

Curriculum Learning for Brain Tumor Segmentation

K Venkat Ramnan¹, Karthik Seemakurthy², and Uthayakumar G³

¹ PES University, Bengaluru, India
venkatramnank@pesu.pes.edu

² LAR/LIAT, University of Lincoln, United Kingdom.
karthikvjit@gmail.com

³ St. Joseph Institute of Technology, Chennai, India
guthayakumar@gmail.com

Abstract. Training the deep networks on the huge MRI volumes is a memory intensive task. This imposes an upper cap on the batchsize as well as the number of variations that can be attempted for competitions like BraTS 2021 challenge. In this paper, we have proposed an curriculum learning based approach which can reduce the training time by almost 30% without any significant compromise on the performance. In this technique, we define a schedule based on the class imbalance for every data point before starting to train the deep network. Also, we obtain the psuedolabels for the given validation set and retrain the model to improve the performance. Our curriculum learning based approach helped us to attempt multiple variations within the duration of competition.

Keywords: Curriculum Learning · Psuedo Labelling · Segmentation.

1 Introduction

Brain tumor segmentation is one of the most memory intensive task in the field of medical imaging due to the huge MRI volume for every data sample. This puts an upper limit on the number of samples that can be used in every batch of data. In this kind of scenarios which are memory intensive, it is very important to bring in optimization at every step of the entire training process. Especially, in the challenges like Brats 2021 [5] [8] [2] [3] [4] where there is a limited time, effective optimizations at every step of the training the deep networks play a very crucial role in testing wide variety of techniques in short duration of time without compromising on the performance. In this paper, we would present a curriculum learning based technique which has the potential ability to reduce the training time without having to compromise on the performance. Here not every sample is used in every epoch. Curriculum learning is one such approach which ranks each data sample and propose a schedule for loading the data samples in a specific order of importance. In this paper, we propose a novel schedule for each data sample based on the class imbalance and reduce the training time of the network by almost 30% without any significant compromise on the performance. This helped us to do more experiments with the architecture we have.

The next step is the choice of architecture. From network architecture point of view, we have used the Brats 2020 winning solutions and picked up the best architecture based on 3D UNET. We have used the architecture as it is and have not performed any major optimizations.

The main contributions of this paper are as follows

- Based on the class distribution, we propose a novel curriculum learning strategy for performing the task of brain tumor segmentation.
- We also show that our strategy can reduce the training time by almost 30%, specifically for the brain tumor segmentation task in BraTS 2021 challenge.
- We show that psuedolabels for segmentation task can improve the overall performance.

In the next sections, we describe about each of our contributions in detail and then followed by the supported experimental results.

2 Proposed approach

The flow of approach is given in Fig. 1. Initially we define a schedule for the entire training data by using class imbalance as a criterion. To our knowledge, using class imbalance as a criterion for defining a curriculum to all the data points is the first of its kind. We explain our curriculum learning approach in detail in Sec. 2.1. Once the schedule is defined for each data sample, we train the UNet model used in previous year challenge [1]. We used the 3D UNet architecture as it is without any changes whose details are given in Fig. 2. It is primarily composed of 3D convolutions, upsampling and downsampling operations. All the four mpMRI scans (T1, T1Gd, T2, T2-FLAIR) corresponding to each data point are stacked and fed as an input to the UNet architecture as shown in Fig. 2 and a three channel segmentation map is obtained as an output. For the first 10 epochs, we trained our model using the top 30% most difficult examples and then we pick the top 60% for 10 more epochs. In the last phase of 10 epochs, we pick all the training samples to train our model. This way we save almost 30% of training time without significant loss in the performance as will be explained in Sec. 3. BCEDice is used as a loss function to train the model. Once the model is trained, we obtain the psuedolabels of the validation set which is further combined with training set to retrain the model. The final segmentation map is obtained by the model which is trained by using the new augmented dataset. In the next subsections, we further describe the details of each of the steps in detail.

2.1 Curriculum learning

Let $X = \{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$ be the training set where each $\mathbf{X}_i \in \mathbb{R}^{C \times H \times W}$ is the i^{th} MRI volume, C is the total number of channels, H and W is the height and width of each channel. Every MRI volume \mathbf{X}_i is associated with a segmentation map $\mathbf{y}_i \in \mathbb{R}^{C \times H \times W}$ where each element value is the label indicating the nature of tumor at that specific location of brain. As per the annotation mentioned

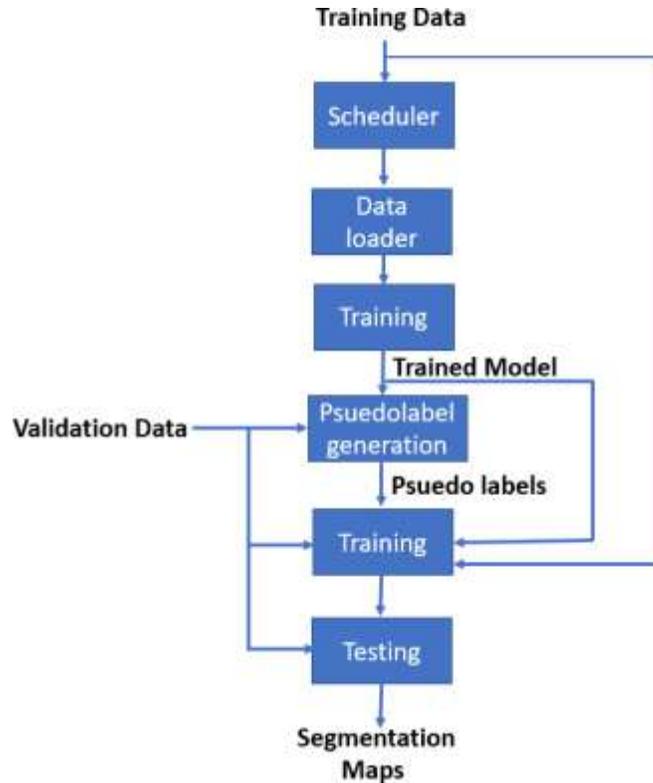


Fig. 1. Flow diagram of proposed approach.

in [2], each pixel can take the labels zero, one, two, four indicating no tumor, necrotic tumor core, peritumoral edematous/inavded tissue, and Gd-enhancing tumor, respectively. In BraTS 2021 dataset, there are a total of 155 channels associated with each MRI volume, where each channel has a resolution of 240×240 . Due to the huge memory occupied by each MRI volume, it is a time intensive task to train the state-of-the-art deep networks with BraTS 2021 dataset.

For a long time, random sampling of data samples was considered as the best strategy for training the deep neural networks to avoid the dataset bias. With the advent of curriculum learning [6], it was proven that a relevant data sampling strategy can significantly reduce the training time as well as have the better convergence. In this paper, we propose a novel curriculum learning strategy based on the class label distribution and show that we can achieve the comparable accuracy as that of using a full training set in much lesser time.

Class imbalance is considered to be the difficult task in the machine learning literature due to the bias it can induce into the models favoring the class label which has got higher frequency of occurrence. Let σ_i be the standard deviation

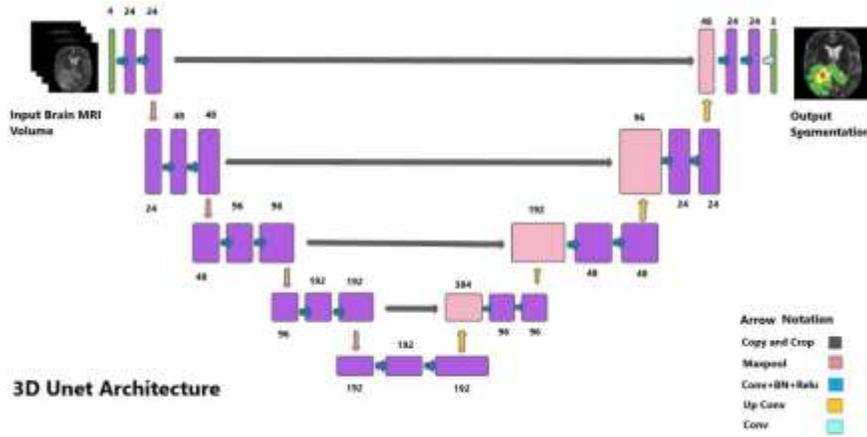


Fig. 2. Architecture used for proposed approach [1].

corresponding to the class distribution in the i^{th} MRI volume. The higher the value of σ_i , the difficult is the task of segmentation. We define a schedule for each data point in such a way that the deep network is trained by using the MRI volumes which got the higher class label standard deviation in the first few epochs followed by the rest in the later epochs. The motivation for this choice is that when we learn difficult examples first, we will be able to handle the simple tasks more confidently and accurately while the reverse might not be true.

3 Experimental Results

3.1 Impact of curriculum learning

In this subsection, we describe the details of an experiment which demonstrates the use of curriculum learning in reducing the training time without significant reduce in the accuracy. For this experiment, we initially split the complete data given in BraTS 2021 Challenge into 80/20 split for training and validation, respectively. The model is initially trained by using the complete training set and then followed by using the proposed curriculum learning strategy by assigning a schedule for every data point in the training split. Fig. 3 depicts the variation of validation loss, dice coefficient, and Jaccard coefficient with respect to number of epochs. By following the curriculum learning strategy it can be clearly seen that the validation loss, Dice coefficient, Jaccard coefficient converges well with the full data validation curves in 40 epochs. But, **by using the proposed curriculum learning strategy, we could save 30% of training time** which is very crucial in competitions like BraTS 2021 where there is a need to try lot of variations to top the leaderboard.

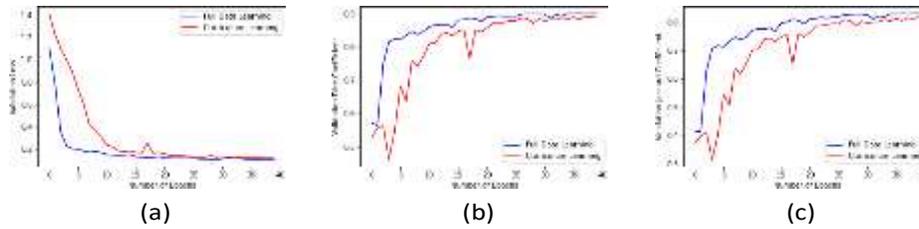


Fig. 3. Impact of Curriculum Learning. (a) Validation Loss, (b) DICE Coefficient, and (c) Jaccard Coefficient

Also, Table 1 shows the performance metrics obtained by submitting the segmentation outputs to the server at the end of 40 epochs. It can clearly be seen that the performance by using curriculum learning is quite close to the one which is obtained by using the full data learning.

Metrics	Full Data Learning	Curriculum Learning
DICE ET	0.7517	0.7410
DICE TC	0.8036	0.7919
DICE WT	0.9125	0.9087
Hausdorff95 ET	26.8026	33.2490
Hausdorff95 TC	11.5999	12.2628
Hausdorff95 WT	4.1710	4.3580
Sensitivity ET	0.7641	0.7626
Sensitivity TC	0.8054	0.8092
Sensitivity WT	0.9230	0.9205
Specificity ET	0.9997	0.9996
Specificity TC	0.9996	0.9995
Specificity WT	0.9991	0.9991

Table 1. Performance comparison between Full Data Learning and proposed Curriculum Learning.

3.2 Pseudo Labeling

In this section, we describe the details of an experiment which signifies the importance of pseudo labels in improving the accuracy of brain tumor segmentation task. The concept of pseudo labeling was proven to be quite successful in the field of medical image segmentation [7][9]. Once we train the model with the available training data, using the evaluation mode in pytorch, we obtain the pseudo labels for the validation set. These pseudo labels are augmented with original training data and the model is retrained with a reduced learning rate. By including the pseudo labels, it can be seen from Table ?? that there is a

significant improvement in the performance of our approach in majority of the metrics. One of the key reason behind the poor performance with respect to few metrics (Hausdorff95 ET, Sensitivity WT, Specificity ET, Specificity TC) might be due to the noisy psuedo labels.

Metrics	Curriculum Learning (Without psuedolabels)	Curriculum Learning (with psuedolabels)
DICE ET	0.7132	0.7451
DICE TC	0.7551	0.7784
DICE WT	0.8727	0.9038
Hausdorff95 ET	29.0983	31.7926
Hausdorff95 TC	23.0086	18.1825
Hausdorff95 WT	8.2637	5.6970
Sensitivity ET	0.6667	0.7730
Sensitivity TC	0.7231	0.7865
Sensitivity WT	0.9402	0.9229
Specificity ET	0.9997	0.9996
Specificity TC	0.9997	0.9995
Specificity WT	0.9983	0.9989

Table 2. Performance improvement with psuedo labels.

3.3 Five fold cross validation

The entire training data initially given in the first phase is split into train and validation sets. Even though through validation sets, we have an option to select the best training model, but this best model has not been trained on the portion of given training data. This leads to poor generalization ability of our trained model. In order to overcome this limitation, we have performed five fold cross validation and trained five different models by using the flow diagram described in Fig. 1 for each fold. Table. 3 reports the results obtained by using the five folds in the first five columns. We take the maximum voting approach and fuse the results obtained by all the five training models and is indicated in last column of Tab. 3. It can be seen that in majority of metrics, the fused results takes an upper hand and this emphasize the need for cross validation.

4 Conclusion

In this paper, we proposed a novel curriculum learning strategy based approach to train for the task of brain tumor segmentation in the BraTS 2021 challenge. We observed that learning the model with the samples ordered in decreasing order of difficulty leads to a solution closer to the one obtained by training the model with the full data. We have fused the outputs obtained by using the

Metrics	Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	Fusion
DICE ET	0.7330	0.7558	0.7451	0.7465	0.7447	0.7632
DICE TC	0.7603	0.7920	0.7784	0.7680	0.8049	0.7877
DICE WT	0.9062	0.9072	0.9038	0.9078	0.9103	0.9113
Hausdorff95 ET	31.6701	26.5285	31.7926	26.805	31.7403	26.2822
Hausdorff95 TC	14.568	15.5412	18.1825	20.5855	10.2154	15.1033
Hausdorff95 WT	4.9272	4.5095	5.6970	4.3570	4.4058	4.4020
Sensitivity ET	0.7185	0.7771	0.7730	0.7491	0.7747	0.7741
Sensitivity TC	0.7626	0.7923	0.7865	0.7535	0.84227	0.7843
Sensitivity WT	0.9183	0.9288	0.9229	0.9193	0.9079	0.9196
Specificity ET	0.9997	0.9996	0.9996	0.9996	0.9996	0.9997
Specificity TC	0.9996	0.9996	0.9995	0.9997	0.9994	0.9997
Specificity WT	0.9991	0.9989	0.9989	0.9991	0.9993	0.9992

Table 3. Performance metrics across multiple folds and fusion.

models trained using five fold cross validation by using a maximum vote criterion. However, it would be an interesting research direction to use a meta learning based approach to assign a weight to each of the trained models obtained in five folds. Also, it would be very useful to identify a better schedules other than class imbalance and which is more closer to human judgements in identifying hard examples. Through our curriculum learning strategy we could show that we can reduce the training time by almost 30%.

References

1. <https://www.kaggle.com/polomarco/brats20-3dunet-3dautoencoder>
2. Baid, U., Ghodasara, S., Bilello, M., Mohan, S., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., Prevedello, L.M., Rudie, J.D., Sako, C., Shinohara, R.T., Bergquist, T., Chai, R., Eddy, J., Elliott, J., Reade, W., Schaffter, T., Yu, T., Zheng, J., Annotators, B., Davatzikos, C., Mongan, J., Hess, C., Cha, S., Villanueva-Meyer, J., Freymann, J.B., Kirby, J.S., Wiestler, B., Crivellaro, P., Colen, R.R., Kotrotsou, A., Marcus, D., Milchenko, M., Nazeri, A., Fathallah-Shaykh, H., Wiest, R., Jakab, A., Weber, M.A., Mahajan, A., Menze, B., Flanders, A.E., Bakas, S.: The rsna-asnr-miccai brats 2021 benchmark on brain tumor segmentation and radiogenomic classification (2021)
3. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. Nat Sci Data **4**, 170117 (2017)
4. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J., Freymann, J., Farahani, K., Davatzikos, C.: Segmentation labels and radiomic features for the pre-operative scans of the tcga-lgg collection. The cancer imaging archive **286** (2017)
5. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 1–13 (2017)

6. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. pp. 41–48 (2009)
7. Li, Y., Chen, J., Xie, X., Ma, K., Zheng, Y.: Self-loop uncertainty: A novel pseudo-label for semi-supervised medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 614–623. Springer (2020)
8. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**(10), 1993–2024 (2014)
9. Wang, L., Guo, D., Wang, G., Zhang, S.: Annotation-efficient learning for medical image segmentation based on noisy pseudo labels and adversarial learning. *IEEE Transactions on Medical Imaging* pp. 1–1 (2020). <https://doi.org/10.1109/TMI.2020.3047807>