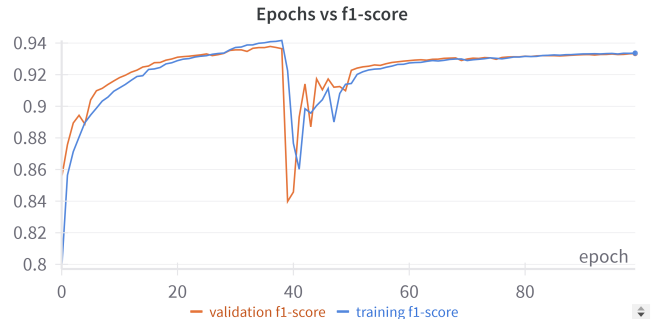

Figure 4. Training and validation F1-score curves over 100 epochs.

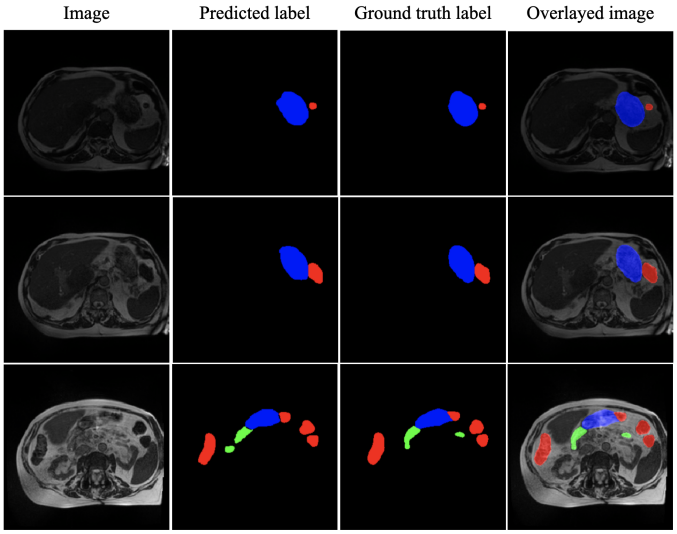

Figure 5. Comparison of MRI images (first column),predicted segmentation (third column), ground truth labels (second column), and overlayed results (fourth column).

These results confirm that SegFormer effectively segments MRI images, achieving high accuracy while maintaining computational efficiency. The analysis of the learning curves further supports the model's stability and convergence. Future work may focus on optimizing hyperparameters to further refine segmentation performance and enhance generalizability across different datasets.

V. Conclusion

This study demonstrates the effectiveness of the SegFormer model in addressing the challenging task of semantic segmentation in medical imaging, specifically on gastrointestinal tract MRI scans. The results highlight several key findings:

- High Accuracy: The SegFormer model achieved a Dice Coefficient and F1-Score of approximately 0.933 for both training and validation sets, showcasing its ability to accurately delineate anatomical structures.
- Efficient Training: The steady decrease in loss function and stabilization of F1-scores around epoch 80 suggest efficient learning and convergence, with no signs of overfitting.
- Robust Segmentation: Qualitative evaluations demonstrate the model's robustness in capturing complex structures, effectively handling both under- and over-segmentation scenarios.
- Computational Efficiency: The lightweight nature of SegFormer makes it a practical choice for clinical applications, balancing accuracy and efficiency.

In conclusion, SegFormer presents a promising solution for medical image segmentation tasks, with potential applications in clinical workflows and diagnostic pipelines. Future work could focus on expanding the dataset, exploring other imaging modalities, and further optimizing the model's architecture to enhance performance across diverse medical imaging tasks.

References

[1] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[2] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision (3DV)*, 2016.

[3] J. Chen *et al.*, "Transunet: Transformers make strong encoders for medical image segmentation," *arXiv preprint arXiv:2102.04306*, 2021.

[4] X. Cao *et al.*, "Swin-unet: Unet-like pure transformer for medical image segmentation," *arXiv preprint arXiv:2105.05537*, 2021.

[5] E. Xie *et al.*, "Segformer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.