Liver Tumor Segmentation Using V-Net Based Approach

Jiaqi Li

***Abstract*—In this paper, we present a deep learning-based approach to liver tumor segmentation using the V-Net architecture. We trained our model on the liver tumor dataset provided by the Medical Segmentation Decathlon. The proposed method demonstrates promising results in terms of accuracy and Dice coefficient, showing potential for application in clinical practice.**

***Keywords*—Deep learning; Liver tumor; Image segmentation; V-Net; Dice coefficient**

I. Introduction

Medical image segmentation is a fundamental process in the analysis of medical images, where images are partitioned into multiple regions, each representing a specific anatomical structure or region of interest. Liver tumor segmentation is a specific application of it. Accurate liver tumor segmentation is crucial for improving patient care, personalizing treatment approaches, and advancing research in liver disease and cancer management. This project applies medical segmentation techniques to a decathlon liver tumour task through deep learning techniques to analyse the results qualitatively and quantitatively.

II. Dataset Description

This dataset, named "Liver," focuses on liver and cancer segmentation from CT images. It consists of 3D tensor images. The dataset contains 131 training images and 70 test images. The labels include "background" (0), "liver" (1), and "cancer" (2). The dataset is designed to facilitate the development and evaluation of medical image segmentation algorithms for liver tumors, which are important for accurate diagnosis and treatment planning.

III. Data Preprocessing

Since the existing test set was ignored, we divided the dataset consisting of 131 volumetric CT images into a training set containing 89 images and a test set containing 42 images, with the aim of segmenting the images into descriptive labels. In addition, the algorithm pipeline is implemented using MONAI.

A. Loading the dataset

I used the CacheDataset and DataLoader from the MONAI database[1], because the CacheDataset is used to speed up the training and validation process, it is 10 times faster than a normal dataset:

- cache_rate = 1.0 (Cache all the data)
- num_workers = 4 (Multithreading)
- Batch size = 2 (Generate 2x8 images per batch)

B. Deterministic Training and Transforms

Using the MONAI open-source database of transformations to augment the dataset[1]:

- Load liver CT images and labels from NIfTI format files (LoadImaged)
- Construct a 'channel-first' shape (EnsureChannelFirstd) and unify the data orientation (Orientationd)
- Pixdim = (1.5, 1.5, 1.0) (Adjust the spacing)
- Intensity range = [-200, 200] and scaled to [0, 1] (ScaleIntensityRanged)
- Focus on the effective subject area of the image and label (CropForegroundd)
- Randomly crop patch samples from large images (RandCropByPosNegLabeld)

A dataset containing 89 images was split into 70 for training and 19 for validation. Since a good implementation of V-Net[1] [2] and the generalized dice score function is already implemented in MONAI by PyTorch [1][3], I used that V-Net model as my backbone network for this task and used that function as my loss function, while choosing the Adam optimizer as my optimizer.

IV. Training and Evaluation Data Analysis

The training loop iterated through a maximum of 600 epochs. And the best metric value and its corresponding epoch were tracked throughout the training process, and the model with the best metric was saved as “best_metric_model.pth”.

A. Results and performance evaluation

1) For the initial training I used a dropout probability of 0.48 in V-Net, the activation function in the loss function was set to Sigmoid and the learning rate of the optimizer was 1e-5. After 300 epochs, the best metric is 0.4001 at epoch 256. Based on the results, the segmentation algorithm pipeline is performing as expected, with training losses converging gradually and the model gradually learning the data and predicting it, but the model may be over-fitted, as shown in Fig. 1.

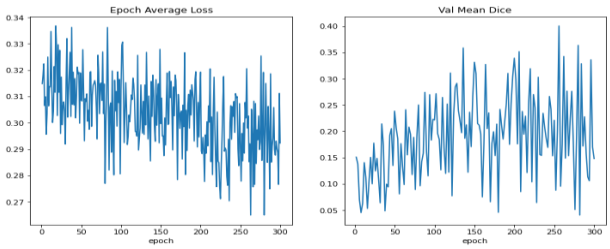

Fig. 1 Plot of initial average Dice coefficient scores on the validation set

2) Since common data enhancement techniques already exist in the MONAI open-source library, I directly invoke data enhancement techniques for Gaussian noise, contrast, and Elastic deformation. By using these data augmentation techniques, the model is exposed to a wider range of variations in the input data, which can improve its generalization performance and help it to better adapt to new and unseen data. After 300 epochs, it was found that the best dice score was obtained at epoch 296 for 0.6320. This is a better performance of the model compared to previous training without the use of data enhancement techniques, rising from a best dice score of 0.4001 to 0.6320, indicating that the training process is more robust to against overfitting and generalize better to unseen data.

3) In the process of the optimisation, I changed the dropout probability in the V-Net network parameters from 0.48 to 0.5, which helps the model to be more robust to overfitting. In the loss function section, I found that the sigmoid function used in the generalised sieve loss function did not perform well and I replaced it with a SoftMax probabilities to calculate the loss[4][5][6][7]. In the optimiser section, I adjusted the learning rate from 1e-5 to 1e-4 to make the model converge more consistently. After 600 epochs, the best metric is 0.7358 at epoch 588. Based on the results, the segmentation algorithm pipeline is performing as expected, the model is learning well, and it means that the model can segment the liver and tumor regions with a good accuracy, as shown in Fig. 2.

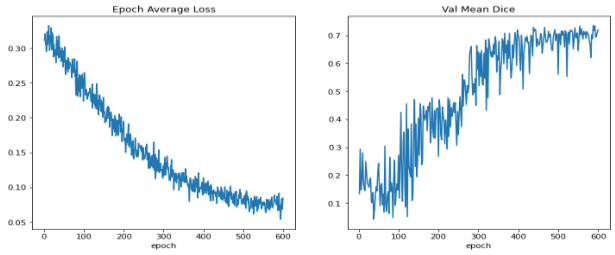

Fig. 2 Plot of average Dice coefficient scores on the validation set after optimizing.

B. *Visualizing segmentation results*

The selected model is evaluated on the test set. Due to the previous neglect of the background class, the output tensor has 2 channels, where the first channel represents the liver (label 1), the second channel represents the cancer (label 2). Visualisation of results based on predictions, only a small number of tumour labels are shown in the colour map, as shown in Fig. 3.

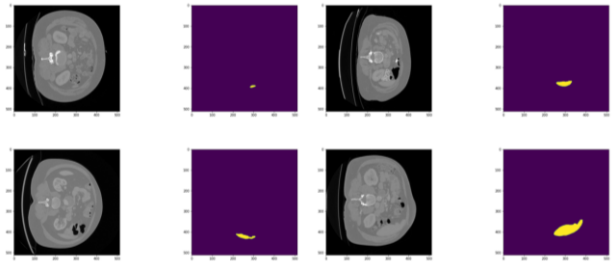

Fig. 3 Prediction graph for the test set

Fig. 3 shows that the tumour area possibly does not appear in most of the slices of the image, or it is too small to be seen in the image. This made it necessary to process the loss function later to improve the network's ability to segment small tumours.

C. *Visualizing segmentation results after adjusting the loss function*

For adjusting my loss function "Generalized Dice Loss", I use Focal Loss in conjunction with it[8]. By combining the advantages of GDL and Focal Loss, this new loss function helps the model to focus on hard-to-classify examples, while also considering the overlap between predicted segmentation and ground truth. The loss function is calculated as GeneralisedDiceFocalLoss = (1 - GDL) * Focal Loss.

This loss function is an updated version of GeneralisedDiceScore and designed to address the class imbalance issue, which is particularly important for small tumors. By doing so, Focal Loss focuses on harder-to-classify examples, such as small tumors, which often have a higher misclassification rate. In this loss function, I set the Focal weight to [0.0, 1.0, 2.0], assigning more weights to the cancer class and weights of 0 and 1 to the background and liver classes, which allows the model to learn more effectively from the underrepresented cancer class to improve the network's ability to segment small tumours. Also, the best Dice score on the validation set was 0.7416, which is an improved performance compared to the previous one.

The selected model is evaluated on the test set again. Visualisation of results based on predictions, more number of tumour labels are shown in the colour map, as shown in Fig. 4.

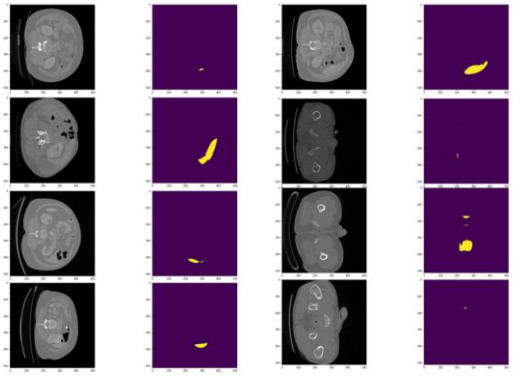

Fig. 4 Prediction graph for the test set after adjusting loss function

Fig. 4 shows that the number of tumour labels displayed in the colour map has increased compared with using only the GeneralizedDiceScore function as a loss function. As the model was previously poor at segmenting smaller liver tumours, with this combination applied, the model's ability to segment small tumours is improved.

V. Uncertainty Estimates

The model suffers from uncertainty in segmenting the target labels, which is common in medical image segmentation. Assessing these uncertainties, in turn, is critical

to improving model performance and increasing the ability of the model to achieve accurate and stable segmentation of target labels.

A. Methods

In the uncertainty estimation visualisation section, I implemented the method presented in the study by implementing TestTimeAugmentation[9] in the MONAI architecture and have defined a function to visualise the results. The change in voxel-wise variance coefficient (VVC) is a measure of the variance or uncertainty of the segmentation predictions obtained using test-time augmentation. For the test-time augmentation, I define a test transformation for TestTimeAugmentation to apply by adding Rand Affined random transform to implement invertible transformations and extract the output via a sigmoid activation function as an inference function. Number of examples is set to 10 and the network will, for a randomly selected image, generate 10 augmented versions while the model generates a prediction for each augmented version. The result is a segmentation due to differences caused by random transformations. Each segment is inverted and the results are averaged. The mode, mean, standard deviation and coefficient of volume change are returned and these are used later in the uncertainty estimation.

B. Visualisation of results

Applying Test Time Augmentation (TTA) to test datasets, we can obtain the following result: VVC is 1.68; The value of mode is 0; The value of mean in other regions is around 0.3 and the value of mean in liver regions is about 0.55; The value of std in liver class is about 0.05 and the value of std in other regions is about 0.25 and 0.35, as shown in Fig. 5.

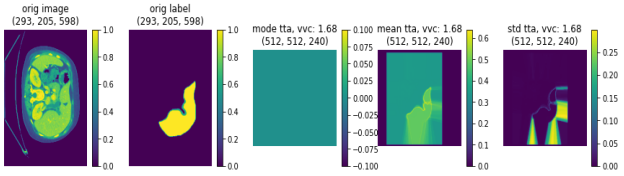

Fig. 5 Uncertainty estimate map

Fig. 5 shows that a moderate level of overall uncertainty across the TTA predictions. The result of mode suggests that the model predicts the background (or liver) class most frequently. And the results of mean and std suggest that the model is more confident about the liver class than cancer class and the predictions for the liver class are more consistent and have less variability compared to the other classes. This may be due to the cancer being too small or the absence of a cancer in the selected section not being shown.

C. Summary

In summary, there is a higher level of uncertainty in the predictions for the cancer class. In the context of medical image segmentation, high uncertainty regions might correspond to areas where the model struggles to differentiate between different anatomical structures or pathologies, such as small tumour. Areas of high uncertainty in the model's predictions are generally more error prone. High uncertainty indicates that the model is less confident in its predictions for cancer regions, which could be due to factors such as the presence of ambiguous features, artifacts, or noise in the image. As a result, the model's performance in these areas might be less accurate compared to liver areas of low uncertainty.

VI. ENSEMBLE

Our group project will focus on the “Liver Tumour” task of the decathlon. This project has two sections: an individual section where my part will tackle this task based on the V-Net. And a team section, where all members solving for the same decathlon task will share each pre-trained models to ensemble.

A. Average Ensemble

For the part of average ensemble, I evaluate the Dice coefficient on a test dataset using an ensemble of six models by averaging their probability[10]. The models are applied to input images in a sliding window manner to generate predictions, and the average of these predictions is then calculated. The final metric value will be the mean of the Dice coefficients calculated for each image in the test dataset. The code iterates over the batches in the test dataset, and for each batch, it applies each of the six models to the input images using a sliding window approach. The resulting predictions are averaged to generate the final predictions for the batch. After iterating over all batches in the test dataset, the mean of the metric values is computed using the mean function, and this value is printed as the final result.

Previously, I used my own model to evaluate the model's predictions on the test dataset and the results showed that the Dice coefficient on the test set of 42 images was 0.7757. By generalising the learning and prediction to new unseen data, this showed that my model predicted the test dataset well. According to the result of the average ensemble, the Dice coefficient is 0.7835, which improves the prediction of the model compared to using only my model, but the improvement is not very significant. However, it also indicates that the ensemble models have a better segmentation accuracy and result in more accurate and stable predictions.

B. Weighted Ensemble

For the part of the weighted ensemble method, I manually choose the weights between different models[10]. Based on each model's Dice score values from previous training and evaluation phases, I set higher weight values for the models that performed well in cancer region and set lower values for the models that performed well in other regions. The weight values for these six models correspond to 0.3, 0.2, 0.2, 0.1, 0.1 and 0.1 respectively. The predictions of each model are weighted with this predefined set of weights, and then the weighted predictions are combined into a single segmentation using the sum operation. The resulting segmentation is then evaluated using the Dice metric. Finally, the mean Dice coefficient is computed over all test samples and printed.

According to the results, a weighted ensemble was used with a Dice coefficient of 0.8139, which is better than the previous result by simply averaging the ensemble. This is because when the models have different strengths and weaknesses, and we want to emphasize the strengths of the better-performing models while downplaying the weaknesses of the worse-performing models. Moreover, weighting can also help to avoid the potential overfitting of models in the ensemble. If one model is trained on a slightly different set of training data and starts to overfit to that particular data, assigning a lower weight to that model can prevent it from dominating the ensemble and leading to suboptimal results.

VII. CONCLUSIONS

Overall, after using data augmentation techniques, adjusting the loss function and tuning the relevant hyperparameters, my V-Net-based 3D segmentation model achieved good performance in both validation and test evaluations, with average Dice scores of 0.7416 and 0.7757 respectively. The segmentation algorithm is performing as expected, the model is learning well, and training process is more robust to against overfitting and generalize better to unseen data. However, although the number of cancer predictions on the test set increased after adjusting the loss function, the model still had high uncertainty in the segmentation of cancer labels based on the uncertainty estimates, so future optimisation of the model and enhancement of its training capacity are needed to achieve high accuracy and robustness in the segmentation of cancer labels. The performance of the segmentation is greatly improved by using a single consensus model rather than using my model alone. In the weighted ensemble, Dice scored 0.8139, which illustrates the importance and robustness of the ensemble. In particular, the performance improvement of the weighted ensemble implementation is very noteworthy. In future research, ensembles can be used to improve the predictive power of the model in order to obtain better segmentation predictions.

ACKNOWLEDGMENT

Jiaqi Li would like to acknowledgment Maria Anastasia Howarth, Dewmini Hasara Wickremasinghe, Emilie Fabienne Aps, Sri Naga KamalaSraavya Kocherlakota, Maha Mohammed and Alshammari for Contribution of this "liver tumour" Task.

REFERENCES

[1] Gologorsky R, Harake E, von Oiste G, et al. Generating novel pituitary datasets from open-source imaging data and deep volumetric segmentation[J]. Pituitary, 2022: 1-12.

[2] Milletari F, Navab N, Ahmadi S A. V-net: Fully convolutional neural networks for volumetric medical image segmentation[C]//2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-571.

[3] . Sudre C H, Li W, Vercauteren T, et al. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3. Springer International Publishing, 2017: 240-248.

[4] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[5] Li W, Wang G, Fidon L, et al. On the compactness, efficiency, and representation of 3D convolutional networks: brain parcellation as a pretext task[C]//Information Processing in Medical Imaging: 25th International Conference, IPMI 2017, Boone, NC, USA, June 25-30, 2017, Proceedings 25. Springer International Publishing, 2017: 348-360.

[6] Kamnitsas K, Ledig C, Newcombe V F J, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation[J]. Medical image analysis, 2017, 36: 61-78.

[7] Wu Z, Shen C, Hengel A. Bridging category-level and instance-level semantic image segmentation[J]. arXiv preprint arXiv:1605.06885, 2016.

[8] Li X, Wang W, Wu L, et al. Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection[J]. Advances in Neural Information Processing Systems, 2020, 33: 21002-21012.

[9] Wang G, Li W, Aertsen M, et al. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks[J]. Neurocomputing, 2019, 338: 34-45.

[10] Kamnitsas K, Bai W, Ferrante E, et al. Ensembles of multiple models and architectures for robust brain tumour segmentation[C]//Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: Third International Workshop, BrainLes 2017, Held in Conjunction with MICCAI 2017, Quebec City, QC, Canada, September 14, 2017, Revised Selected Papers 3. Springer International Publishing, 2018: 450-462.