Universal Lesion Segmentation Challenge 2023: A Comparative Research of Different Algorithms

Kaiwen Shi[1], Yifei Li[2], Binh Ho[1], Jovian Wang[2], and Kobe Guo[2]

[1]Mathematics Department, Vanderbilt University, Nashville, TN, USA
[2]Department of Computer Science, Vanderbilt University, Nashville, TN, USA

March 30, 2024

Abstract

In recent years, machine learning algorithms have achieved much success in segmenting lesions across various tissues. There is, however, not one satisfying model that works well on all tissue types universally. In response to this need, we attempt to train a model that 1) works well on all tissue types, and 2) is capable of still performing fast inferences. To this end, we design our architectures, test multiple existing architectures, compare their results, and settle upon SwinUnet. We document our rationales, successes, and failures. Finally, we propose some further directions that we think are worth exploring. codes: https://github.com/KWFredShi/ULS2023NGKD.git

Keywords: Universal Lesion Segmentation, UNet

1 Introduction

Medical image segmentation is a crucial task in medical image processing. Thanks to the advent of CNN[11], U-Net [17], and their variants such as V-Net[13], 3D U-Net[5], Res-UNet[14], Dense-UNet[12], we are able to perform segmentation task with precision. More recently, with implementations of transformer-based models, the medical imaging community enjoyed satisfying success in segmentation tasks. Networks like Medical Transformers[18] and SwinUnet[1] push the front-line boundary to another degree. Others have implemented learning methodologies from other fields, such as dictionary learning, to work on medical images. KEN[15] - knowledge embedding network - for example, takes advantage of the fruitfulness of information embedding in each layer via dictionary learning to provide a more semantically meaningful network.

While most of the networks have achieved very promising results on datasets composed of one tissue type, none of them works well universally across various tissue types. Meanwhile, facing a growing need for CT exams and their lesion segmentation, radiologists have suffered much from intensive human labor. In response to these two situations, Max de Grauw, Bram van Ginneken, and Alessa Hering of the Diagnostic Image Analysis Group, partnering with Mathai Tejas, Pritam Mukherjee, and Ronald Summers of the National Institutes of Health, launched the Universal Lesion Challenge 23 (https://uls23.grand-challenge.org/).

To address the dire situation, we attempt to adopt/devise an algorithm that is 1) precise enough based on the Dice score, 2) adequately robust to respond to variation in tissue types, and 3) light-weight in model complexity so that the inference runs around 5 seconds per input. To this end, we test the effectiveness of various algorithms, including nnUNetv2[10], DeepLabV3+[3], Medical Transformer[18], SwinUnet[1], and TransUNet[2]. We compare their results and fine-tune our final model on TransUNet, the one that has worked best in our testing.

Contributions. 1. We tested the effectiveness of nnUNetv2, DeepLabV3+, Medical Transformer, SwinUnet, and TransUNet on the Bone Lesion dataset. 2. We fine-tuned the TransUNet model, with well-written data augmentations and transformations. 3. We included a discussion on potential improvements that can be made to our project.

2 Task Description

The task at hand is to design, implement, and train a model that can segment lesions universally. To be more specific, the input of the model will be a 256x256x128 data point, with only one channel. The input is named VOI, or volume of interest. The output of the model should be a segmentation of the lesion, and the performance will be measured with inference speed, segmentation accuracy, and robustness.

Figure 1: The Proposed Pipeline from ULS Challenge Host

Once the model is trained and deployed, radiologists can locate the VOI manually, pass the VOI into the model pipeline, and get the returned lesion segmentation almost immediately.

3 Experiments

In this section, we describe the rationale behind the use and test of these models, the methodologies and highlights of each model, and their corresponding results.

3.1 Rationale

We started the project by testing the effectiveness of each model on the Bone dataset, Part 1. We split the data into a 0.7-0.2-0.1 train-test-validation compartmentalization. We trained the models on the train data and compared the validation scores. We understood that the model that worked well on the Bone dataset did not necessarily tell its effectiveness on the entire data population, so after having initial results that did not deviate too much, we further trained the working models on the whole Part 1 data. We again compared the results and went on with the most promising model to complete the project.

3.2 Models and Results

3.2.1 nnUNetv2 - Baseline

nnUNetv2[10] is an advanced neural network architecture designed for medical image segmentation tasks. Built upon the success of its predecessor, nnUNet, nnUNetv2 incorporates several enhancements and optimizations, making it a state-of-the-art solution in the field of medical image analysis.

nnUNetv2 adopts a cascaded architecture, consisting of multiple processing stages, each refining the segmentation output progressively. The network architecture comprises deep convolutional neural networks (CNNs), augmented with advanced modules such as residual connections, dilated convolutions, and attention mechanisms. These components enable nnUNetv2 to capture complex spatial dependencies and contextual information crucial for accurate segmentation. Some of its highlights include performance, versatility, and robustness.

nnUNetv2 served as the baseline model in the ULS 23 challenge. It has produced very promising results on data points from one tissue type. However, when training on multi-typed data, the model exhibited a large variation across tissue types and even within each tissue type.

This model was evaluated using 10% of each fully-annotated dataset, split on a patient-level. The scores are aggregated per lesion type.

Lesion Type	DICE
Kidney	0.77 ± 0.21
Lung	0.71 ± 0.14
Lymph Node	0.70 ± 0.18
Bone	0.68 ± 0.24
Liver	0.65 ± 0.17
Pancreas	0.64 ± 0.19
Colon	0.55 ± 0.21

Figure 2: Baseline Result

Since the nnUNetv2 pipeline had written optimizers to tune its hyperparameters, we did not continue fine-tuning it. We instead use it as a benchmark against which we test our experiments.

3.2.2 Medical Transformer

Medical Transformer[18] utilized the LOGO learning strategy and Gated Axial Attention. The former strategy helps to discern the local features while also caring for the global context[6], and the latter Attention mechanism saves time complexity while giving more attention to the axial components.[9]

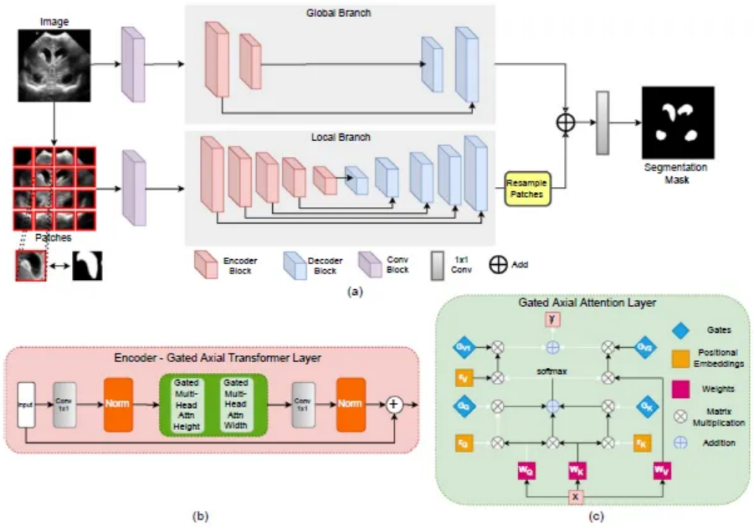


Figure 3: LOGO Learning and Gated Axial Attention

The structure of Medical Transformer gave it prime effectiveness in segmentation tasks on MoNuSeG, GLAS, and Brain Anatomy US datasets. The problem of adopting this model, however, lay with its fixed input dimension and type. The input of the original Medical Transformer was a 128x128 image, while the VOI that was put into the model pipeline was a 256x256x128 volume. The disagreement with regard to the inputs called for modification of the existing model.

The team primarily experimented with three different potential fixes. The first one involved simple flattening of the VOI into 2D images and adjusting the model parameters according to the size of the image; the second one utilized an FCNN to extract the features from the input and fit the extracted

feature into the model; the last one, being the most straightforward solution, was to implement a 3D version of Medical Transformer.

None of the fixes, unfortunately, worked well to address the current challenge at hand. The first two fixes, no matter where the feature extraction filters were put or simple flattening, took more than 100 GB of RAM to initialize the model. The last solution, even though it seemed more promising, suffered from the same problem of exploding model size. Therefore, the Medical Transformer failed to solve our problem.

3.2.3 SwinUnet

The SwinUNet model[1] differs from UNet due to its use of the Swin transformer blocks. It uses a hierarchical Swin transformer with shifted windows as the encoder and a symmetric Swin transformer as the decoder. This allows the model to adeptly capture complex patterns in medical images. The hierarchical design lets the model efficiently process different scales of spatial context.

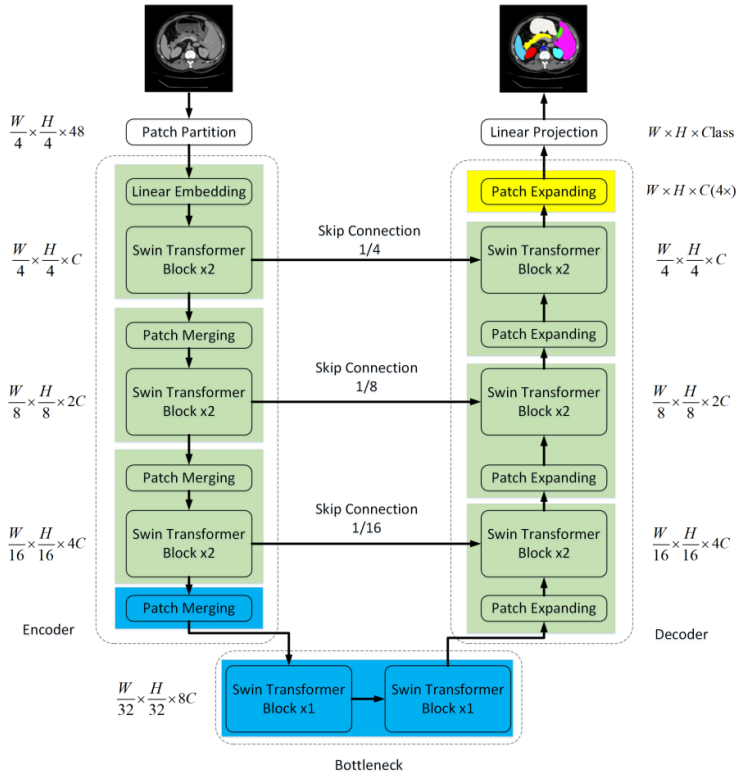


Figure 4: SwinUNet Architecture

Our experimental process began with training the model on the ULS23 bone dataset. Through the training, we were able to track the model's training dice loss and validation dice loss metrics. The training dice loss had a consistent downward trend, which indicated an increase in dice score, but the validation dice loss had upward spikes throughout the iterations. This showed us that the model could have been overfitting on the dataset.

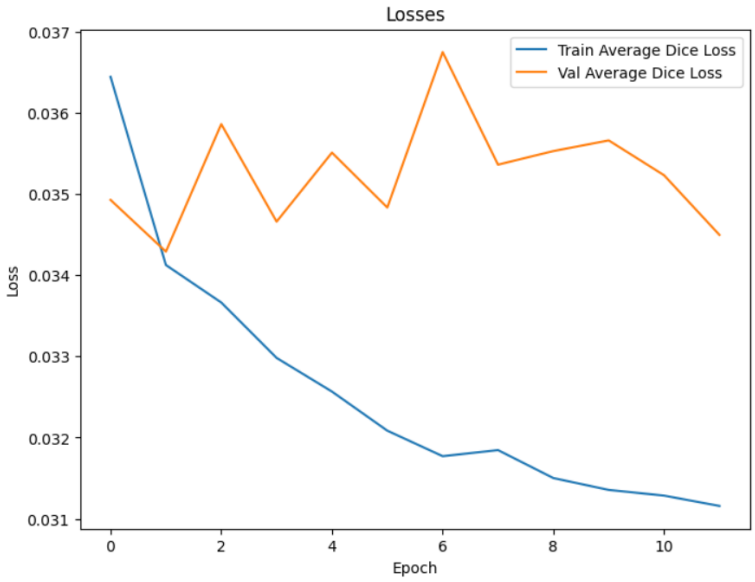


Figure 5: SwinUNet Result

To mitigate overfitting, we employed data augmentation techniques such as random rotations, horizontal and vertical flips, and Gaussian blurs onto the medical images in the dataset. Even through this, however, the validation training loss did not improve much, which led to this model not being our best one.

3.2.4 DeepLabV3+

The DeepLabV3+[4] is designed with a four-layer architecture featuring downsampling and upsampling blocks, integrated with skip connections to facilitate information flow. In the downsampling path, convolutional layers extract and learn features using convolutional filters followed by max pooling. The upsampling path then reconstructs the spatial dimensions through transposed convolution. Skip connections bridge these paths, allowing information to be directly passed from downsampling blocks to the corresponding upsampling blocks to mitigate the loss of finer details during the simplification of the image.

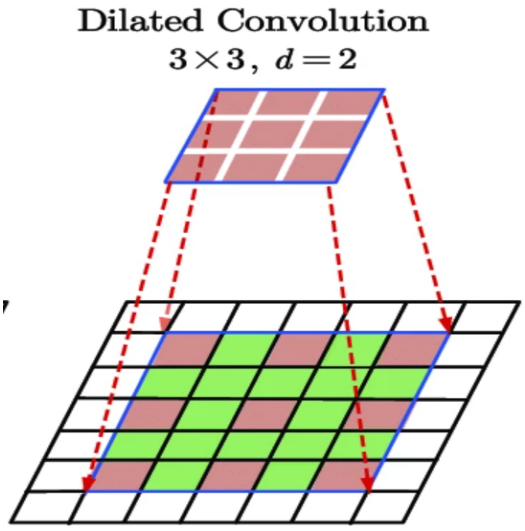


Figure 6: Atrous/Dilated Convolution

We began the experiment with the 2D DeepLabV3[3] model. We flattened the input directly and sent it into the inference pipeline. Initially promising, this model displayed tendencies of overfitting, as evidenced by its attainment of a 0.82 Dice coefficient following 2500 epochs of training—surpassing the baseline model's 0.72 Dice score significantly on the bone dataset.

In response, we ventured into a 3D U-Net model, referred to as DeepLabV3+, owing to its atrous convolution (or dilated convolution) features similar to the DeepLabV3 architecture. However, the complex structure of this 3D model introduced significant challenges, notably extending the training duration. To overcome the challenges, we tried various local pooling techniques before inferencing, but the result turned out bleak: After 25 epochs of training, the dice score was only around 0.1, which made this model nearly useless for our purpose.

3.2.5 TransUNet

TransUNet[2] is a model that aims to improve the performance of the UNet architecture by leveraging Transformer blocks as the encoders for medical image segmentation tasks. In particular, while the UNet architecture is better than pure Transformers-based models in extracting low-level details, the UNet architecture alone is not sufficient in modeling long-range dependencies, a drawback that is directly addressed by the introduction of Transformer blocks. Therefore, by using Transformer blocks as the encoders in a UNet-like architecture, TransUNet is able to efficiently tackle both long-range dependencies as well as low-level details. In their original paper, the authors of TransUNet show that TransUNet greatly outperforms other contemporary models in segmentation tasks related to the aorta, left kidney, pancreas, and spleen.

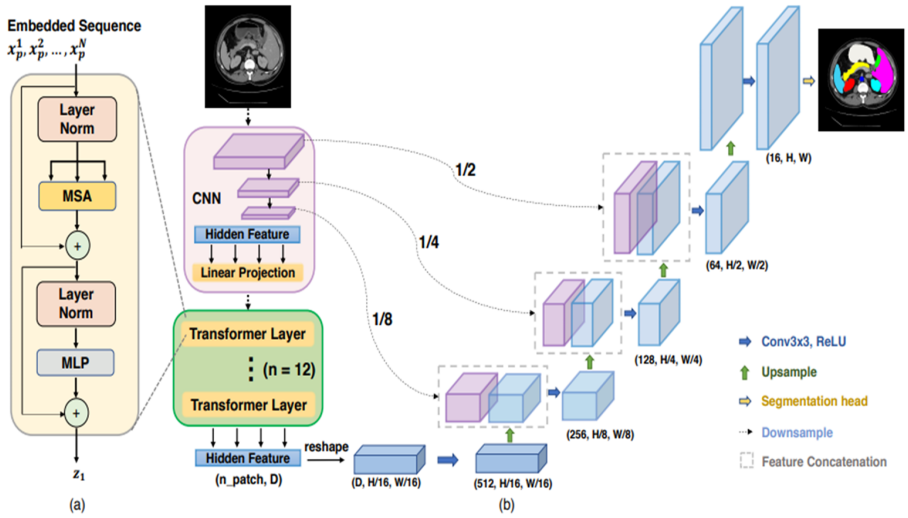


Figure 7: TransUNet Model

In our experiments, we observed that TransUNet achieved a high dice score on the bone dataset, with a training and validation dice score of 0.87.

Figure 8: TransUNet Result 1

However, when we expanded the dataset to also include data from the pancreas dataset, the training dice score dropped to 0.80 and the validation dice score dropped to 0.79. We speculated that this is because the TransUNet architecture is able to perform well when it is tasked with a single type of tissue and that its performance deteriorates when it has to work with a dataset that contains different types of tissues.

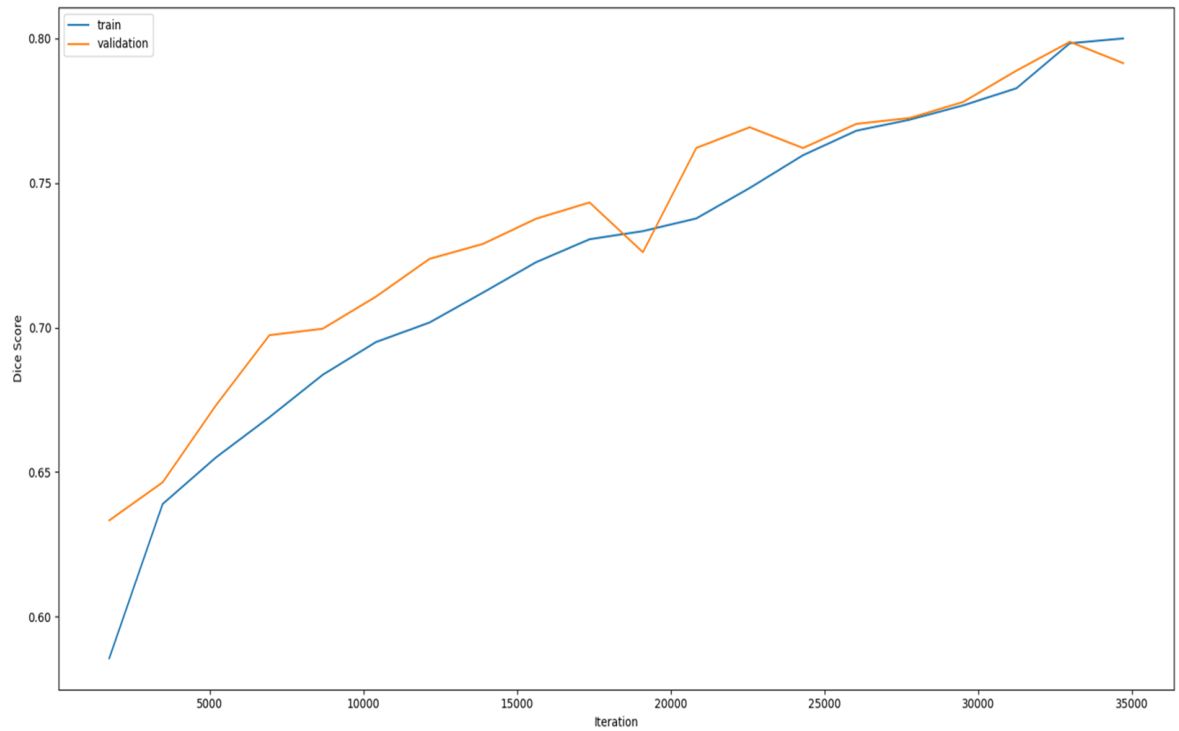


Figure 9: TransUNet Result 2

3.3 Results

The validation results of the working models on the Bone dataset are shown below.

Model Comparison on Novel Bone Lesion Dataset

	Baseline	Candidate Algorithms		
Model	Residual 3D U-net	Modified DeepLabV3+	Swin-Unet	TransUNet
Dice Score	0.680 ± 0.24	0.823	0.579	0.864

Figure 10: Model Result Comparison

Based on the comparison, we decided to further train and fine-tune the TransUNet model. The last time we looked at its training curve showed its potential to generate a comparable result against the baseline. Below are some quantitative and qualitative results of TransUNet.

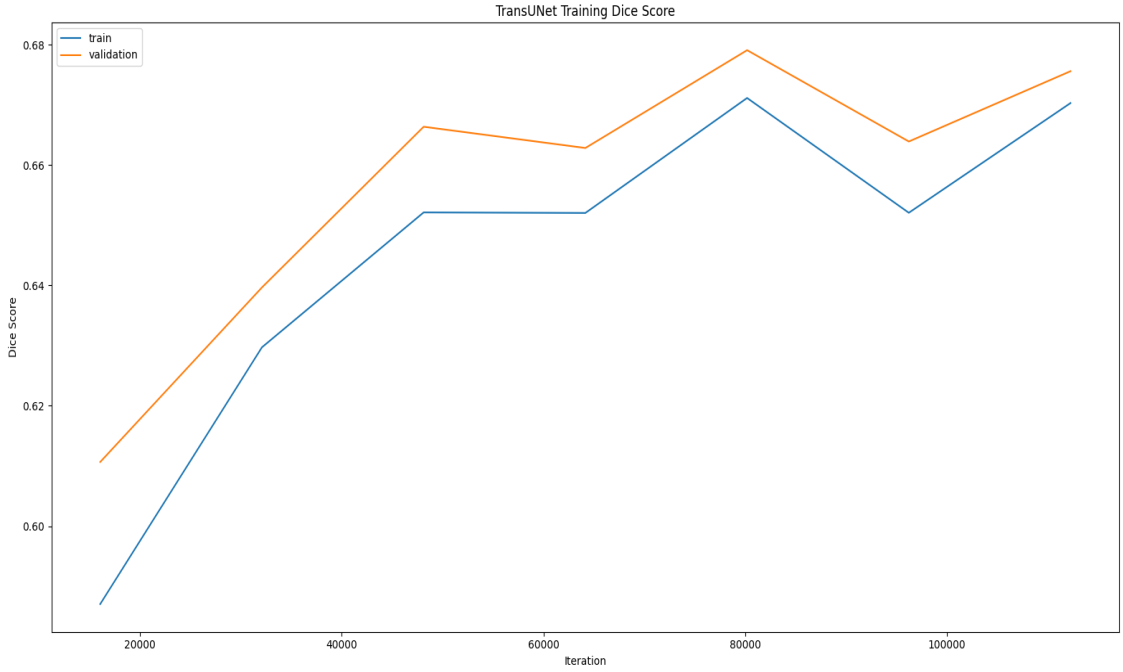
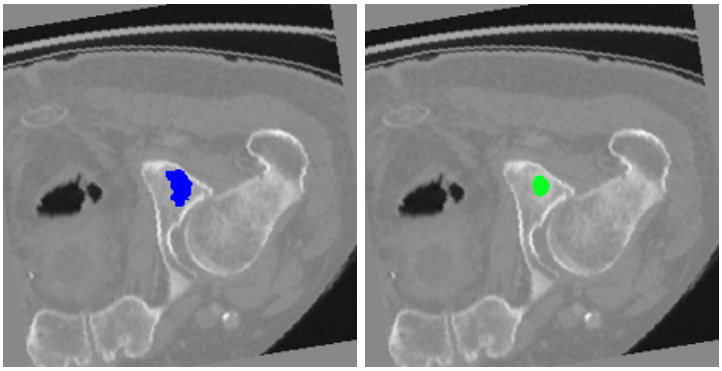
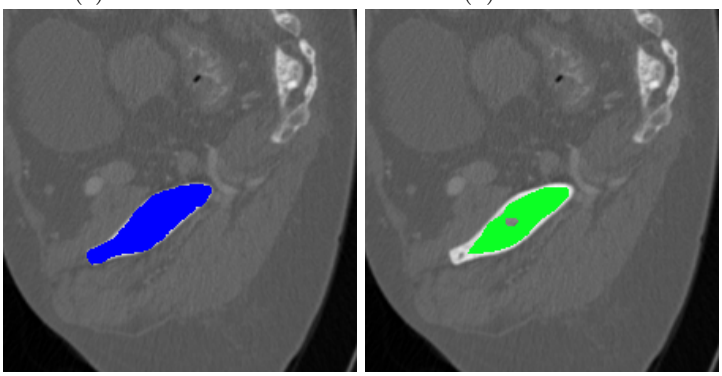


Figure 11: TransUNet Result 3



(a) Ground Truth 1 (b) Inference 1



(c) Ground Truth 2 (d) Inference 2

Figure 12: TransUNet Qualitative Results

4 Discussion

Most of our experimented models did not work as intended and failed due to out-of-memory error. Some of them were able to run but suffered very badly when we trained our models across various parts of the dataset. In this section, we will attempt to find the reasons behind these phenomena and propose potential remedies.

4.1 Model Size

Model sizes have been always an issue during our experimentation. Many models failed to run because the RAM requirements for their initialization were too ambitious, exceeding our current 50 GB capacity.

The reason behind this gigantic memory consumption is straightforward: the original models typically work on small data points, such as a 128x128 image; however, once we adopted it to work on the data from the ULS challenges, the immense increase in the size of the data jumped the magnitude of the parameters up, thus creating a barrier for initializing the models.

Now, there are some potential remedies. One is to combine data parallelism, pipeline partition, and model parallelism.[7] This way the model can be split across many devices, and potentially, we can train a more complex model that can provide better results.

After getting a well-performing giant model, we can distill its knowledge to small-scale, easy-to-deploy specialist models.[8] After all the specialist models are trained, we can train a gated model on the giant model, distributing the data points to its corresponding specialists. This essentially composes a "mixture of experts" model[16], which could deviate from the original intent of the challenge host. However, it sounds to us a very effective solution to boost up the segmentation accuracy without sacrificing much of the inference speed.

4.2 Contextual Information

Another compelling reason behind the inability of the models to perform well is, perhaps, innate to the problem at hand. One single model, with its predefined architecture, may fail to capture the variation of all the tissue types. That is to say, with the limited computation resources and strict requirements on the inference speed, it is impossible to have a universal model working well on all tissue types.

To address this issue, we found Knowledge Embedding Network[15], which follows dictionary learning principles to carefully select a collection of vocabularies and incorporate the context information into the inference layers using that collection. The model, trained for several epochs, was not satisfying to us concerning the training and validation loss. We encourage further investigation.

5 Conclusion

In response to the growing demand for robust and efficient lesion segmentation models across various tissue types, a comparative research study was conducted as part of the Universal Lesion Segmentation Challenge 2023. We aimed to develop a model capable of universal lesion segmentation while maintaining fast inference times. Several state-of-the-art architectures were evaluated, including nnUNetv2, DeepLabV3+, Medical Transformer, SwinUnet, and TransUNet. Most of them did not perform well, and none of the experiments beat the baseline. Problems involved a lack of RAM in initializing models, insufficient segmentation accuracies, and etc. Based on our results, we continued training TransUNet and included a qualitative demonstration of its inferences on some slices. Finally, we discussed what could be done to potentially have a better model for the universal segmentation task.

References

[1] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation, 2021.

[2] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L. Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *CoRR*, abs/2102.04306, 2021. URL https://arxiv.org/abs/2102.04306.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *CoRR*, abs/1706.05587, 2017. URL http://arxiv.org/abs/1706.05587.

[4] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. *CoRR*, abs/1802.02611, 2018. URL http://arxiv.org/abs/1802.02611.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016. URL http://arxiv.org/abs/1606.06650.

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

[7] Amir Gholami, Ariful Azad, Kurt Keutzer, and Aydin Buluç. Integrated model and data parallelism in training neural networks. *CoRR*, abs/1712.04432, 2017. URL http://arxiv.org/abs/1712.04432.

[8] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[9] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *CoRR*, abs/1912.12180, 2019. URL http://arxiv.org/abs/1912.12180.

[10] Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203 – 211, 2020. URL https://api.semanticscholar.org/CorpusID:227947847.

[11] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, R. Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. In D. Touretzky, editor, *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann, 1989. URL https://proceedings.neurips.cc/paper_files/paper/1989/file/53c3bce66e43be4f209556518c2fcb54-Paper.pdf.

[12] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: Hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018. doi: 10.1109/TMI.2018.2845918.

[13] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, 2016. doi: 10.1109/3DV.2016.79.

[14] Xiao nan Xiao, Sheng Lian, Zhiming Luo, and Shaozi Li. Weighted res-unet for high-quality retina vessel segmentation. *2018 9th International Conference on Information Technology in Medicine and Education (ITME)*, pages 327–331, 2018. URL https://api.semanticscholar.org/CorpusID:57190934.

[15] Yu Qiu and Jing Xu. Delving into universal lesion segmentation: Method, dataset, and benchmark. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 485–503, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20074-8.

[16] S. J. Nowlan R. A. Jacobs, M. I. Jordan and G. E. Hinton. Adaptive mixtures of local experts., 1991.

[17] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, 2015.

[18] Jeya Maria Jose Valanarasu, Poojan Oza, Ilker Hacihaliloglu, and Vishal M. Patel. Medical transformer: Gated axial-attention for medical image segmentation. *CoRR*, abs/2102.10662, 2021. URL https://arxiv.org/abs/2102.10662.