MED_Segmentation: A Medical Image Segmentation Pipeline for 3D Volumetric Data

Atif Abedeen aabedeen@umass.edu

Abstract

This report describes the development of **MED_Segmentation**, a pipeline for segmenting 3D medical images with applications in spleen segmentation from MRI data. The pipeline incorporates state-of-the-art deep learning models (UNET, VNET, UNETR) and emphasizes reproducibility through tools like MLFlow and DVC. Regularization techniques such as weight decay and dropout are utilized to improve model generalization and prevent overfitting. Additionally, it integrates uncertainty quantification techniques to improve reliability in clinical applications. This report provides details on methods, experiments, and results achieved using this pipeline. GitHub repository: link.

I. Introduction

Accurate segmentation of medical images is essential for many clinical applications, including organ quantification, treatment planning, and disease monitoring. **MED_Segmentation** is a comprehensive pipeline developed for the task of 3D spleen segmentation from MRI data. This pipeline leverages advanced deep learning models and robust experimental tracking tools to ensure reproducibility and high performance.

II. Dataset

The dataset used in this project is the Spleen Dataset from the Medical Segmentation Decathlon [1]. It consists of:

- **Training Data:** 28 3D MRI scans with manual segmentations of the spleen.
- **Validation Data:** 8 3D MRI scans for evaluating model performance during training.
- **Test Data:** 5 3D MRI scans for evaluating model performance.

Each 3D scan has been preprocessed to standardize voxel spacing and intensity values. The dataset is challenging due to variations in spleen size, shape, and position across patients.

III. Pipeline Description

A. Data Ingestion

The dataset is downloaded, extracted, and organized into directories for training and testing. Hidden files are removed, and 3D volumes are validated for integrity. DVC is employed to version control the data, ensuring seamless collaboration and reproducibility across the pipeline.

B. Data Preprocessing

The preprocessing pipeline for the dataset includes the following key steps:

- **Load and Normalize:** Load 3D volumetric images and scale intensity values to the range [0, 1].
- **Crop Foreground:** Remove irrelevant background regions based on intensity values, focusing the input on the organ of interest.
- **Orientation Alignment:** Align all volumes to a consistent orientation using RAS (Right-Anterior-Superior) coordinate system.

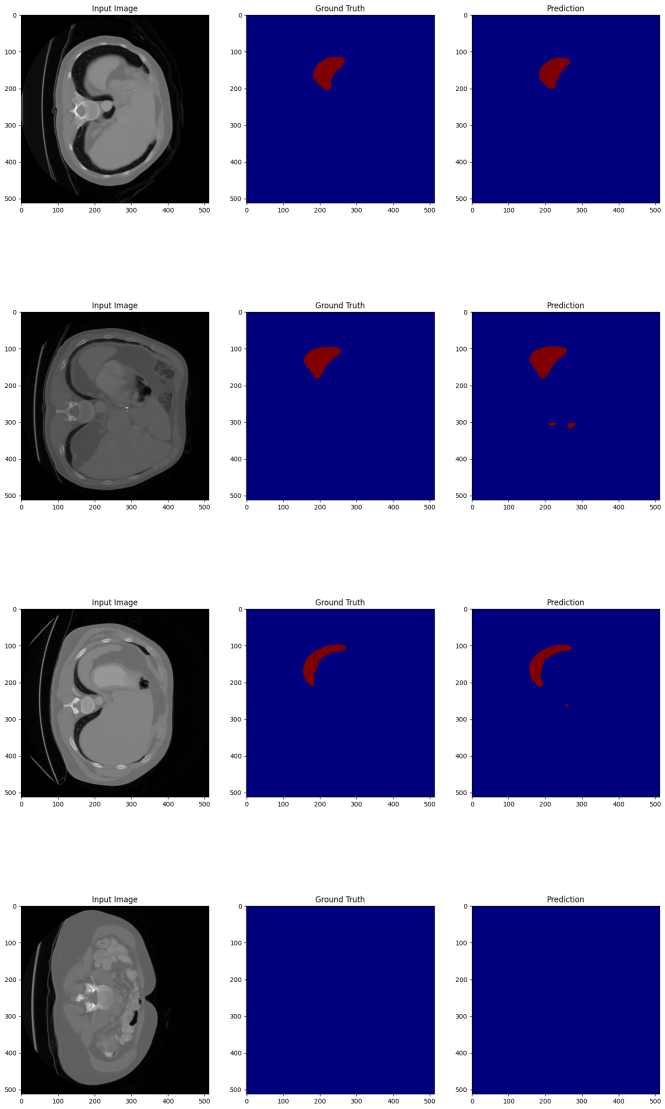

Fig. 1: Inference on test set using UNET. Each slice is from different volumes.

- **Resampling:** Standardize voxel spacing to (1.5mm, 1.5mm, 2.0mm) using bilinear interpolation for images and nearest-neighbor for labels.
- **Random Cropping:** Extract random patches of fixed dimensions from the images based on positive and negative label sampling.
- **Intensity Scaling:** Adjust intensity values within a specified range using min-max scaling to improve feature uniformity.
- **Data Augmentation:** Apply transformations including random flipping, rotation, and cropping to enhance generalization and mitigate overfitting.

This preprocessing ensures that the input data is normalized, aligned, and augmented for robust training and evaluation.

C. Model Training

Three deep learning models (UNET, VNET, and UNETR) are trained using the Dice Loss function. The training pipeline includes:

- **Optimizer:** Adam optimizer with weight decay to prevent overfitting.
- **Regularization:** Dropout layers are used during training to improve generalization by randomly deactivating neurons, ensuring better robustness against overfitting.
- **Learning Rate Scheduler:** Reduces the learning rate periodically.
- **Checkpointing:** Saves the best model based on validation performance.
- **Early Stopping:** Prevents overfitting by halting training when performance plateaus.

MLFlow is integrated into the training process to track experiments, including hyperparameters, training metrics, and model checkpoints. This ensures that experiments are reproducible and allows comparisons between different configurations.

D. Evaluation and Metrics

Model performance is evaluated using the following metrics:

- **Dice Score:** Measures the overlap between the predicted and ground truth masks. It is particularly sensitive to segmentation accuracy and is a commonly used metric in medical image segmentation.
- **Hausdorff Distance:** Evaluates the boundary alignment between the predicted and ground truth masks by calculating the maximum distance between points on their boundaries. The 95th percentile is often used for robustness against outliers.
- **Intersection over Union (IoU):** Measures the similarity between the predicted and ground truth masks by calculating the ratio of their intersection to their union.
- **Precision:** Represents the proportion of correctly predicted positive pixels (true positives) among all pixels predicted as positive.
- **Recall:** Measures the proportion of correctly predicted positive pixels (true positives) among all actual positive pixels in the ground truth.
- **Specificity:** Represents the proportion of correctly predicted negative pixels (true negatives) among all actual negative pixels in the ground truth, indicating the model's ability to avoid false positives.

E. Uncertainty Quantification

Uncertainty quantification (UQ) in this pipeline is implemented using **Monte Carlo (MC) Dropout**, a technique that estimates predictive uncertainty by enabling dropout layers during inference. This approach provides insights into areas of low confidence in the segmentation predictions.

Steps in the UQ Process:

1) **MC Dropout Activation:** During inference, all dropout layers are explicitly set to `train` mode to introduce stochasticity in predictions.
2) **Multiple Forward Passes:** For each input image, 10 stochastic forward passes are performed, generating multiple predictions.
3) **Mean and Variance Calculation:**
   - The **mean** of these predictions is computed to produce the final probabilistic segmentation map, highlighting areas where the model is most confident.
   - The **variance** across predictions quantifies the uncertainty, identifying regions with variability and lower prediction confidence.
4) **Post-processing:** The mean and variance maps are spatially restored to their original dimensions and saved for visualization using MONAI's `Invertd`.

Visualization Example: Figure 2 showcases an example of uncertainty quantification:

- **Left Panel:** Original MRI slice, providing the anatomical context for segmentation.
- **Middle Panel:** Mean prediction map from MC Dropout, with high-intensity regions corresponding to confident segmentations. The prediction focuses on the identified regions of interest with minimal ambiguity in the center.

- **Right Panel:** Uncertainty map (variance), where brighter regions indicate areas with higher predictive uncertainty, particularly along the boundaries of the segmented regions ad other incorrectly segmented regions. These regions reflect areas where the model struggles to confidently distinguish between foreground and background.

This information is critical for clinicians to focus on less reliable areas during diagnosis.

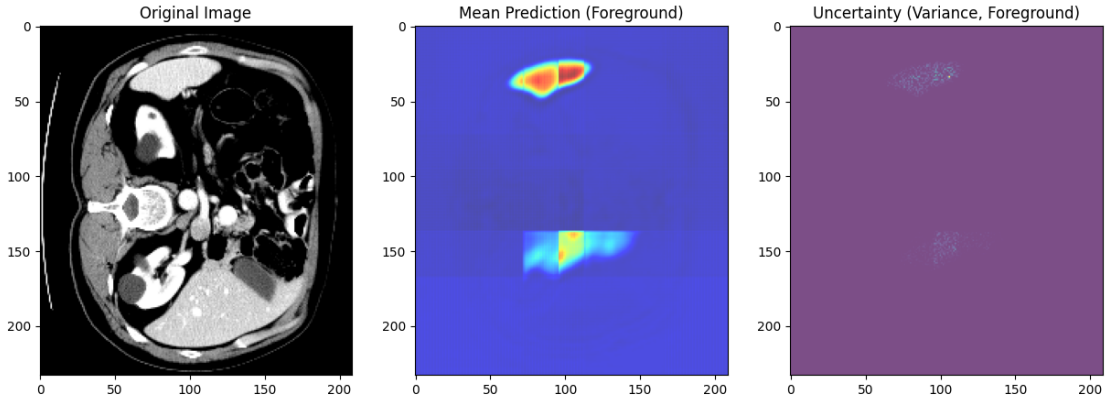

Fig. 2: Visualization of Uncertainty Quantification. Left: Original Image. Middle: Mean Prediction (foreground channel). Right: Uncertainty Map (variance in the foreground channel).

F. User Interface

A Streamlit-based interface allows users to:

- Upload 3D MRI files for inference.
- Select a model (UNET, VNET, or UNETR).
- Visualize segmentation results of each slice overlayed on the input.

IV. Results

Table I summarizes the performance of the models on the Spleen Dataset.

TABLE I: Model Performance Metrics (Average values from test set inferences)

Model	Dice Score	Hausdorff Distance	IoU	Precision	Recall	Specificity
UNET	0.77	134.97	0.99	0.996	0.996	0.996
VNET	0.215	199.61	0.920	0.958	0.958	0.958
UNETR	0.944	4.673	0.999	0.999	0.999	0.999

From Table I, it is evident that the UNETR model outperforms both UNET and VNET across all evaluation metrics. The Dice score of UNETR is the highest at 0.944, with the lowest Hausdorff distance of 4.673, indicating superior segmentation quality and spatial accuracy. On the other hand, VNET lags significantly with a Dice score of only 0.215 and a high Hausdorff distance of 199.61, suggesting poor segmentation quality and spatial misalignments.

The precision, recall, and specificity metrics also affirm the dominance of UNETR, which achieved near-perfect scores (0.999). Although UNET performs reasonably well, its metrics, including a Dice score of 0.77 and Hausdorff distance of 134.97, fall short when compared to UNETR. VNET, with a Dice score of 0.215, shows notable underperformance, further validated by its low precision, recall, and specificity values of 0.958 each.

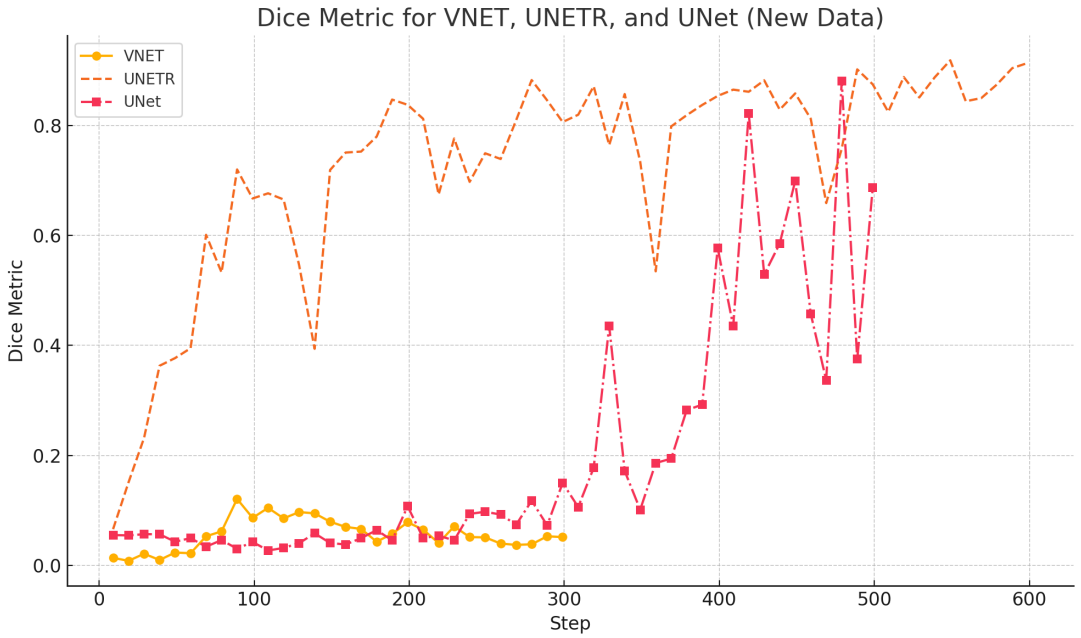

Fig. 3: Dice Metric Calculated on the validation set. Comparison for VNET, UNETR, and UNET Models.

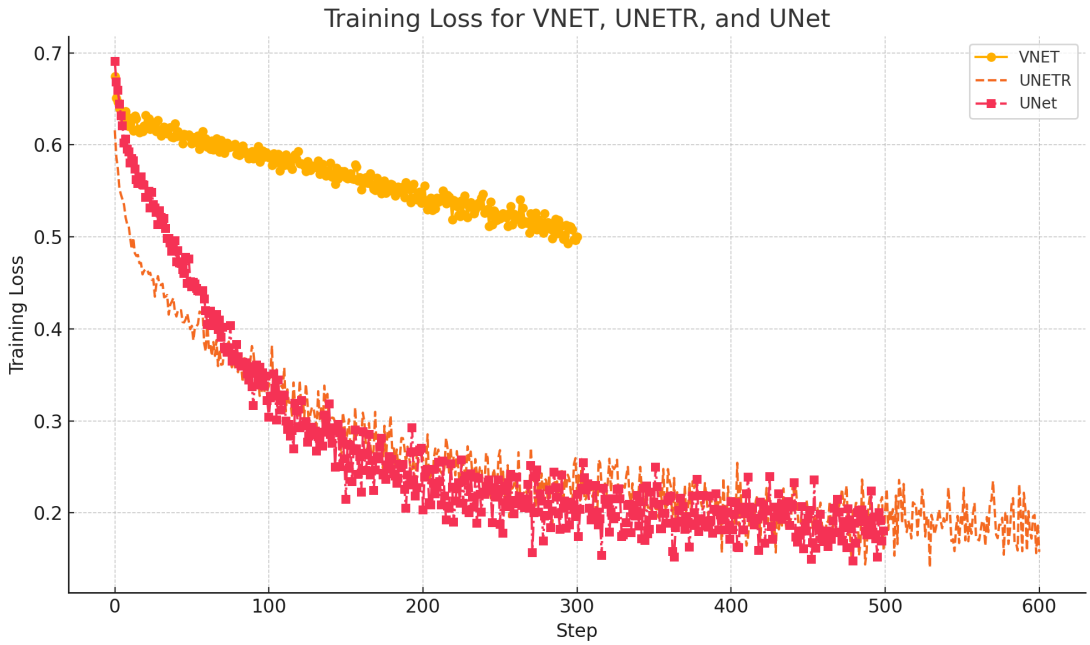

Fig. 4: Training Loss for VNET, UNETR, and UNET Models.

Analyzing Figures 3 and 4, UNETR consistently demonstrates better convergence behavior in training loss (Figure 4) and significantly higher Dice metric values (Figure 3). UNET and VNET exhibit higher fluctuations in both loss and Dice metrics, reflecting lower robustness and stability.

Overall, UNETR emerges as the best-performing model for the Spleen Dataset, offering exceptional segmentation accuracy, robustness, and generalization capabilities, while VNET falls short in all evaluation aspects.

V. Conclusion

MED_Segmentation is a robust and modular pipeline for spleen segmentation, achieving high accuracy and reliability. Regularization techniques, such as weight decay and dropout, were critical in improving model generalization. MLFlow and DVC were instrumental in ensuring reproducibility and experiment tracking. Future work will explore additional uncertainty quantification methods and model improvements.

Acknowledgments

This work was supported by the Hospital for Special Surgery and leveraged datasets from the Medical Segmentation Decathlon.

References

[1] Medical Segmentation Decathlon. http://medicaldecathlon.com