SOLVING THE GRAND-CHALLENGE HAN-SEG-CHALLENGE THROUGH FINE-TUNING THE MONAI FOUNDATION MODEL

Sabrina Caspary (3695797)
Foundation Models, Winter Term 2023/24
Institute for Artificial Intelligence
University of Stuttgart

ABSTRACT

Medical image segmentation plays an important role in various aspects of the medical field. In this paper we want to fine-tune a foundation model from MONAI and evaluate the performance of the fine-tuned model on a medical image segmentation task for the Han-Seg Challenge in order to investigate if we are able to solve the challenge by using a fine-tuned foundation model. The goal of the challenge is to segment 30 Organs-at-risk within the head and neck region from CT images. For that we use a Foundation Model from MONAI and apply several fine-tunings onto it to increase its performance. We then compare the results of the fine-tuned and the not fine-tuned foundation model. We show that the fine-tuned models are able to learn the task of medical image segmentation with the given fine-tuning data. However, the performance of the our fine-tuned models are disappointing when compared to the not fine-tuned foundation model with regards to the challenge task. We further propose some ideas how to increase the performance of the fine-tuned models.

1 INTRODUCTION AND MOTIVATION

In the ever-evolving field of medical research and healthcare, the significance of biomedical image segmentation cannot be overstated. This critical process plays an important role in various aspects of the medical field, primarily focusing on the identification of Regions of Interest (ROIs). The implications extend far beyond visual analysis, including substantial improvements in diagnosis and treatment methodologies. The motivation behind investigating this area lies in the pursuit of reducing human errors, a huge concern in healthcare.

Recently, Foundation Models gained a lot in popularity. These models refer to large-scale pre-trained models that serve as the basis or foundation for various downstream tasks. These models are typically trained on vast amounts of data, often encompassing diverse sources from for example the internet, books, articles, and other material.

The performance of the foundation models can be increased through fine-tuning. By opting for fine-tuning, practitioners unlock the ability to customize models to specific domains, ensuring a more tailored and contextually relevant application. This approach facilitates customization and also contributes to performance increases, aligning the segmentation process more closely with the unique requirements of diverse medical scenarios.

1.1 STRUCTURE OF THIS PAPER

In this paper, we utilize the capabilities of a foundation model provided by MONAI to conduct a comprehensive evaluation of its performance when fine-tuned and to participate at the Grand Challenge. The MONAI Foundation Model serves as the cornerstone of our investigation, providing a robust and versatile framework for deep learning in healthcare imaging. In the following subsection, we describe the details of the MONAI Foundation Model, highlighting its architecture and features that make it a compelling choice for our study.

Section 2 of our paper describes the problem and the methodology employed in our study.

In Sections 3 and 4, we detail our experimental setup and present the corresponding results. We offer a comprehensive overview of the performance metrics and benchmarks achieved by the fine-tuned MONAI Foundation Models.

The following Section 5 discusses some challenges that we encountered during the project.

In Section 6 of our paper, we explore the potential applications and broader implications arising from the work presented in the preceding sections. This critical section serves as a bridge between our empirical findings and the real-world impact of our research.

Lastly, Section 7 encapsulates our conclusions and summarizes the key takeaways from our study.

1.2 The MONAI Foundation Model

Introducing MONAI (Medical Open Network for AI) (`https://github.com/Project-MONAI/MONAI`), a PyTorch-based, open-source framework deep learning framework in healthcare imaging. MONAI is equipped with a suite of features tailored specifically for the demands of medical data.

At the core of MONAI lies its flexible pre-processing capabilities, crafted to handle multi-dimensional medical imaging data. This adaptability ensures easy integration with diverse datasets, accommodating the complexity inherent in medical images.

MONAI includes comprehensive implementations for networks, providing a robust foundation for constructing and fine-tuning models tailored to specific medical imaging tasks. The framework offers a large array of tools, empowering researchers and practitioners to explore novel architectures that cater to the intricacies of healthcare data.

Navigating the loss functions is a crucial aspect of any deep learning framework, and MONAI includes a suite of loss functions tailored for medical imaging applications.

Moreover, MONAI shines with its large set of evaluation metrics. These metrics provide a quantitative measure of model performance, ensuring a thorough assessment of the algorithms' effectiveness in diverse medical imaging scenarios.

One of MONAI's standout features is its repository of pre-trained foundation models specifically designed for medical imaging tasks. This collection accelerates the development process by providing a starting point for practitioners, reducing the need for extensive manual labeling and training data. Leveraging these pre-trained models, users can speed up their research and development of applications in healthcare imaging, fostering innovation in diagnostics, treatment planning, and beyond.

Furthermore, we want to note that we also tried to use the Segment Anything Model (SAM) by Meta (`https://segment-anything.com/`). However, this was not possible since the challenge required use to be to able to handle 3D data, which SAM was not capable of.

2 Problem Description and Method

We want to participate at the *Grand Challenge HaN-Seg Challenge* (`https://han-seg2023.grand-challenge.org/`). This section states the problem and describes our approach of using a fine-tuned foundation model for solving it.

2.1 The Challenge

The objective of the *Grand Challenge HaN-Seg Challenge* (Head and Neck Segmentation Challenge) is to automatically delineate 30 Organs-at-risk (OARs) within the HaN region from CT images in the given test set. This set comprises 14 CT and MR images of the same patients. The challenge leverages the information available in a training set, which includes 42 CT and MR images of the same patients, accompanied by reference 3D OAR binary segmentation masks for CT

Number of epochs	30
Loss	Binary Cross Entropy
Optimizer	Adam
Learning rates	0.0001 - 0.0005

Table 1: Overview of the training hyperparameters

images. The test is maintained as private, and its release is withheld from potential participants to prevent fine-tuning of algorithms. The dataset can be found at Podobnik et al. (2023).

2.2 METHODOLOGY

To solve the presented challenge we utilize the *ct pleen* Foundation Model from MONAI. The model is based on a 3D U-Net architecture, a widely used framework for medical image segmentation. The U-Net consists of an encoder and a decoder, connected by a bottleneck layer. The architecture parameters are as follows:

- Spatial Dimensions: 3 (for 3D)
- Input Channels: 1 (grayscale images)
- Output Channels: 2 (background and spleen)
- Channels: (16, 32, 64, 128, 256) in encoder and symmetrically in decoder
- Strides: (2, 2, 2, 2) for down-sampling
- Number of Residual Units: 2
- Normalization: Batch Normalization

The weights of the neural network are provided by MONAI.

Then, we compare three different configurations of the Foundation Model:

- The pre-trained model without any fine-tuning
- We add one additional layer to the neural network of the pre-trained model and fine-tune it with the training set (set 1) of the HaN-Seg challenges
- We fine-tune the pre-trained model by adding two additional neural network layers training with the training set (set 1) of the HaN-Seg challenges

The foundation model has one input and two output channels. These numbers of channels have to adapted to fit our data. We need two input channels for both the CT and MR data and furthermore we need 30 output channels instead of two, matching the 30 Organs-at-risk within the HaN-Seg data. For that we use convolutional layers to combine the two input channels and to adjust the number of final output channels. We do this by modifying the model by freezing all layers except the last one or two. This means that we kept the architecture of the original model intact and only allowed the last one or two layers to be trainable, while keeping all other layers frozen.

For training we use the hyperparameters that are depicted in table 1.

3 EXPERIMENTAL SETUP

For evaluating the performance of our three different models we use two metrics: Binary Cross-Entropy Loss and the Dice Coefficient. This section describes the experimental process and the metrics used for evaluation.

3.1 METRICS

The choice of loss functions plays an important role in training models effectively. Two widely used loss functions, Cross-Entropy Loss and Dice Coefficient, are particularly noteworthy for their effectiveness in addressing different aspects of the segmentation task.

(Binary) Cross-Entropy Loss is a commonly employed loss function in image segmentation tasks. It evaluates the difference between the predicted probability distribution and the ground truth distribution. In the context of segmentation, where the goal is to classify each pixel into different classes, Cross-Entropy Loss measures the dissimilarity between the predicted probability of each class for a pixel and the actual class label. This loss function is effective in penalizing models for making confident yet incorrect predictions, making it suitable for tasks where pixel-wise accuracy is crucial.

The formula for Cross-Entropy Loss involves calculating the negative logarithm of the predicted probability assigned to the correct class. It encourages the model to assign high probabilities to the correct class and low probabilities to incorrect ones. The minimization of this loss function during training results in a model that accurately assigns pixel-wise labels.

$$\text{CE} = -\sum_{\text{classes } c} [y_c \cdot \log(p_c) + (1 - y_c) \cdot \log(1 - p_c)]$$

In the binary case, i.e., when the number of classes equals two, the BCE can be calculated as:

$$\text{BCE} = -y \cdot \log(p) + (1 - y) \cdot \log(1 - p)$$

The **Dice Coefficient**, often referred to as the F1 score or Sørensen-Dice coefficient, assesses the overlap between the predicted segmentation and the ground truth. It is particularly useful when dealing with imbalanced datasets, where certain classes or regions of interest might be underrepresented. The Dice Coefficient is calculated as twice the intersection of the predicted and ground truth regions divided by the sum of their areas.
The formula for the Dice Coefficient emphasizes both precision and recall, making it a suitable metric for segmentation tasks where capturing the true positives while minimizing false positives and false negatives is crucial. Its values range from 0 to 1, with 1 indicating a perfect overlap between the predicted and ground truth segmentations.

$$\text{DSC} = \frac{2|X \cap Y|}{|X| + |Y|}$$

In summary, while Cross-Entropy Loss is effective for pixel-wise classification tasks, Dice Coefficient provides a valuable metric for evaluating the spatial overlap and segmentation accuracy, especially in scenarios where imbalanced datasets or nuanced segmentation boundaries are prevalent.

3.2 SETUP

Our primary testing function evaluates the segmentation models on the test dataset (set 2). It sets the models to evaluation mode, iterates through batches in the test loader, applies the models to input images, computes the Binary Cross Entropy loss for the entire batch and each channel, thresholds model outputs to obtain binary masks, calculates the Dice coefficient for each channel, and prints information about the overall test loss, average Binary Cross Entropy loss per channel, and average Dice coefficient per channel. Additionally, the testing function visualizes slices with labels for the samples in the batch.

4 RESULTS

We evaluate the performance of all three model configurations. The results can be seen in figures 4, 4 and 4. We can see from them that the two layer fine-tuned model is able to learn way more than the model on which we only fine-tuned the last layer. However, the not fine-tuned model performs the best on the challenge task. This could indicate that the foundation model from MONAI is too inflexible to be able to adapt to the new task.

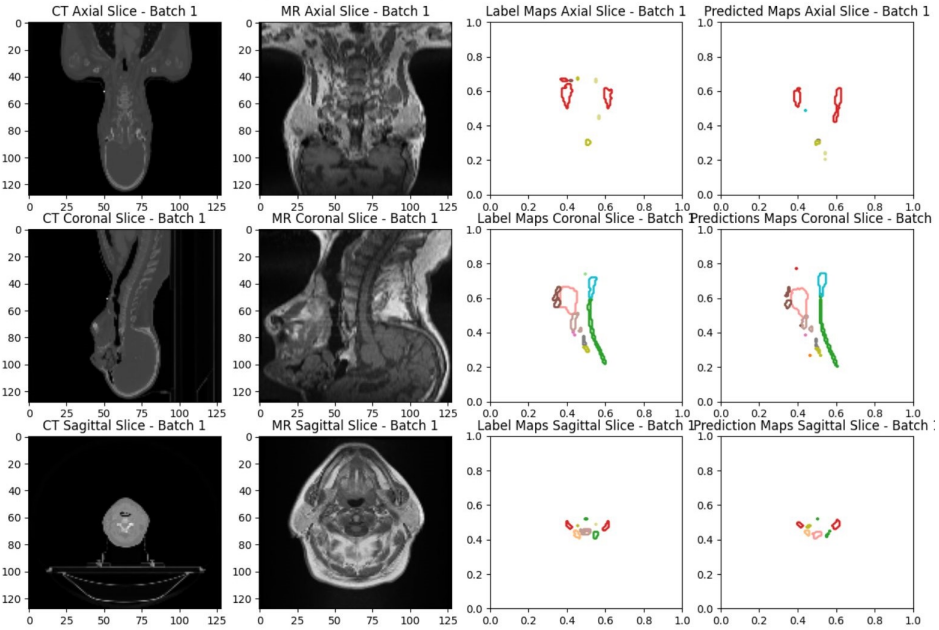

Figure 1: The results of the not fine-tuned model

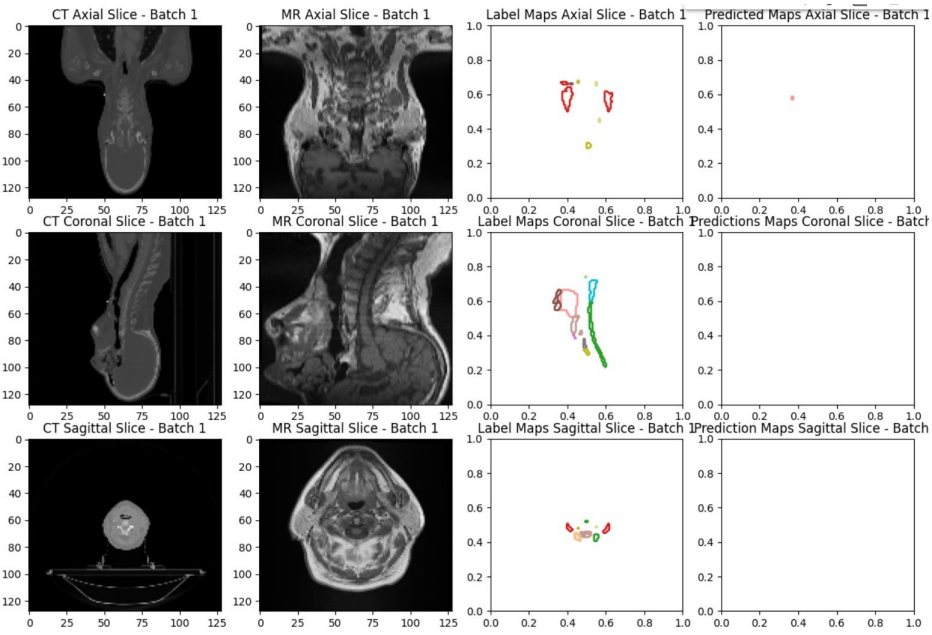

Figure 2: The results of the one layer fine-tuned model

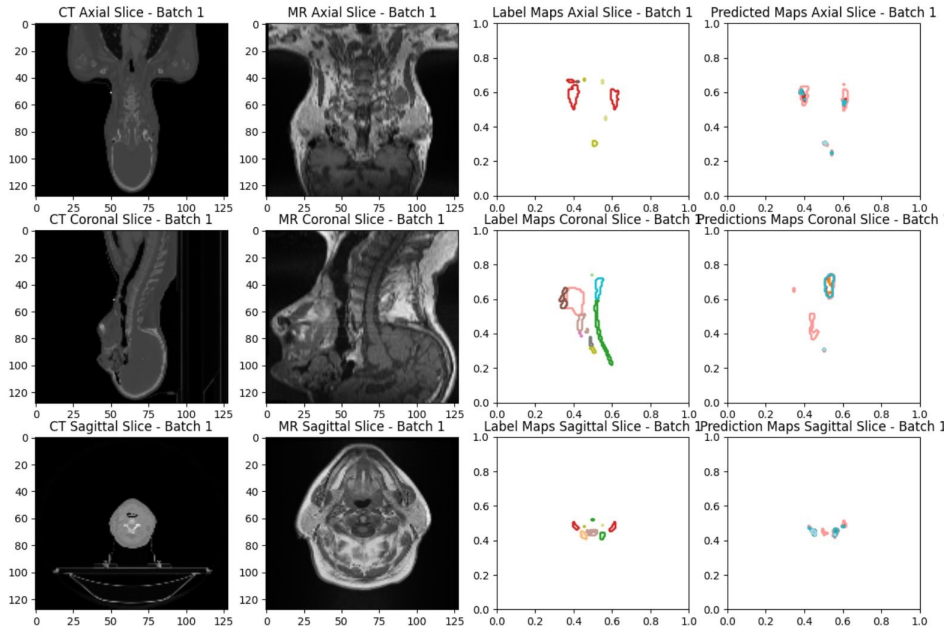

Figure 3: The results of the two layer fine-tuned model

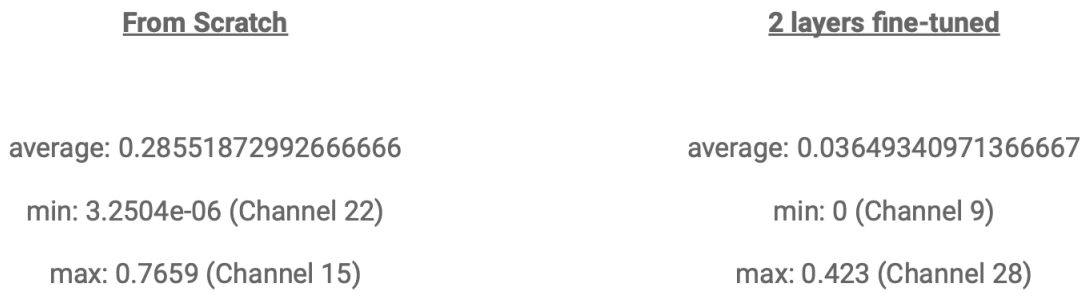

Figure 4: Results of the Model from Scratch vs Pretrained

5 Discussion

During the course of our project, we encountered several challenges that influenced our approach and outcomes. In this section, we discuss these challenges encountered during the fine-tuning process of our segmentation model and propose some ideas how to possibly address them.

Reshaping Data Instances

The size of our data instances posed a significant hurdle, necessitating reshaping to enable effective fine-tuning of our model. This step was crucial for accommodating the data within the model architecture and optimizing its performance.

Limited Training Data

Another challenge arose from the limited number of training instances. With only 42 data instances available, concerns were raised regarding the adequacy of the sample size for robust model training.

Computational Resource Requirements

Fine-tuning the models proved to be computationally intensive, requiring substantial computational resources. Increasing the computational resources could benefit the models' performance.

Adapting to Multilabel Segmentation

Adapting the pre-trained model for multilabel segmentation presented additional challenges, primarily due to significant differences from binary tasks. Considering the complexity involved, transitioning to multiclass segmentation could potentially be simpler approach for model adaptation and training, with the expectation of yielding more promising results.

Addressing Performance Limitations

The performance of the model in certain channels was suboptimal, possibly stemming from downsampling and small segmentation masks. To address this, strategies such as leveraging more powerful computational resources for full-size training, employing data augmentation techniques, and conducting hyperparameter tuning could be considered to enhance overall performance.

6 Applications and Implications

In the field of medical diagnosis, the fine-tuned model might emerge as a powerful ally, capable of providing enhanced accuracy and precision in identifying anomalies and patterns within medical images. This, in turn, has far-reaching implications for treatment planning, as the model's ability to discern intricate details can contribute to the development of more personalized and effective treatment strategies. The fine-tuned model might then become an indispensable tool in the hands of healthcare professionals, facilitating a more refined and targeted approach to patient care.

The usage of the fine-tuned model is not without its ethical considerations, particularly in the context of AI integration into medical imaging. As healthcare becomes increasingly reliant on AI technologies, addressing ethical considerations becomes very important. Issues such as data privacy, transparency in decision-making processes, and the potential impact on patient-doctor relationships must be carefully navigated. A thoughtful exploration of the ethical dimensions associated with the fine-tuned model ensures that its deployment aligns with the principles of beneficence, non-maleficence, autonomy, and justice in the medical field.

7 Conclusion

Fine-tuning foundation models for medical image segmentation is is a powerful technique in using the power of deep learning for healthcare applications. This approach enables the adaptation of pre-trained neural network architectures to the nuanced characteristics of medical imaging data. By leveraging knowledge acquired from diverse datasets during initial training, fine-tuning addresses the challenge of limited medical data, facilitating faster convergence and efficient model development. In the scope of this paper we participated in the Grand Challenge Han-Seg Challenges. Its objective is to automatically segment 30 Organs-at-risk within the Head and Neck region from CT images. To solve the challenge task we used the MONAI Foundation Model and fine-tuned it in various ways with the given data from the challenge. Unfortunately, the performance of the fine-tuned models on the challenge was disappointing compared to the pre-trained model without any fine-tunings. We conclude that the foundation model from MONAI is not flexible enough to adapt to the new task.

The code of this project can be found at `https://github.com/SabCas/FoundationModel/blob/main/mtctvalidationtest.py`.

7.1 Future Work

Possibilities for future work involve addressing the need for more data in our project. Expanding the dataset can contribute significantly to the robustness and generalization capabilities of the model.

Collecting additional diverse and representative data instances can help in training a more comprehensive model that is capable of handling a broader range of variations and scenarios. Moreover, the inclusion of more data can potentially alleviate challenges related to overfitting and improve the model's performance on unseen instances.

Furthermore, there is a growing demand for flexible foundation models in the medical domain. Existing foundation models, such as those found in the MONAI Zoo, offer valuable starting points. However, they may not always be flexible enough to adapt to the diverse and evolving needs of medical applications. In the future, addressing this challenge could present an interesting opportunity. Developing more flexible foundation models tailored specifically for medical tasks could lead to advancements in various areas, including medical imaging analysis, patient diagnosis, and treatment planning. Such models would empower researchers and practitioners to tackle complex medical problems with greater efficiency and accuracy.

REFERENCES

G. Podobnik, P. Strojan, P. Peterlin, B. Ibragimov, and T. Vrtovec. Han-seg: The head and neck organ-at-risk ct & mr segmentation dataset. *Medical Physics*, 2023. doi: 10.1002/mp.16197.