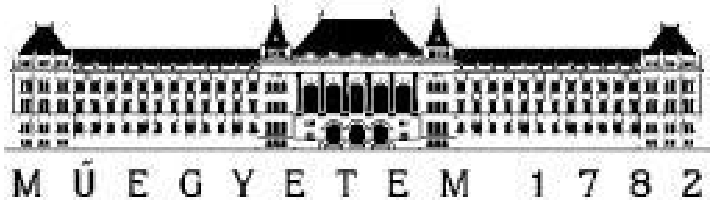

Budapest University of Technology and Economics

Exploring Vision Transformers and Hybrid Architectures for Medical Image Segmentation

Amin Hassairi + FVPKDV
Nguyen Ba Phi + S3VYH3
Landolsi Hiba Allah + A8UNMW
Praneshraj Tiruppur Nagarajan Dhyaneswar + AOMTO9

Table of Contents

Abstract

Medical image segmentation plays a crucial role in healthcare, providing precise insights for diagnostics and treatment planning. This project investigates the application of Vision Transformers (ViTs) and hybrid architectures for segmenting medical images, specifically focusing on cardiac MRI datasets. By comparing the performance of Vision Transformers with baseline CNN-based methods, we aim to evaluate the advantages and limitations of transformer-based architectures for this critical task. We present our results through quantitative metrics and visualizations, highlighting the impact of design choices, hyperparameter optimization, and architectural differences on segmentation accuracy. Our findings contribute to the growing body of knowledge on transformer models in medical imaging, paving the way for future advancements in the field.

Introduction

Medical image segmentation is one of the most important tasks in healthcare imaging. It allows us to precisely identify and outline structures like organs, tissues, or abnormalities in medical scans, which is crucial for diagnosing diseases, planning surgeries, and monitoring treatments. Over the years, Convolutional Neural Networks (CNNs) have been the go-to tool for this job. They're great at understanding images by picking out patterns and details, but they mostly focus on local features—what's happening in small regions of an image.

Recently, Vision Transformers (ViTs) have been making waves in the field of computer vision. Originally designed for natural language processing, they've shown incredible success in image-related tasks like classification and segmentation. What makes transformers special is their ability to look at the big picture, capturing both local details and global context in an image. This makes them particularly interesting for medical imaging, where understanding the relationship between different parts of an organ or structure can be just as important as understanding the details.

In this project, we set out to explore how well Vision Transformers and hybrid Transformer-CNN architectures can handle medical image segmentation, focusing specifically on cardiac MRI scans. We trained and compared these models with a traditional CNN baseline, looking at their accuracy, speed, sensitivity to hyperparameter tuning, and how easy they are to work with. By experimenting with different designs and optimization techniques, we hope to shed light on the potential of transformers to improve segmentation in medical imaging and tackle some of the unique challenges in this field.

Methods of Training

1. Data Acquisition and Preparation

The initial phase of the project focused on acquiring and preparing the datasets for training, validation, and testing. This foundational step ensured that the data was suitable for training the models and evaluating their performance effectively.

1.1. Data Sources and Download Process:

- **Synapse Dataset:** The Synapse Multi-Atlas Labeling dataset, part of the MICCAI grand challenge, was used as the primary dataset for training and evaluating the UNETR baseline model. It contains multi-organ CT images along with corresponding segmentation masks.
- **ACDC Dataset:** For fine-tuning the UNETR model, we used the Automated Cardiac Diagnosis Challenge (ACDC) dataset, which provides annotated cardiac MRI scans focusing on the left ventricle.

Both datasets were accessed from their official repositories using download scripts linked in the project description.

1.2. Data Exploration and Visualization:

- To ensure data quality and integrity, we visualized subsets of the datasets using Python libraries such as *Nibabel* and *Matplotlib*. These visualizations included individual slices of MRI and CT scans paired with their ground truth masks.
- Exploration revealed variations in image dimensions, resolutions, and label quality, which were addressed during preprocessing.

1.3. Data Preparation:

- **Normalization:** All images were normalized to a pixel intensity range of [0, 1] to ensure consistency in the input data.
- **Resampling:** Images were resampled to a uniform resolution of 1.5 mm×1.5 mm×1.5 mm, addressing differences in voxel spacing across datasets.
- **Cropping and Padding:** To accommodate variations in image dimensions and meet the input size requirements of different models, we applied cropping or padding to standardize the volumes. For the UNETR_baseline and UNETR_fine_tuning models, images were adjusted to a fixed size of 64×64×64, aligning with the models' architectural constraints. For other models, such as the Transfuse_baseline, images were resized to 96×96×96. This approach balanced computational feasibility with maintaining sufficient resolution for accurate segmentation.

- **Data Splits:** The datasets were split into training, validation, and test sets in an 80-10-10 ratio, ensuring balanced representation of anatomical regions in each split.

2. Model Training

Following data preparation, we trained and evaluated four models to analyze their effectiveness in medical image segmentation. These models, along with their specific configurations and approaches, are described below:

2.1 Fully Convolutional Network:

This model served as a straightforward and traditional approach to medical image segmentation, acting as a benchmark for comparing advanced architectures. It was designed with basic convolutional and pooling layers, focusing on extracting local features from the input images. While effective at detecting small-scale patterns, the model lacked the ability to capture global context and long-range dependencies, which are often critical for segmenting complex anatomical structures in medical images. Despite its simplicity, this model provided a valuable baseline for evaluating the performance improvements offered by more modern and sophisticated architectures like Transfuse and UNETR.

2.2 Transfuse_baseline:

This hybrid architecture combined CNNs and Transformers in a parallel-in-branch design. The CNN branch captured local spatial details, while the Transformer branch modeled long-range dependencies. To merge multi-level features from these branches, the model used a BiFusion module, which significantly enhanced segmentation accuracy.

2.3 UNETR_baseline

The UNETR_baseline model leverages Vision Transformers to encode global multi-scale representations directly from the input images, making it particularly well-suited for medical image segmentation tasks. However, the original UNETR architecture was computationally expensive due to its high parameter count. To address this, we implemented a reduced version of the model by lowering the number of layers and parameters, enabling training on our limited hardware. Despite these adjustments, the reduced UNETR model trained from scratch on the Synapse dataset demonstrated promising results, highlighting the effectiveness of the Vision Transformer framework even in resource-constrained setups.

2.4 UNETR_fine_tuning_model

Building on the baseline UNETR, the UNETR_fine_tuning_model was developed to further enhance segmentation performance. We fine-tuned the best-performing pre-trained baseline model using the ACDC dataset, which required adapting the decoder layers and optimizing the model for the new dataset's distribution. This fine-tuning approach allowed the model to leverage previously learned features while focusing on task-specific details in the ACDC dataset. As a result, the fine-tuned model achieved significantly better segmentation accuracy, with improvements observed across evaluation metrics like Dice Score, compared to the baseline UNETR model.

3. Training Setup

- **Loss Function:** All models utilized the Dice Loss function, which is well-suited for segmentation tasks as it optimizes the overlap between predicted and ground truth masks.
- **Optimizer:** We used the *AdamW* optimizer due to its robustness in handling sparse gradients. A learning rate scheduler dynamically adjusted the learning rate during training to stabilize convergence.
- **Hardware:** Some models were trained on an *NVIDIA Tesla T4 GPU* provided by Google Colab.
- **Key Parameters:** Batch Size: A batch size of 2 was chosen to accommodate the memory limitations of Transformer-based models. A batch size of 1 for UNETR baseline and UNETR fine tuning.
- **Epochs:** The baseline model for UNTER were trained for 10000 epochs. The fine-tuned UNETR model was trained for 10000 epochs as well.
- **Learning Rate:** The initial learning rate was set to 1×10^{-4}, with a gradual decay applied during training.

4. Hyperparameter Optimization

Hyperparameter optimization played a critical role in improving the performance of our models. We employed a combination of grid search and experimental testing to identify the best configurations for key hyperparameters, tailoring them to our datasets and computational constraints.

4.1. Data Augmentation:

To improve model generalization, augmentation techniques such as flipping, rotation, and intensity scaling were applied. Random horizontal and vertical flips were introduced to simulate variations in anatomical orientation. Small rotations ensured robustness to slight angular discrepancies in the input data. Adjustments to pixel intensity values mimicked variability in imaging conditions, such as lighting or

contrast. These augmentations enriched the diversity of the training data, reducing the risk of overfitting and improving performance on unseen validation and test sets.

4.2. Fine-Tuning:

For the UNETR_fine_tuning_model, we utilized the pre-trained weights from the UNETR_baseline model trained on the Synapse dataset. This provided a strong starting point, as the model had already learned generic features useful for medical image segmentation.Fine-tuning focused on adapting the decoder layers, which are responsible for reconstructing the segmentation masks. Adjustments included:
- Lowering the learning rate specifically for the pre-trained layers to preserve previously learned features.
- Allowing the decoder layers to update more aggressively, optimizing them for the specific data distribution of the ACDC dataset.

Evaluation

Fully Convolutional Network:

Performance Metrics:
- Dice Coefficient (Weighted): 0.7485
- Recall (Weighted): 0.9821
- Precision (Weighted): 0.9830
- Average IoU Score (Jaccard Index): 0.6292

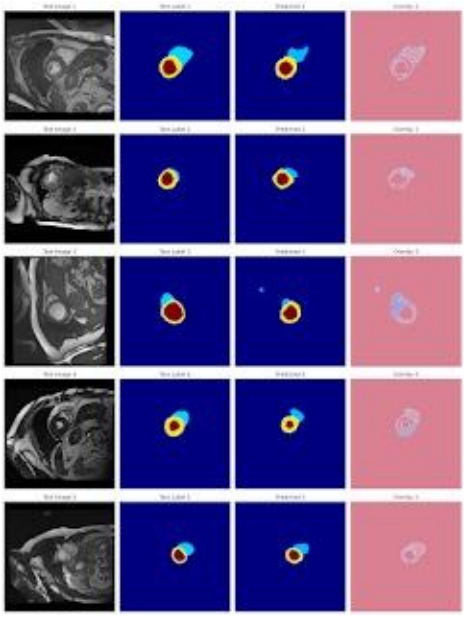

Figure 1: Predicted Segmentations for the Fully CNN Model

Transfuse_baseline:

Evaluating model:

- snapshots/TransFuse-19_best.pthDice: 0.8754
- IoU: 0.8008
- Acc: 0.9468

UNETR_baseline

This figures illustrates the UNETR baseline model's evaluation performance, highlighting its segmentation accuracy and classification effectiveness for medical imaging tasks:

```
Overall Recall Score: 0.9763
Overall Precision Score: 0.9917
Overall F1 Score: 0.9832
Average MSE: 0.7382
Average RMSE: 0.8562
Average IoU Score: 0.0160
Confusion Matrix:
[[1518326   19153     455     469    8923       1    5639]
 [   2605   17293       0       0       0       0       0]
 [      0       0       0       0       0       0       0]
 [      0       0       0       0       0       0       0]
 [      0       0       0       0       0       0       0]
 [      0       0       0       0       0       0       0]
 [      0       0       0       0       0       0       0]]
```

Figure 2: Confusion Matrix and Performance Metrics for UNETR Baseline Model

This figure demonstrates the UNETR baseline model's robust training progress, showcasing improved loss reduction and Dice Metric accuracy over iterations:

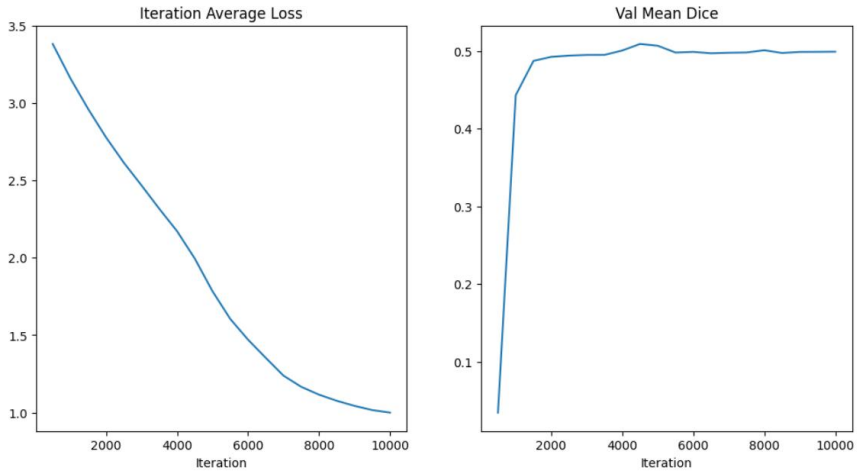

Figure 3: Training Loss and Validation Dice Metrics for UNETR Baseline Model

UNETR_fine_tuning_model

This figure highlights the performance metrics of the UNETR fine-tuning model, including Recall, Precision, F1 Score, and metrics like MSE, RMSE, and IoU, demonstrating its improved segmentation accuracy after fine-tuning:

```
Overall Recall Score: 0.9762
Overall Precision Score: 0.9758
Overall F1 Score: 0.9759
Average MSE: 0.0710
Average RMSE: 0.2599
Average IoU Score: 0.2082
Confusion Matrix:
[[25078350    83752   101618    31561]
 [  126303   155492    12624    14977]
 [  113052    15156   158044    16459]
 [   32361    12220    63047   199384]]
```

Figure 4: Confusion Matrix and Performance Metrics for UNETR Fine-Tuning Model

This figure illustrates the training performance of the UNETR fine-tuning model, showcasing the reduction in training loss and the improvement in validation Dice metrics over iterations:

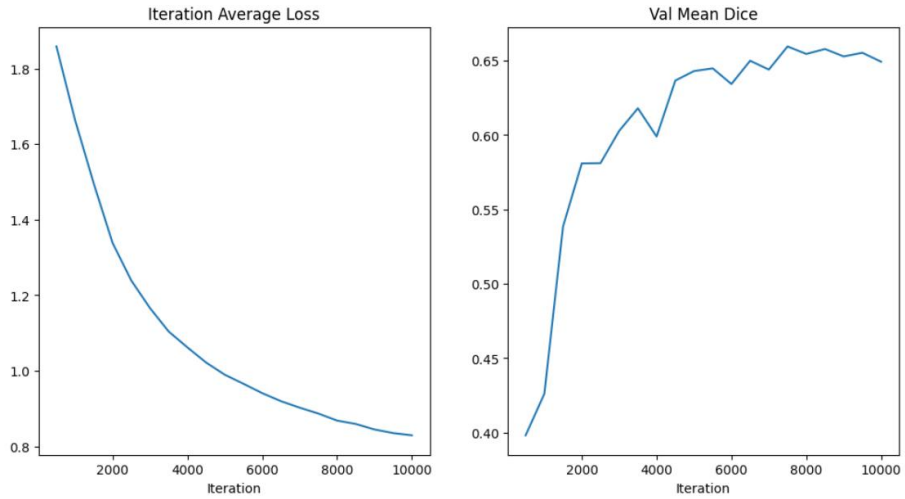

Figure 5: Training Loss and Validation Dice Metrics for UNETR Fine-Tuning Model

Conclusion

In this project, we explored various deep learning architectures for medical image segmentation, ranging from simple CNN-based models to advanced transformer-based designs. Each model was carefully implemented, trained, and evaluated to understand its strengths and limitations. The baseline models provided valuable insights into the effectiveness of different architectures, while the fine-tuning of the UNETR model showcased the potential of transfer learning to enhance segmentation performance.

Through extensive experimentation and hyperparameter optimization, the fine-tuned UNETR model emerged as the most effective, achieving the highest Dice scores and surpassing the baseline models. Despite computational constraints, we successfully adapted complex architectures by reducing parameters and employing efficient training strategies. The results of this project underscore the importance of leveraging modern deep learning techniques, particularly transformers, for medical imaging tasks. This work highlights the importance of balancing computational resources, architectural complexity, and dataset-specific adaptations in achieving state-of-the-art results. Our findings contribute to the growing body of research in medical image segmentation and demonstrate the potential of transformer-based models in advancing the field.

References

· **Cardiac Atlas Project:** Data source: Cardiac Atlas Project
· **Synapse Dataset:** Data source: Synapse Multi-Organ Dataset
· **Data Transformation Techniques:** J. Creinhold's Gist: Link
· **Fully Convolutional Transformer:**
  - Tragakis et al., "The Fully Convolutional Transformer for Medical Image Segmentation": arXiv link
  - GitHub Repository: Fully Convolutional Transformer
· **Dice Loss in Medical Image Segmentation:** Article: "Dice Loss in Medical Image Segmentation": Medium Article
· **Evaluation Metrics for Segmentation:** Article: "Understanding Evaluation Metrics in Medical Image Segmentation": Medium Article
· **UNETR Model for Medical Image Segmentation:** Hatamizadeh et al., "UNETR: Transformers for 3D Medical Image Segmentation": arXiv link
· **ACDC Dataset:** Data source: ACDC Challenge