Spleen Segmentation Task – Final Project Report

Zou Han
King's College London

STEP 1

Full Volume Evaluation Results: Average Dice Score: **0.4267.**

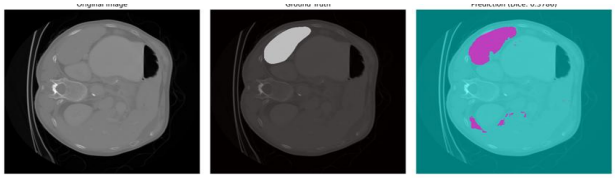

Fig. 1. Full Volume Evaluation Results

Train Loss and Val Loss & Validation Dice Score

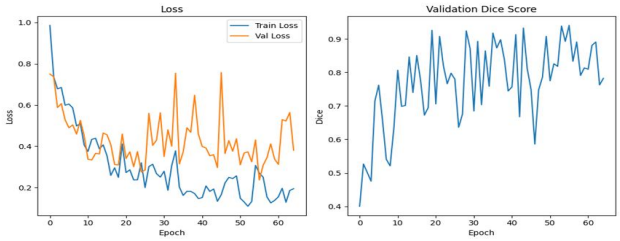

Fig. 2. Loss and Dice Score Metrics

Description of the training process

The spleen segmentation task was implemented using the 3D HighResNet model combined with the Top-K Loss function:

The spleen segmentation task used a 3D HighResNet model with a Top-K loss function. The dataset consists of 41 CT volume images, with 32 for training and 9 for validation (80%-20% split). To fit GPU memory limits, the data was divided into (64 × 64 × 64) voxel patches. The model ' s backbone is 3D HighResNet, made up of three sets of residual blocks with dilation rates of 1, 2, and 4. Dilation convolutions expand the receptive field while maintaining parameter count and resolution. The Top-K loss focuses on the top 10% of the most difficult pixels, helping the model learn challenging areas like boundaries. The Adam optimizer was used with a learning rate of 1e-4, training for 65 epochs with a batch size of 8 to prevent memory issues. After each epoch, the Dice coefficient was calculated on the validation set, and the model with the highest Dice score was saved, tracking both training and validation loss trends.

Training Process

The model demonstrated a staged learning process. In the initial phase (epochs 1-5), training loss dropped quickly from 0.9855 to 0.5990, and the validation Dice score improved significantly from 0.4004 to 0.7147. In the middle phase (epochs 10-30), training loss continued to decrease, while validation Dice fluctuated between 0.6 and 0.9, indicating ongoing optimization and exposure to more complex samples. In the later phase (epochs 30-65), training loss decreased to 0.1948, and the validation Dice score peaked at 0.9404, signaling model stabilization and convergence.

Result Description

After 65 epochs, the model achieved excellent results on the validation set, with a peak Dice score of 0.9404 at epoch 56. The final training loss was 0.1948, and validation loss was 0.3809, indicating some overfitting. When evaluated on 9 full-sized validation volumes, the average Dice score was 0.4267, lower than the patch-based validation score. The Dice scores for individual volumes ranged from 0.2216 (spleen_27) to 0.5718 (spleen_2), with spleen_14 at 0.5323. This discrepancy suggests challenges in processing full 3D volumes, likely due to the complexity of organ shapes and interference from surrounding structures.

The model struggled to generalize from local features to the global context, as shown by the gap between the high validation patch Dice score and lower full volume score. Misclassification of neighboring organs, like the liver, was common, likely due to similar textures and grayscale values. For example, spleen_46 had an accurate visual segmentation but a low Dice score, suggesting that traditional metrics may not fully capture clinical relevance.

Due to memory and GPU limitations, issues such as large learning rates, inappropriate batch sizes, or suboptimal optimizer configurations contributed to the fluctuations in the training curves. Additionally, the lack of data augmentation led to overfitting, further affecting the model's generalization ability.

STEP 2

Full Volume Evaluation Results: Average Dice Score: **0.5295.**

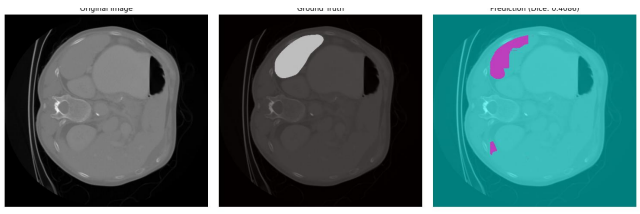

Fig. 3. Full Volume Evaluation Results

Train Loss and Val Loss & Validation Dice Score

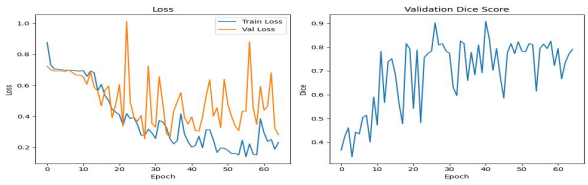

Fig. 4. Loss and Dice Score Metrics

Description of the augmentation and parameters
(Using Noise, Contrast Adjustment, and Gaussian Blur)
Noise Addition
Parameters:

- Probability: 50% chance of applying noise to each image
- Noise level: Random value between 0.01 and 0.05 (standard deviation of Gaussian noise)
- Clipping: Applied to maintain the [0,1] intensity range

Rationale:
Noise simulates CT imaging quality variations due to scanner settings. The noise levels are subtle to avoid damaging anatomical features while introducing enough variance to improve model robustness. The 50% chance ensures the model also trains on clean images.

Contrast Adjustment
Parameters:

- Probability: 50% chance of applying contrast adjustment
- Contrast factor: Random value between 0.8 and 1.5
- Reference point: Mean intensity value of the image
- Clipping: Applied to maintain the [0,1] intensity range

Rationale:
Contrast variations are common in medical imaging due to different acquisition protocols and patient characteristics. The contrast factor range (0.8-1.5) was carefully selected to create realistic variations without causing extreme distortions that would compromise anatomical structures. Decreasing contrast (factors < 1.0) simulates low-contrast scans, while increasing contrast (factors > 1.0) improves visibility of certain structures. The operation preserves the mean intensity, avoiding overall brightness changes.

Gaussian Blur
Parameters:

- Probability: 30% chance of applying Gaussian blur
- Sigma: Random value between 0.4 and 0.8 (controls blur amount)

Rationale:
Gaussian blur simulates variations in image resolution and partial volume effects commonly encountered in medical imaging. The lower probability (30%) was chosen because excessive blurring could remove critical edge information needed for accurate segmentation. The sigma range (0.4-0.8) introduces moderate blur that represents realistic resolution variations without severely degrading image quality.

Description of the performance gains
The training dynamics showed slow but consistent improvement during the initial epochs. Around epoch 12, there was a significant boost in the validation Dice scores, reaching approximately 0.78. The best validation Dice score of 0.9071 was achieved at epoch 40.

Performances Comparing (Table 1):

- Without augmentation, there was a large gap between the validation Dice (0.9404) and full volume evaluation (0.4267). With augmentation, the gap reduced, showing better generalization despite a slight drop in validation score.
- The full volume Dice improved by 24% (from 0.4267 to 0.5295), indicating that augmentation enhanced real-world performance and generalization.
- Augmented training showed more fluctuations in validation metrics, which is expected due to data variability, but these fluctuations improved the model's generalization ability.

Metric	Without Augmentation	With Augmentation
Final Training Loss	0.1948	0.2310
Best Validation Dice	0.9404	0.9071
Full Volume Dice	0.4267	0.5295
Training Time	Comparable	Slightly longer

Case ID	Without Augmentation	With Augmentation	Change
spleen_8	0.4907	0.7915	+0.3008
spleen_46	0.3786	0.4086	+0.0300
spleen_14	0.5323	0.7151	+0.1828
spleen_2	0.5718	0.8410	+0.2692
spleen_41	0.4180	0.4577	+0.0397
spleen_33	0.4999	0.6945	+0.1946
spleen_27	0.2216	0.1010	-0.1206
spleen_63	0.3815	0.4216	+0.0401
spleen_45	0.3461	0.3349	-0.0112
Average	**0.4267**	**0.5295**	**+0.1028**

Table.1. The Comparison Data

Performance improved conclusion:
The model showed improved robustness to variations in contrast, noise, and resolution, reflected by better full-volume performance. The lower validation Dice but higher test Dice indicates reduced overfitting, with the model focusing on generalizable features. The 24% improvement in full-volume Dice confirms that augmentation effectively addressed the limited training data challenge, allowing the model to handle diverse spleen appearances and imaging conditions.

STEP 3

Description of the optimisation process
The optimization adopted a staged grid search approach, optimizing each parameter individually to reduce computational complexity, though this may have missed interactions between parameters. After each stage of optimization, the best model was saved and the parameter values from that stage were used for the next stage of optimization.

Through experimentation, the optimal k value for the TopK Loss was found to be 10, influencing the model's ability to learn from difficult regions. The best learning rate was 1e-4, affecting convergence speed and stability. Finally, 16 filters were found to provide the best performance, balancing model complexity and feature representation.

Hyperparameter Optimization Results:

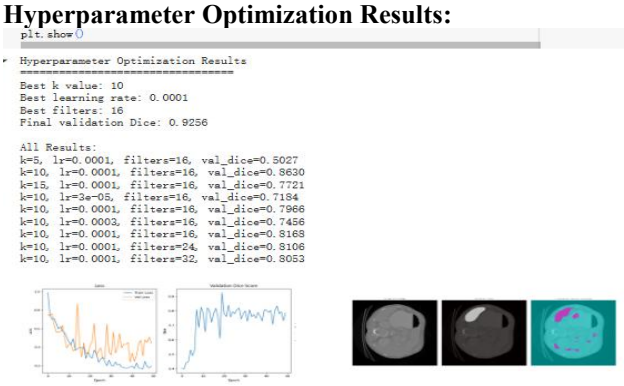

Fig. 5. Step 3 Results

Description of the performance gains

The optimized model from Step 3 achieved a validation Dice score of 0.9256, improving from 0.9071 in Step 2, highlighting the benefits of hyperparameter optimization.

However, the model scored only 0.3826 on full-volume evaluation, compared to 0.5295 from the data augmentation model, indicating that validation improvements did not translate to better full-volume results. The model still struggled with misclassifying non-spleen regions, underscoring the challenge of generalizing from validation to real-world segmentation.

My analysis of the occurrence of this issue :

The problem likely occurred because the hyper_parameter optimization focused too much on patch features, leading to overfitting, and didn't take full-volume performance into account. Full-volume evaluations were not conducted regularly during training, causing misdirected optimization. Additionally, the parameter search range was limited, and interactions between parameters were overlooked.

STEP 4

I fine-tuned and retrained the Step 2 model using the optimization results from Step 3 and incorporated augmentation consistency training. To better visualize the results, I used the axial, coronal, and sagittal planes, making the outcomes more intuitive. (See **Fig. 6.**)

Description of the performance gains

In Step 4, consistency training was introduced using unlabeled data from the test set, with key components including a pre-trained model from Step 3 (epoch 47, validation Dice: 0.9318), a gradually increasing consistency weight (from 0.1 to 0.5), and a shorter training duration (15 epochs vs. 60). The training included 40 labeled batches, 6 validation batches, and 25 unlabeled batches, with a loss function combining supervised TopK loss for labeled data and consistency loss for unlabeled data.

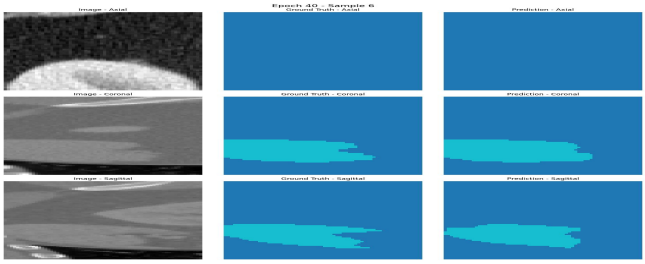

Fig. 6. Axial, Coronal, and Sagittal Planes for Visualization (optimized model)

Using Hyperparameter Optimization Results tain again:

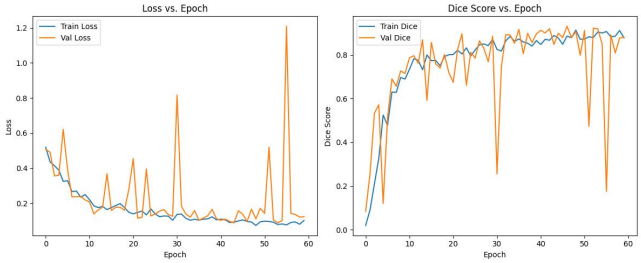

Fig. 7. New Training Loss and Dice Score Metrics

Performance Metrics Comparison

Metric	Step 2 (Data Augmentation)	Step 3 (Hyperparameter Optimization)	Step 4 (Consistency Training)
Best Validation Dice Score	0.9071	0.9318	0.8983
Training Loss	0.2310	0.0953	7.9548+
Full Volume Dice Score	0.5295	0.3826	0.0000

Table.2. The Comparison Data (STEP 2-4)

However, the training exhibited unstable patterns: the training loss was abnormally high (ranging from 7.9548 to 16.5202, compared to 0.0953 in Step 3), and the validation Dice fluctuated significantly, peaking at 0.8983 in epoch 5, then dropping to 0.3187 by epoch 15. Full-volume evaluation failed with a Dice score of 0.0000, indicating complete segmentation failure. Visualizations showed that the model struggled to identify the spleen in the validation volumes after consistency training.

GPU performance limitations severely impacted the consistency training in Step 4, leading to a catastrophic full-volume Dice score of 0. The restricted computational resources forced the use of smaller batch sizes, fewer training epochs, and a simplified network architecture, resulting in unstable gradient estimates and limited model capacity. Additionally, the mixed-precision and gradient

accumulation techniques used to address memory constraints may have introduced further numerical instability. The inability to fully explore hyperparameters, such as different consistency weight strategies, learning rates, or more complex architectures, hindered the model's performance. The excessive consistency weight likely led to over-regularization, causing the model to lose focus on the spleen segmentation task. Furthermore, the extremely high loss values (up to 16.5202) suggest significant numerical instability, and the model might have overfitted on noise or irrelevant features from the unlabeled data.

STEP 5

Visulization

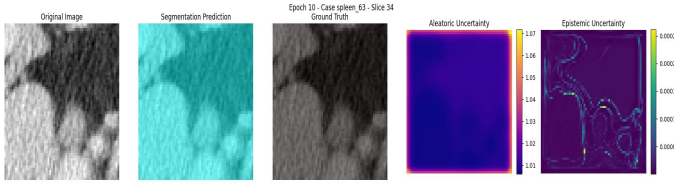

Fig. 8. Visulization Results

From left to right: Original Image, Segmentation Prediction, Ground Truth, Aleatoric Uncertainty, Epistemic Uncertainty

Description

In the spleen_63 sample at Epoch 10 (Slice 34), the uncertainty distribution shows the following patterns: Aleatoric uncertainty is primarily concentrated around the image edges and segmentation contours, with a noticeable difference between the center and the edges of the image. Meanwhile, Epistemic uncertainty is more prominent along the spleen contour edges, especially in areas with complex contours or abrupt grayscale changes.

STEP 6

One person missed the extended deadline, and another person failed to upload the required data on the last day.

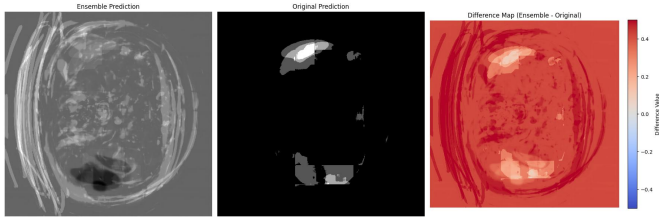

Fig. 9. Ensembled and Original Model Comparison

Description the differences in performance

The difference between the Ensemble and Original predictions indicates that both models perform well in predicting the spleen region, with the colorbar showing lighter shades, suggesting minimal differences. This implies that the predictions from both models are quite similar. However, the original model outperforms the ensemble model as it successfully captures the main spleen structure, even if some misclassifications occur. While the ensemble model is supposed to leverage the strengths of multiple models, in practice, it introduces considerable noise. Averaging the models results in significant over-segmentation (Volume difference percentage: 4,435.11%), which reduces the overall performance.

The ensemble model may exhibit higher sensitivity, capturing more potential spleen regions, which can be valuable in certain clinical scenarios. Additionally, it might better capture the spleen's boundary areas or less obvious extensions, which could be important in specific pathological conditions.

STEP 7

One person missed the extended deadline, and another person failed to upload the required data on the last day.

Manual weights used:

Zouhan: 0.7000 (Me) Karthik: 0.2000

Danman: 0.0500 Natthaya: 0.0500

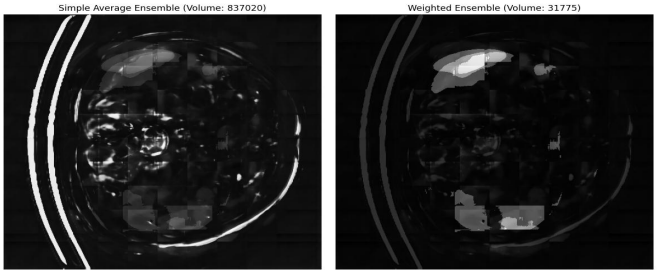

Fig. 10. Simple Average and Weighted Ensembled Comparison

The simple average approach led to severe over-segmentation, with the volume inflating to 26 times that of the original model, and the Dice coefficient being only 0.0211. In contrast, the manual weighting method, by assigning 70% of the dominant weight to the Zouhan model and sharing the remaining 30% among the other models, effectively limited the impact of models that might introduce over-segmentation. This successfully produced a high-quality segmentation result with only a 1.5% volume difference from the reference prediction, while achieving a Dice coefficient close to 1 (0.9914) and a Jaccard index of 0.9829.

Assigning 70% of the weight to the Zouhan model may lead to over-reliance on a single model, potentially limiting the useful information that could be obtained from other models. The weight distribution is also subjective, lacking a systematic approach to determine the optimal weights. Additionally, while this weighting strategy performs well on the current data, its generalization to new data has not been validated.

Reference:

Topkloss: https://arxiv.org/pdf/1605.06885.pdf
HighResNet: https://arxiv.org/abs/1707.01992
The Aleatoric and Epistemic BNN method: https://arxiv.org/pdf/1703.04977.pdf
Model ensembling: https://arxiv.org/pdf/1711.01468.pdf