Scheduling Cross Entropy and Dice Loss for Optimal Training of Segmentation Models

Chaitra V. Hegde*
Center for Data Science
New York University
New York, NY 10011
cvh255@nyu.edu

Aakash R. Kaku*
Center for Data Science
New York University
New York, NY 10011
ark576@nyu.edu

Sohae Chung
Department of Radiology
New York University School of Medicine
New York, NY 10016
Sohae.Chung@nyulangone.org

Xiuyuan Wang
Department of Radiology
New York University School of Medicine
New York, NY 10016
Xiuyuan.Wang@nyulangone.org

Yvonne W. Lui
Department of Radiology
New York University School of Medicine
New York, NY 10016
Yvonne.Lui@nyulangone.org

Narges Razavian
Department of Radiology
New York University School of Medicine
New York, NY 10016
Narges.Razavian@nyulangone.org

Abstract

Brain MRI Segmentation is a challenging task partly due to severe class imbalance and large number of segments. Previous work tackled the class imbalance issue by using weighted cross entropy or weighted dice loss. In this work, we show that only using a fixed loss function for the entire training period is not an optimal strategy, and propose a novel and simple scheduling method for loss function optimization, that leads to more robust and optimal segmentation model. Using this technique, we show that a standard U-Net architecture is able to surpass the more sophisticated state-of-the-art QuickNAT architecture when tested on MICCAI Multi-Atlas Labeling challenge data set under the similar conditions. We also compare our results to the widely used tool, Freesurfer, and show that our method provides systematically superior results.

1 Introduction

Anatomical brain segmentation is an important task for almost all the neuroimaging analysis. Traditional and widely used software, Freesurfer [1], require hours [8] to perform the segmentation task for each brain scan, which in turn restricts its usability in the clinic. Similarly, non-deep learning based models like STAPLE[12] and PICSL[11] also takes hours to perform the inference task. Therefore, deep learning based models can be used to achieve massive performance gains in terms of speed and accuracy. The current state-of-the-art deep learning model for the brain MRI segmentation is QuickNAT[8], which combines results of three views (axial, sagittal and coronal) to perform segmentation on brain volume. QuickNAT is a U-Net [7] style architecture with four dense blocks

*Equal contribution

Preprint. Work in progress.

used for encoding and four dense blocks used for decoding. The encoding and decoding blocks are connected using skip connections.

All the existing deep learning models for segmentation use a fixed loss function while training their network. We showcase that using different loss functions at different stage of training can lead the model to a better generalization performance. We term this procedure of training the network as loss scheduling.

2 Scheduling of Weighted Cross Entropy and Weighted Dice Loss

In segmentation task, the dice score is often the metric of importance. A loss function that directly correlates with the dice score is the weighted dice loss. But often the network trained with only weighted dice loss gets stuck in a local optima and doesn't converge at all. In addition, empirically it is seen that the stability of model in terms of convergence decreases as the number of classes increases. An alternative loss function is weighted cross-entropy (w-cel), with the drawback that it is sensitive to the class weights and often suffers from a problem of over prediction for some classes and under prediction for other classes [2]. Therefore, a linear combination of the two loss functions is often considered as the best practice [8] [10][9].

Intuitively, the combination of loss function is still sub-optimal as there is a component of w-cel loss which is very sensitive to the class weights which are usually calculated heuristically[2]. Therefore, we propose to gradually change the loss function from w-cel to w-dice loss as the training advances. This will ensure that once the model reaches a local optima using the w-cel, the w-dice loss can help the model to reach a better minima which helps to maximize the dice score. This hypothesis was validated using a U-Net[7] which was trained on MRIs of 20 healthy patients from Human Connectome Project (HCP)[4]. For this stage, the Freesurfer segmentation results processed by HCP were used as auxiliary ground truth.

3 Experiments and Results

3.1 Datasets

Two datasets were used during the experiment, Human Connectome Project Dataset [4] and MICCAI Multi-Atlas Labeling challenge data set [6]. The MRI volumes from first data set are 3D MPRAGE images acquired at multiple sites using 3T Connectome Skyra scanners (FOV = 224mm x 224mm, resolution = 0.7mm isotropic, TR/TE = 2400/2.14 ms, bandwidth = 210 Hz/pixel). Each MRI volume has a dimension of $256 \times 256 \times 256$. For training the U-Net, a 2-D axial slice of the MRI volume is used as the input. Freesurfer segmentation results processed by HCP were used as auxiliary ground truth for pre-training the models. MICCAI data set contains T1 MRI from 30 healthy patients with 15 patients in the train set and 15 patients in the testing set. Each MRI volume is manually segmented, and is of the dimension $256 \times 256 \times \geq 256$. Last dimension is along the sagittal direction. We use the training set to fine-tune the pre-trained models.

3.2 Implementation Details

The segmentation model architecture is a standard U-Net [7] as described in the original paper with only difference being the input image size is 256×256 and output image size being $256 \times 256 \times 28$ (number of classes + one for None). All the parameters of the U-net were initialized using Xavier initialization[5].

For all the models, the class weights were calculated based on median frequency balancing method[3]. The model that was trained using only the w-dice Loss did not converge. As seen in Figure 1, the model reached a better optima after switching from a combination of w-cel and w-dice loss to pure w-dice loss. We also confirmed the performance gain was significant by testing our trained model on MICCAI Multi-Atlas Labeling challenge test set[6]. As it can be seen in Table 1, the pre-trained performance of the model trained using loss scheduling is better than all other models including the state-of-art QuickNAT[2] model which has much more complex architecture as compared to the vanilla

[2]U-Net is compared to the pretrained performance of QuickNAT because the U-Net and QuickNAT both are trained on MRI of healthy patients with Freesurfer labels as the ground truth and tested on MICCAI test data set

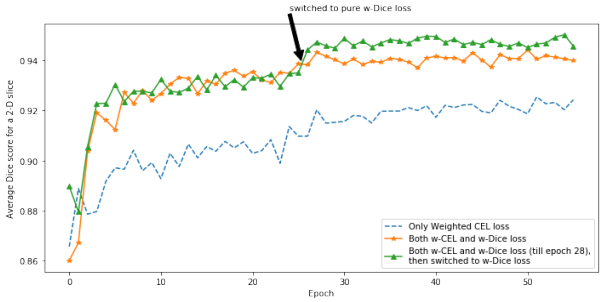

Figure 1: Plot showing average dice vs epoch for different training procedures

Table 1: Performance on MICCAI test data. Pre-Trained refers to results when training on auxiliary (Freesurfer) labels, and Fine-Tuned refers to results of fine-tuned model on MICCAI training set.

Models			
Name	Loss Function	Pre-Trained	Fine-Tuned
U-Net	Fixed (w-Dice Loss)	Did not converge	
U-Net	Fixed (w-Cel)	0.7602 ± 0.085	
U-Net	Fixed (w-Cel + w-dice loss)	0.7819 ± 0.072	
U-Net	Loss scheduling	$\mathbf{0.8049 \pm 0.067}$	0.885 ± 0.042
QuickNAT	Fixed (w-Cel + dice loss + Boundary Loss)	0.798 ± 0.097	$\mathbf{0.901 \pm 0.045}$
U-Net	Fixed (w-Cel + dice loss + Boundary Loss)	0.681 ± 0.193	0.857 ± 0.079

U-Net. Moreover, QuickNAT was pre-trained on 581 healthy patients with Freesurfer segmentation as the auxiliary ground truth while the U-Net model was pretrained on only 20 healthy patients. Hence, it can be seen that an appropriate loss function and training methodology is as important as the architecture of the model.

4 Systematic Bias in Freesurfer Labels

For the purpose of understanding the systematic biases of Freesurfer and also to improve model's performance on the test data, we fine tuned the U-Net model using the MICCAI training set. Although the performance of the fine-tuned U-Net model (trained using loss scheduling) is not as good as fine-tuned QuickNAT, it performs better than the fine-tuned U-Net model trained with fixed loss function.

As seen in Table 1, the pre-trained U-Net model which was trained using Freesurfer labels doesn't perform as well as the model which was fine-tuned using the manual segmentation labels. This indicates an inherent bias in how the Freesurfer segments the MRI. Hence, we were interested to know the segments for which the difference of performance between pre-trained and fine-tuned model is significant. As it can be seen from Figure 2, there is a considerable difference in the

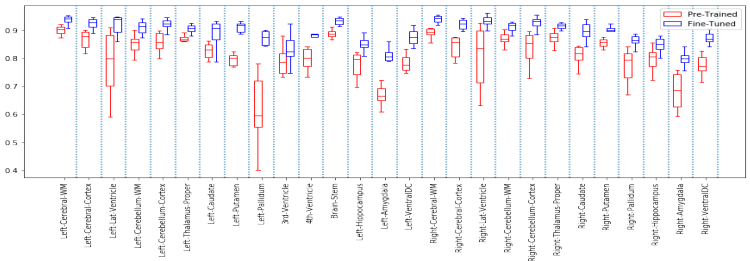

Figure 2: Box-Plot showing the difference in the Dice scores of Pre-trained U-Net (Red) and Fine-Tuned U-Net (Blue) for all the segments on the test data set of MICCAI challenge

performance for all the segments. But, particularly for segments like Left-Pallidum, Left-Amygdala, Left-Lat-Ventricle and Right-Lat-Ventricle, the difference in the performance is very large suggesting that Freesurfer has some systematic biases while labelling the above-mentioned segments.

5 Conclusion

In this work, we show that with an appropriate loss function and a training methodology, a simple model architecture can outperform the state-of-the-art architecture for segmenting the imbalanced and diverse segments under similar conditions. The simple and yet effective techniques demonstrated in this study can be used for training of any segmentation model. We also showed that a wildly used tool, Freesurfer, has systematic biases for some segments and our method can have superior performance which overcomes the systematic biases of Freesurfer labels. As future work, we plan to improve the performance on larger manually segmented data, open source and release our tool, and evaluate impact of loss scheduling on other applications and other publicly available data sets.

References

[1] Rahul S. Desikan, Florent Ségonne, Bruce Fischl, Brian T. Quinn, Bradford C. Dickerson, Deborah Blacker, Randy L. Buckner, Anders M. Dale, R. Paul Maguire, Bradley T. Hyman, Marilyn S. Albert, and Ronald J. Killiany. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, 31(3):968 – 980, 2006.

[2] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *CoRR*, abs/1804.10851, 2018.

[3] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. *CoRR*, abs/1411.4734, 2014.

[4] Matthew F. Glasser, Stamatios N. Sotiropoulos, J. Anthony Wilson, Timothy S. Coalson, Bruce Fischl, Jesper L. Andersson, Junqian Xu, Saad Jbabdi, Matthew Webster, Jonathan R. Polimeni, David C. Van Essen, and Mark Jenkinson. The minimal preprocessing pipelines for the human connectome project. *NeuroImage*, 80:105–124, oct 2013.

[5] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS'10). Society for Artificial Intelligence and Statistics*, 2010.

[6] Bennett A. Landman and Simon K. Warfield. Miccai 2012 workshop on multi-atlas labeling (volume 2). *CreateSpace Independent Publishing Platform(EDS)*, 2012.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[8] Abhijit Guha Roy, Sailesh Conjeti, Nassir Navab, and Christian Wachinger. Quicknat: Segmenting MRI neuroanatomy in 20 seconds. *CoRR*, abs/1801.04161, 2018.

[9] Chen Shen, Holger R. Roth, Hirohisa Oda, Masahiro Oda, Yuichiro Hayashi, Kazunari Misawa, and Kensaku Mori. On the influence of dice loss function in multi-class organ segmentation of abdominal CT using 3d fully convolutional networks. *CoRR*, abs/1801.05912, 2018.

[10] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sébastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. *CoRR*, abs/1707.03237, 2017.

[11] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, March 2013.

[12] Simon K Warfield, Kelly H Zou, and William M Wells. Simultaneous truth and performance level estimation (staple): an algorithm for the validation of image segmentation. *IEEE transactions on medical imaging*, 23(7):903–921, 2004.