Benchmarking BraTS Baseline Using nnU-Net

Keru Zheng
The Chinese University of HongKong,Shenzhen

November 7, 2024

Abstract

This report evaluates the nnU-Net framework on the Brain Tumor Segmentation (BraTS) dataset in 2017 by comparing its performance against the established BraTS baseline. The focus is on understanding the framework's adaptability and effectiveness in segmenting brain tumors through advanced metrics and direct comparison.

1 Introduction

1.1 BraTS Dataset

The Brain Tumor Segmentation (BraTS) dataset is a collection of multimodal MRI scans that are used widely for the evaluation of algorithms that perform the segmentation of brain tumors. BraTS focuses on the segmentation of gliomas from pre-operative MRI scans, delineating the boundaries of high-grade gliomas (HGG) and low-grade gliomas (LGG). The dataset includes MRI scans along with manually-verified tumor annotations and has been used in several benchmarking challenges within the medical image analysis community to improve the state-of-the-art in brain tumor segmentation. In this paper, nnU-Net shows state of art performance for in BRATS 2017 training, validation and testing dataset.

1.2 nnU-Net Framework

The nnU-Net framework, developed by Isensee et al., represents a significant advancement in medical image segmentation. It builds on the foundational architecture of traditional 2D and 3D U-Nets, enhancing them through a self-adapting system that automates much of the parameter tuning typically required in neural network configurations (Isensee et al., 2018). This automation includes dynamic adjustments of input patch sizes and network architectures to optimally fit diverse datasets, thus eliminating the need for manual intervention across varying medical segmentation tasks.

Critically acclaimed for its performance, nnU-Net has demonstrated high efficiency in the Medical Segmentation Decathlon challenge, showcasing its capability to achieve superior results across a range of medical imaging tasks, modalities, and dataset complexities. The framework's ability to maintain state-of-the-art performance without dataset-specific modifications underscores its robustness and versatility in handling the intrinsic variabilities of medical image data.[2] The diagram depicted in Figure 2 illustrates the architecture of the nnU-Net framework. This schematic provides a detailed view of the various components and operational layers that constitute the nnU-Net, highlighting its modular and adaptive design tailored for biomedical image segmentation.

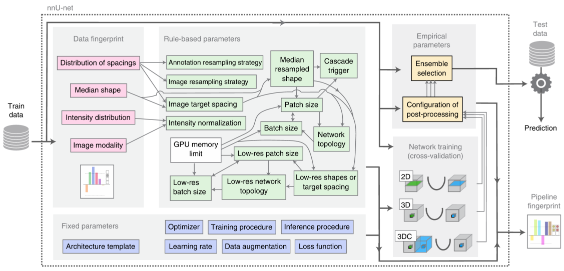

Figure 1: A detailed diagram of the nnU-Net framework.

2 Data Preparation

2.1 Data Acquisition

The dataset is 'Task01 BrainTumor' from the BraTs 2017 dataset. It was download from `http://medicaldecathlon.com/`, which is preferred for its stability and efficient access to the data. Here are more detailed introduction.

Each case within the dataset encompasses four modalities (multimodal) and is segmented into three parts (labels), encompassing a background inclusion which results in four labels in total. The four modalities—T1, T2, FLAIR, and T1ce—represent commonly employed imaging techniques in medical imaging that depict the same anatomical structures but through different acquisition methods or "dimensions". This multimodal approach provides a comprehensive representation of the targeted anatomical features.

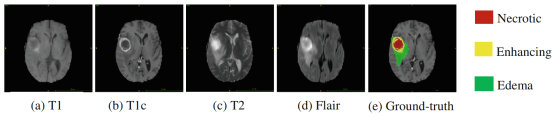

Figure 2: A slice of BraTS 2017 training data with four modalities such as (a) T1, (b) T1c, (c) T2, (d) Flair, (e) Ground-truth with experts' segmentation.[3]

The segmentation labels include the background, necrotic and non-enhancing tumor tissue (NET), peritumoral edema (ED), and enhancing tumor tissue (ET). These labels collectively facilitate the delineation of the whole tumor (WT), enhancing tumor (ET), and tumor core (TC), enabling a detailed and nuanced understanding of the tumor's structural composition and progression.

2.2 Subset Creation

It's necessary to create a subset to reduce training time and prevent GPU memory overload. The subset follows the original dataset's training to testing ratio of approximately 1:0.64, resulting in 100 training cases and 64 testing cases. This strategic reduction not only optimizes GPU usage but also ensures efficient handling during training. Random sampling ensures the statistical integrity and minimizes selection bias, which effectively captures the diversity and true characteristics of the raw dataset.

2.3 Data Preprocessing

In the context of utilizing nnU-Net, a robust framework for medical image segmentation, it is essential to adhere to specific organizational protocols for dataset management:

1. **Dataset Storage Requirements:** Datasets must be located within the `nnUNet_raw` directory. nnU-Net mandates rigorous structuring of training data storage, as delineated by the framework's specifications. This requires strict adherence to a predefined directory hierarchy encompassing raw, processed, and model data to ensure consistent data management and facilitate systematic training and validation processes.

2. **Dataset Format and Conversion:** The dataset sourced from the Medical Decathlon website (`http://medicaldecathlon.com/`), specifically 'Task01_BrainTumour', is formatted to be compatible with the initial version of nnU-Net. With the evolution of nnU-Net to its second version (nnUNetv2), as hosted on GitHub, it becomes necessary to convert datasets to be compatible with the newer version. This transition is crucial due to significant format discrepancies between the two versions.

3. **Correct Conversion Command:** It is critical to recognize that the dataset conversion command in the nnU-Net GitHub repository is incorrect. Instead, the correct command for converting a version 1 dataset to nnUNetv2 format is `nnUNetv2_convert_MSD_dataset`. Utilizing this command effectively converts and stores the dataset under a new designation, such as 'Dataset900_BrainTumour', within the `nnUNet_raw` directory.

By following these structured steps and utilizing the correct commands, we ensure the datasets are properly formatted and stored, thereby facilitating the successful application of the nnUNetv2 framework for medical image segmentation tasks.

3 Training

3.1 Operating System & Hardware requirements

As the training takes for long time and requires substantial memory resources, we choose Linux (Ubuntu 18.04) as operating system. Also, we use RTX 3090, A6000 as our GPU.

3.2 Adapt the configuration of nnU-Net

The training was initiated with nnU-Net default configuration - 1000 epoch, batch size 105. The average duration per epoch is 60 to 70 seconds. If we run in the default configuration, the total training duration is at least for 25 hours and more!

The graph shows that between the 50th and 100th epochs, both the training loss (blue line) and validation loss (red line) exhibit a trend towards stabilization. The validation loss stabilizes at a lower level after initial significant fluctuations, indicating an improvement in the model's adaptation and generalization to new data during this phase. The Dice coefficient stabilizes after about 100 epochs, further verifying that the model's performance gradually stabilizes with ongoing training and is approaching its optimal performance.

Given the analysis of the behavior of the loss and validation metrics, it is indeed sensible to adjust the definition value of epochs to 100 for this training scenario. This adjustment allows the model to stabilize and generalize better before further evaluations or deployment.

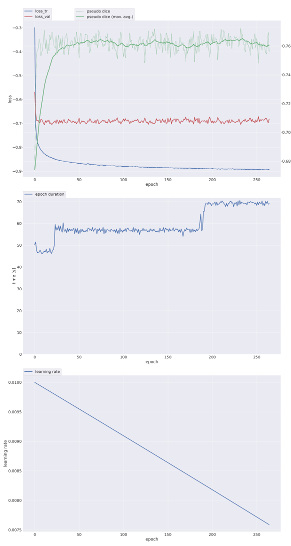

Figure 3: 2d training with 250 epoch

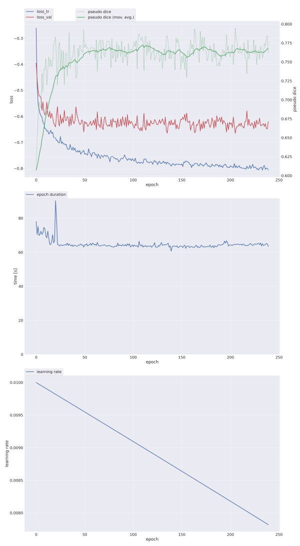

Figure 4: 3d_fullres training with 250 epoch

3.3 Training

nnU-Net employs 5-fold cross-validation to train all U-Net configurations, enabling it to ascertain post-processing techniques and ensemble strategies for the training dataset. This methodology enhances the generalizability and robustness of the model by averaging the performance across different subsets, there by mitigating overfitting and ensuring that the model performs well on unseen data.

We use the nnUNet_find_best_configuration to find the best configuration of U-Net with 2d and 3d_fullres. Running the prediction command and obtain the prediction results in the infer fold. Then we can make evaluation based on the prediction result.

4 Evaluation

4.1 Evaluation Metrics

Performance metrics included Dice Score, Sensitivity and Specificity. These were calculated to assess segmentation accuracy and precision.

4.2 Baseline Comparison

Dataset	**whole**	**core**	**enh.**
BraTS 2017 Train	0.895	0.828	0.707
BraTS 2017 Val	0.896	0.797	0.732
subset_100_2d	0.834	0.557	0.790
subset_100_3d_fullres	0.790	0.551	0.741

Table 1: Training and Validation Dice Scores [1]

Dataset	**Sensitivity**		
	whole	**core**	**enh.**
BraTS 2017 Train	0.890	0.831	0.800
BraTS 2017 Val	0.896	0.781	0.790
subset_100_2d	0.892	0.802	0.812
subset_100_3d_fullres	0.872	0.813	0.803

Table 2: Training and Validation Sensitivity[1]

Dataset	**Specificity**		
	whole	**core**	**enh.**
BraTS 2017 Train	0.995	0.997	0.998
BraTS 2017 Val	0.996	0.999	0.998
subset_100_2d	0.997	0.998	0.999
subset_100_3d_fullres	0.998	0.999	0.999

Table 3: Training and Validation Specificity[1]

5 Analysis

In our study, the BraTS 2017 dataset was utilized for training and validation to assess the performance of our model across different subsets. Specifically, the BraTS 2017 training and validation sets were compared with two subsets consisting of 100 cases each. These subsets were processed using 2D U-Net and 3D Full-resolution U-Net models to explore the impact of model architecture on performance.

The Dice scores indicate that the model performs comparably on the BraTS 2017 training and validation sets, and generally better than the two subsets. This may suggest a higher homogeneity in the standard datasets compared to the random subsets. Differences in Sensitivity across the datasets are minimal, but in terms of Specificity, the 3D fullres model slightly outperforms the 2D model on subsets, demonstrating the potential advantages of 3D models in handling complex structures.

6 Conclusion

In this study, we evaluated the nnU-Net framework on the BraTS 2017 dataset, confirming its robustness and efficiency in medical image segmentation. The results showcase nnU-Net's capacity to achieve competitive Dice scores and high specificity, particularly when using 3D configurations that better handle complex anatomical details. While the framework demonstrates strong baseline performance, further refinement is suggested to enhance its applicability across varied data subsets, aiming for broader generalization in clinical applications.

References

[1] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In *Brainlession: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, Lecture Notes in Computer Science, pages 287–297. Springer, 2018.

[2] Fabian Isensee, Jens Petersen, Andre Klein, David Zabala-Blanco, Paul F. Jaeger, Simon Kohl, and Klaus H. Maier-Hein. nnu-net: Self-adapting framework for u-net-based medical image segmentation. *arXiv preprint arXiv:1809.10486*, 2018.

[3] M. Islam and H. Ren. Multi-modal pixelnet for brain tumor segmentation. In *Lecture Notes in Computer Science, Brainlession: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 298–308. Springer, 2018.