Preliminary Results: Automated Glioma Segmentation using mixed-precision 3D UNet

Diedre Carmo, Letícia Rittner

***Abstract*—Segmentation of Glioma from three dimensional magnetic resonance imaging (MRI) is necessary for diagnosis and surgical treatment of patients with brain tumor. Manual segmentation is expensive, requiring medical specialists. In the recent years, the Brain Tumor Segmentation Challenge (BraTS) has been calling researchers to submit automated glioma segmentation methods for evaluation and discussion over their public, multimodality MRI dataset, with manual annotations. This paper presents a variation of the famous encoder-decoder architecture, UNet, employing modern deep learning techniques in its architecture and optimization, including the use of mixed-precision. Preliminary results are reported.**

***Index Terms*—deep learning, brain tumor segmentation, brats**

I. Introduction

ASSESSMENT of brain tumors is important in the diagnostic of Cancer [1]. Automatic segmentation can aid in this assessment, allowing for description of relevant tumor features such as its volume. However, tumors are very heterogeneous in shape, having different associated grades and classifications. Due to this variance, automatic segmentation of brain tumors is still a challenge [2].

A source of public glioma type brain tumors is the BraTS challenge [3]. This challenge expects high quality automatic segmentations of glioma regions, annotated over the provided four modalities of MRI, T1, contrast enhanced T1, T2 and FLAIR (see Figure 1). Submitted methods have to provide three segmentation maps: Whole Tumor (WT), Tumor Core (TC) and Enhancing Tumor (EC) (see Figure 1). The conception of the challenge came from the high inter-rater disagreement between expert raters in 2012 of up to 0.74 Dice [4]. Currently, most top-ranking methods in the challenge use Deep Learning [5] based methods.

Adaptation of the famous UNet [6] architecture is a common approach in recent years, with many succesful methods [7], [8], [9], [10], [11], [12].

[7] achieved top perfomance using an ensemble of four medical image segmentation CNN architectures, including 2 U-Net based ones, winning BRATS in 2017. The author proposes that its ensemble strategy aims to reduce the impact of different hyperparameters and bias employed to each architecture, by averaging their results.

Isensee Et al. adapted the UNet for 3D convolutions, with more skip connections, less channels, intensive augmentation, and a multi-class adaptation of DICE Loss. Interestingly, this is one of the leading methods from the 2017 [8] and the 2018 [9] challenge using mostly a single U-Net architecture, showing that a well trained U-Net can be superior to complex ensemble approaches. This work seems to have inspired a lot of the 2019 submissions, which used similar hyperparameters and attempt to use a modified 3D UNet.

The winner of the 2018 challenge also used an U-Net like architecture [10]. The main novelty of this work consisted of using a second branch in the decoder part of the architecture, reconstructing the original image as a means of regularization of the encoder. Another difference to basic U-Net is the use of a larger encoder, while most works keep the symmetry between encoder and decoder.

Myronenko in 2019 [11] explored variations in the traditional 3D encoder-decoder architecture, repeatedly used in BraTS. This work uses group normalization instead of batch normalization. A custom loss of 3 terms is used, optimized with Adam and progressive learning rate reduction over 300 epochs. Hyperparameters are in general similar to Isensee's work. The outputs consists of sigmoid nested tumor subregions. The main novelty is the exploration of the custom loss with dice, focal and Acl loss, which resulted in good relative performance to other submitted methods.

The winner of the 2019 challenge [12] used two UNets, one producing a coarse segmentation and other refining that segmentation, using multi stage loss applications. Labels learned are directly the overlapping regions, with a modified Dice Loss where DICE per region is simply added, using similar parameters in optimization to Isensee's work.

3D UNet like architectures have been achieving top performance in BraTS year after year. This paper proposes to further explore the potential of the UNet encoder-decoder architecture, in a mixed-precision ambient. Instructions and access to data, code and enviroments to reproducing this research can be found in https://github.com/MICLab-Unicamp/BTRSeg.

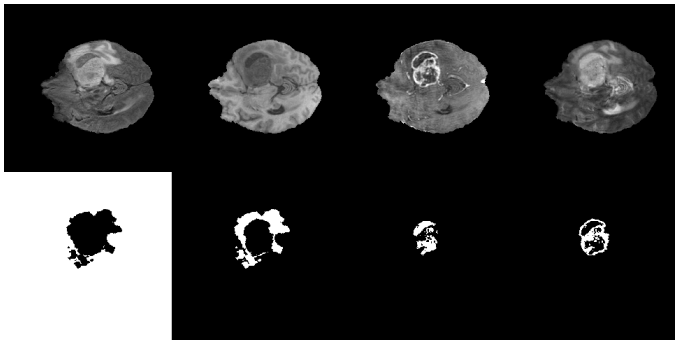

Fig. 1. The four modalities are showcased, in order: FLAIR, T1, T1 with Contrast and T2. Also displayed in the bottom row are manual annotations, in order: background, edema (ED), non-enhancing tumor (NET) and enhancing tumor (ET).

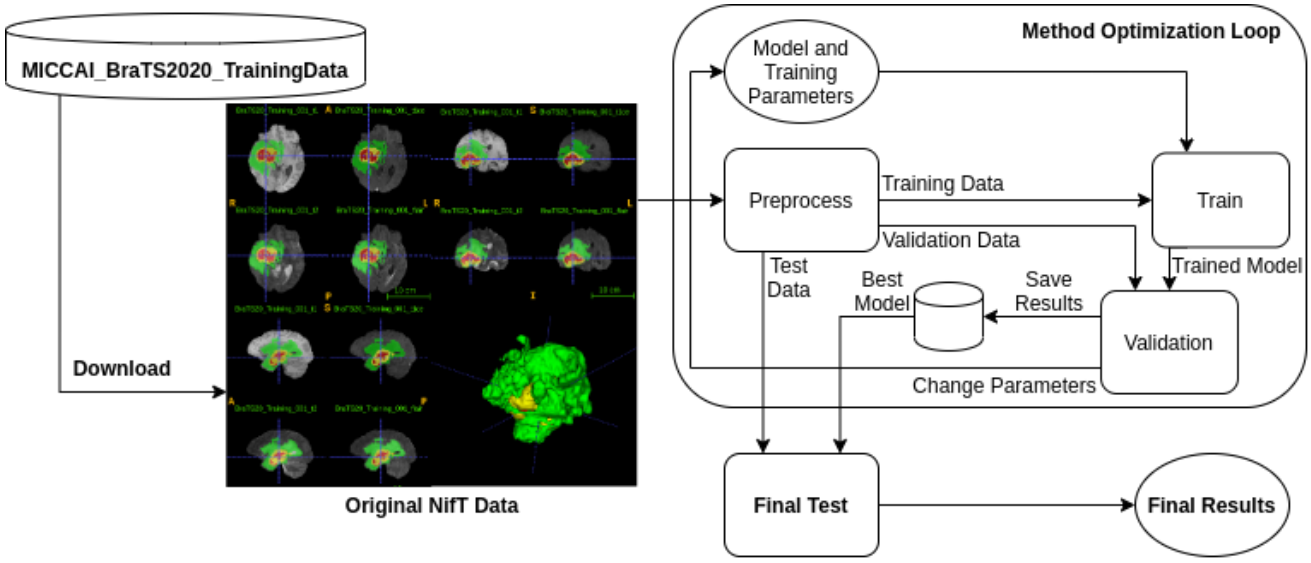

Fig. 2. Workflow illustration of this research. The core machine learning loop is illustrated by the repeated training and validation with different parameters over the training and validaiton sets. The best validation results are used in the final test over the test set.

II. Data

The BraTS 2020 dataset contains 369 MRI scans of various modalities: T1, post-contrast T1, T2, and FLAIR volumes (see Figure 1). All scans are of Low or High grade gliomas (LGG/HGG), acquired with different clinical protocols and various scanners from multiple institutions.

All Subjects have manual segmentations, performed by one to four raters, following the same protocol, with the resulting segmentation being approved by experienced neuro-radiologists. Annotations comprise the GD-enhancing tumor (ET), the peritumoral edema (ED), and the necrotic and non-enhancing tumor core (NET), as described in the latest BraTS summarizing paper [2]. Note that the challenge's evaluation is performed over three targets: the ET, the tumor core (TC) composed of ET + NET, and the Whole Tumor (WT) composed of ET + NET + ED. The provided data are distributed after pre-processing: co-registration to the same anatomical template, interpolation to the same resolution and skull-stripping.

Additional pre-processing applied in this methodology follows [8]. The images are subtracted by the mean and divided by the standard deviation of the brain region, and clipped inside the interval -5 to 5. Finally, they are min-max normalized to the interval 0 to 1. The targets are organized in a softmax manner, including background, resulting in four channels (background, ED, NET and ET). Experiments were also made with using sigmoid activations and optimizing directly over the three evaluation targets (WT, TC and ET).

This paper will split the challenge's (shuffled) training data in a hold-out approach of 70% training, 10% validation and 20% for the final test set, since the challenge validation and test sets are not available yet for 2020. This results in 251 training, 36 validation and 72 test subjects.

III. Methodology

This method, named BTRSeg, leverages modified, fully 3D, UNet like encoder-decoder CNNs. This architecture is based on previous experiments with hippocampus segmentation and 2D UNets [13]. We basically extended the E2DHipseg architecture to 3D convolutions, compensating the increased memory overhead with less channels. Instead of using the consensus of three networks, in this paper we purely explore the performance of one 3D architecture. The architecture still has residual connections [14] in the convolution blocks, however the input is a 4-channel volume containing the FLAIR, T1, T1ce and T2 volumes. In theory, batch normalization [15] may degrade training when using a small batch size. Therefore, experiments were performed in replacing batch normalization with group normalization [16].

In regards to training methodology, we employ similar hyperparameters to [8]. A learning rate of 0.0005, weight decay of 1e-5, max epochs of 300, exponential LR decay by 0.985. Adam [17] and RAdam [18] are experimented with for optimizers. The inputs are random patches of 4x128x128x128 for training, and center crops of 4x128x128x128 for validation. Group Normalization is also tested, however, it seems like its implementation requires more memory than Batch Normalization, which limits the possibility of increasing batch size. The used loss function consists of $1 - (WT_{dice} + TC_{dice} + ET_{dice})/3$, basically 1 minus the mean of Dices for each BraTS evaluation target.

One constant problem in 3D CNNs is the high usage of memory, limiting the size of batch size. The main hypothesis we want to analyze in this work is that using mixed precision training might enable higher batch sizes and better performance. To enable this, we used the AMP library from NVidia, built-in the pytorch-lightning framework. More optimal use of memory is enabled by replacing some representations to 16 bit precision instead of 32 bit, in what is called mixed-precision. Parts of training that benefit from higher precision such as softmax and loss still use 32-bit float, but some basic operations use 16 bit such as add and multiply.

Method	Optimizer	Batch Size	Precision	Normalization	Val Loss	WT (DICE)	TC (DICE)	ET (DICE)
Isensee	Adam	2	full	Instance Norm	0.22	0.88	0.78	0.69
BTRSeg	Adam	2	full	Batch Norm	0.28	0.84	0.67	0.66
BTRSeg	Adam	6	mixed	Batch Norm	0.19	0.90	0.81	0.72
BTRSeg	Adam	3	mixed	Group Norm	0.18	0.89	0.84	0.73
BTRSeg	RAdam	2	mixed	Group Norm	**0.17**	0.91	0.85	0.74
BTRSeg	RAdam	3	mixed	Group Norm	0.18	0.91	0.85	0.71

TABLE I
EACH ROW IS ONE FULL TRAINING, WITH THE BEST VALIDATION LOSS. LOWER IS BETTER. BEST METHOD IN BOLD.

IV. EXPERIMENTS AND RESULTS

Table I shows that improving batch size with mixed precision and using group norm improved results in relation to batch normalization and batch size 2, with performance better than Isensee's method. Using RAdam seemed not to improve results. Figure 3 shows the progression of validation loss per epoch of training. All experiments were run in a Xeon E3-1220 v2 CPU with 32 GB of RAM and a 12 GB Nvidia GTX Titan X. Instructions and access to data, code and enviroments to reproducing this research can be found in https://github.com/MICLab-Unicamp/BTRSeg.

Test metric	Dice
Whole Tumor	0.91
Tumor Core	0.84
Enhancing Tumor	0.81

TABLE II
THE BEST VALIDATION MODEL PERFORMED EVEN BETTER IN THE TEST SET, SHOWING THAT THERE IS NO OVERFITTING. THE TEST LOSS WAS 0.15.

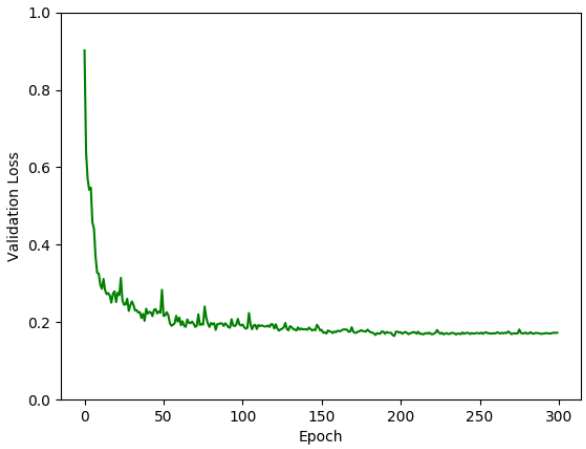

Fig. 3. Validation loss over epochs for the best model.

Figure 3 plots the validation loss over epochs for the best model so far (bold in Table I). The best validation method was selected to be exposed to test data, with final test results being displayed on Table II.

V. DISCUSSION

Batch normalization, as shown by other works, does not deal well with low batch size. Group normalization improved results, however, the current official PyTorch implementation of group normalization uses more memory than batch norm, which limited our possibilities of increasing batch size even more. Although increasing batch sized improved Batch Normalization results, whether increasing batch size with mixed precision improves results or not is not clear yet, needing more investigation. The bigger impact on our architecture performance came from switching to Group Normalization. Final test results show that the model is not overfitting to the validation set.

VI. CONCLUSION

As expected the use o Group Normalization improved results, however the impact of using mixed precision and high batch sizes is still not clear, requiring more investigation on the matter.

Future work intends to explore more variations on the architecture and hyperparameters, in an attempt to achieve similar or better performance than the top BraTS 2019 methods. Note that the validation and test data for 2020's challenge hasn't been release yet, which will help this research with more training data.

ACKNOWLEDGMENT

The author would like to thank FAPESP (2018/00186-0) for funding this research. This work was produced as a deliverable for the IA369Z discipline at UNICAMP.

REFERENCES

[1] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Scientific Data*, vol. 4, p. 170117, 2017.

[2] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki *et al.*, "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge," *arXiv preprint arXiv:1811.02629*, 2018.

[3] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE transactions on medical imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[4] C. H. Sudre, W. Li, T. Vercauteren, S. Ourselin, and M. Jorge Cardoso, "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, ser. Lecture Notes in Computer Science, M. J. Cardoso, T. Arbel, G. Carneiro, T. Syeda-Mahmood, J. M. R. Tavares, M. Moradi, A. Bradley, H. Greenspan, J. P. Papa, A. Madabhushi, J. C. Nascimento, J. S. Cardoso, V. Belagiannis, and Z. Lu, Eds. Cham: Springer International Publishing, 2017, pp. 240–248.

[5] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, number: 7553 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/nature14539

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, ser. Lecture Notes in Computer Science, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds. Cham: Springer International Publishing, 2015, pp. 234–241.

[7] K. Kamnitsas, W. Bai, E. Ferrante, S. McDonagh, M. Sinclair, N. Pawlowski, M. Rajchl, M. Lee, B. Kainz, D. Rueckert, and B. Glocker, "Ensembles of Multiple Models and Architectures for Robust Brain Tumour Segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi, S. Bakas, H. Kuijf, B. Menze, and M. Reyes, Eds. Cham: Springer International Publishing, 2018, pp. 450–462.

[8] F. Isensee, P. Kickingereder, W. Wick, M. Bendszus, and K. H. Maier-Hein, "Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge," *arXiv:1802.10508 [cs]*, Feb. 2018, arXiv: 1802.10508. [Online]. Available: http://arxiv.org/abs/1802.10508

[9] ——, "No New-Net," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 234–244.

[10] A. Myronenko, "3D MRI Brain Tumor Segmentation Using Autoencoder Regularization," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham: Springer International Publishing, 2019, pp. 311–320.

[11] A. Myronenko and A. Hatamizadeh, "Robust Semantic Segmentation of Brain Tumor Regions from 3D MRIs," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, ser. Lecture Notes in Computer Science, A. Crimi and S. Bakas, Eds. Cham: Springer International Publishing, 2020, pp. 82–89.

[12] Z. Jiang, C. Ding, M. Liu, and D. Tao, "Two-Stage Cascaded U-Net: 1st Place Solution to BraTS Challenge 2019 Segmentation Task," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, A. Crimi and S. Bakas, Eds. Cham: Springer International Publishing, 2020, vol. 11992, pp. 231–241, series Title: Lecture Notes in Computer Science. [Online]. Available: http://link.springer.com/10.1007/978-3-030-46640-4_22

[13] D. Carmo, B. Silva, C. Yasuda, L. Rittner, and R. Lotufo, "Hippocampus Segmentation on Epilepsy and Alzheimer's Disease Studies with Multiple Convolutional Neural Networks," *arXiv:2001.05058 [cs, eess]*, Jan. 2020, arXiv: 2001.05058. [Online]. Available: http://arxiv.org/abs/2001.05058

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: http://arxiv.org/abs/1512.03385

[15] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *arXiv:1502.03167 [cs]*, Mar. 2015, arXiv: 1502.03167. [Online]. Available: http://arxiv.org/abs/1502.03167

[16] Y. Wu and K. He, "Group normalization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[17] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[18] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han, "On the Variance of the Adaptive Learning Rate and Beyond," *arXiv:1908.03265 [cs, stat]*, Apr. 2020, arXiv: 1908.03265. [Online]. Available: http://arxiv.org/abs/1908.03265