Model Ensemble for Medical Image Segmentation

Bárdos-Deák Botond - HS2APO
Bodai Adrián Tibor - OU1G79

December 8, 2024

Chatgpt was used to create plots, convert images and organize the code, and create descriptions.

1 Introduction

The aim of our project was to segment different parts of the heart using cardiac MRI images. To do this, we used the output of several models to estimate the overall performance, a method called model ensemble.

2 Description of the dataset

Our Dataset is the ACAD Dataset[1] .The data set was originally designed to identify different heart diseases, but is perfect for visual segmentation of parts of the heart. The training set contains 100 patients and the test set contains 50 patients. The images are annotated with left and right lateral ventricles and myocardium.

3 Data preprocessing

the data come in *.nni.gz* file types, what contains 3d images created from layers of 2d images. For easier use we convert the slices into .png files. We also split the files by saving all 3 mask types in separate files.

4 Models what we used

Several models were set up with different parameters. Different model architectures were tried out, the parameters, the loss function, the optimizer, the learning rate were changed. There were models that performed better and worse. To make the ensemble model we followed two main directions, one was to make a model for each of the three objects (left ventricle, right ventricle and myocardium) that would recognize them. Then to make a model that looks for these organs simultaneously. Then, using a combination of these, try to make the most accurate model possible.

4.1 Used model architectures

1. **U-net:** Unet is a fully convolution neural network for image semantic segmentation. Consist of encoder and decoder parts connected with skip connections. Encoder extract features of different spatial resolution (skip connections) which are used by decoder to define accurate segmentation mask. Use concatenation for fusing decoder blocks with skip connections. [2]

[1] See `https://www.creatis.insa-lyon.fr/Challenge/acdc/databasesClassification.html`
[2] See `https://arxiv.org/pdf/1505.04597`

2. **DeepLabV3:** DeepLabV3 is a segmentation model developed by Google. It's designed to perform pixel-level classification tasks efficiently, using atrous (dilated) convolutions to capture multi-scale contextual information and an Atrous Spatial Pyramid Pooling (ASPP) module to improve performance across objects of varying sizes. Segmentation"[3]

3. **MAnet:** Multi-scale Attention Net. The MA-Net can capture rich contextual dependencies based on the attention mechanism, using two blocks:

   - Position-wise Attention Block (PAB), which captures the spatial dependencies between pixels in a global view
   - Multi-scale Fusion Attention Block (MFAB), which captures the channel dependencies between any feature map by multi-scale semantic feature fusion

   [4]

4. **Linknet:** Linknet is a fully convolution neural network for image semantic segmentation. Consist of encoder and decoder parts connected with skip connections. Encoder extract features of different spatial resolution (skip connections) which are used by decoder to define accurate segmentation mask. Use sum for fusing decoder blocks with skip connections. [5]

4.2 Model training parameters

We tried different loss functions: the DiceLoss, JacardLoss, CrossEntropy Loss.

- **Dice Loss** measures the overlap between the predicted area and the ground truth.
- **Jaccard Loss** measures the same as Dice loss, only it uses different operations between the intersection and the union of the predicted area and the ground truth.
- **Cross entropy loss** measures the difference between the true and predicted probability distributions.

For optimizer we used Adam optimizer, and a cyclic learning rate scheduler, with minimum learning rate was 0.00006 and the maximum learning rate was 0.001.
For trank the trainings we used *wandb.ai*.
For the best 4 model we use you can see the trainng loss on Figure 3, and validaton loss on Figure 4

4.3 Final models

For the final model, 7 models were used, one each of the 'DeepLabV3', 'LinkNet', 'Unet', 'MAnet' models predicting all three organ parts simultaneously, and one model predicting only one organ part for each of the three outputs were 'Unet' models.

5 Results of the models

There were some models that completed well and some that completed badly On Figure 1 you can see the predictions prom our final models, you can see that different model can give different inputs on the same picture.
On Figure 2 shows a model that completed very badly. The first row shows the predictions, and the second row shows the original masks.

[3]See `https://arxiv.org/pdf/1706.05587`
[4]See `https://arxiv.org/pdf/2209.14145`
[5]See `https://arxiv.org/pdf/1707.03718`

5.1 Evaluation methods

- **F1 Score:** The F1 score is the harmonic mean of precision and recall. It is useful for evaluating segmentation tasks where class imbalance can occur. The F1 score is defined as:

  scss Kód másolása

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

  Where:

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

- **Pixel Accuracy:** Measures the percentage of correctly classified pixels (both foreground and background). It is calculated as:

$$\text{Pixel Accuracy} = \frac{\text{Number of Correctly Classified Pixels}}{\text{Total Number of Pixels}}$$

- **Intersection over Union (IoU):** IoU measures the overlap between the predicted and ground-truth masks for each class. It is defined as:

$$\text{IoU} = \frac{|\text{Prediction} \cap \text{Ground Truth}|}{|\text{Prediction} \cup \text{Ground Truth}|}$$

  The IoU can be computed separately for each class (left ventricle, right ventricle, myocardium) to provide a more detailed evaluation.

6 Ensemble Model

6.1 Construction of the Ensemble Model

To improve the accuracy and robustness of our segmentation results, we constructed an ensemble model by combining the outputs of four different architectures: **DeepLabV3**, **LinkNet**, **UNet**, and **MAnet**. Each of these models produced segmentation masks for the left ventricle, right ventricle, and myocardium. The ensemble model was created by aggregating the outputs of these models using a threshold-based approach.

For each model, we applied specific thresholds to the predicted outputs to convert them into binary masks (foreground and background). The thresholds were determined based on the performance of each model during validation:

- **DeepLabV3:** Threshold of 5.31
- **LinkNet:** Threshold of 3.33
- **UNet:** Threshold of 6.87
- **MAnet:** Threshold of 5.36

6.2 Ensemble Strategy

We evaluated two primary strategies for combining the thresholded outputs:

1. **Averaging Method:** We computed the average of the binary masks produced by each model and took the mean value at each pixel location. This method helps smooth out discrepancies between model predictions and provides a more consistent output.

2. **Voting Method:** For each pixel, we performed majority voting based on the binary outputs of the models. If more than 50% of the models predicted a pixel as foreground, it was classified as foreground in the ensemble prediction. This method is robust to individual model errors.

7 Results

The ensemble model and the individual models were evaluated using the **F1 Score**, **Pixel Accuracy**, and **Intersection over Union (IoU)**. The average scores reflect the overall segmentation performance across all test images.

The results indicate that all models performed similarly, with slight variations. The **average F1 Score** was approximately **0.32**, while the **average Pixel Accuracy** was around **0.98** across all models. The high accuracy score demonstrates that most pixels were classified correctly, but the lower F1 Score highlights the challenge of accurately segmenting minority classes (left ventricle, right ventricle, myocardium) due to class imbalances.

Table 1 presents the detailed results for each model:

Model	**F1 Score**	**Pixel Accuracy**	**IoU**
DeepLabV3	0.31	0.98	0.29
LinkNet	0.32	0.98	0.30
UNet	0.33	0.98	0.31
MAnet	0.32	0.98	0.30
Ensemble	0.34	0.98	0.32

Table 1: Performance metrics for individual models and the ensemble model.

The ensemble model slightly outperformed the individual models, achieving the highest F1 Score and IoU, demonstrating that combining multiple models improves segmentation performance.

8 Figures, and plots

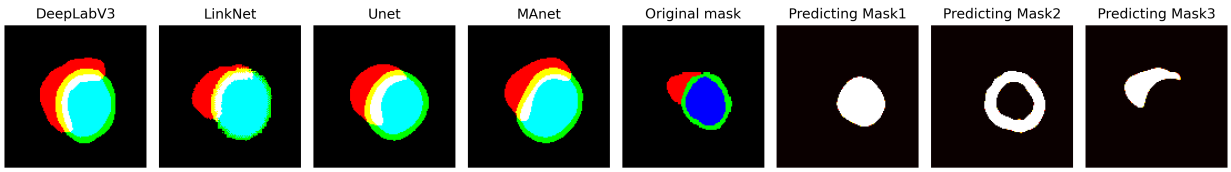

Figure 1: Prediction of the model

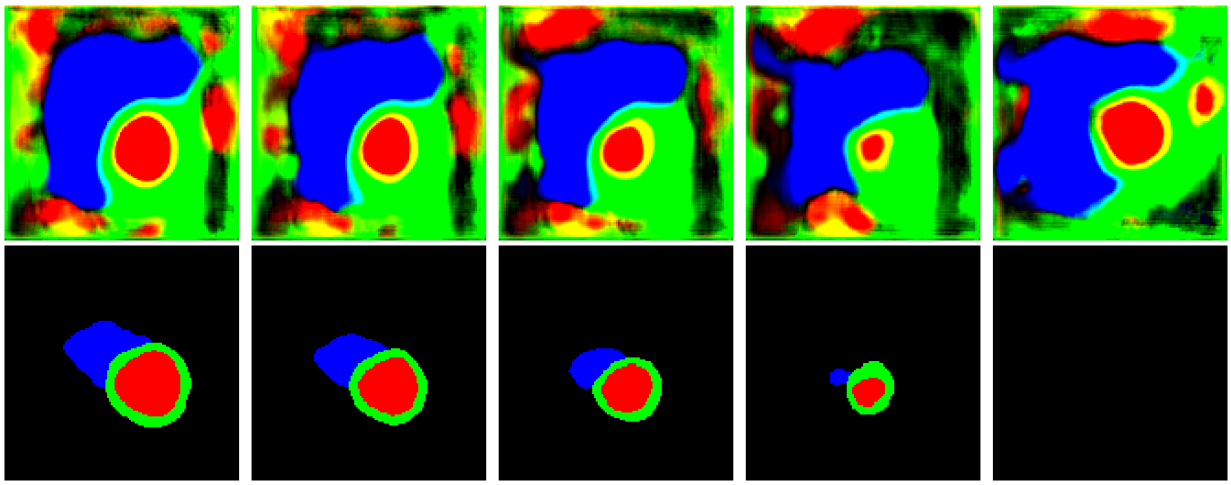

Figure 2: Model with a bad output

Figure 3: training loss

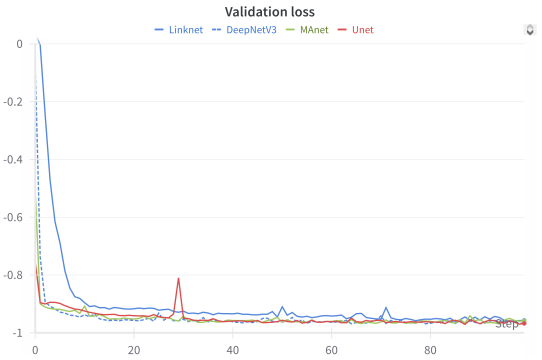

Figure 4: Validation loss