Segmentation of gliomas in MRI scans

Shailja
Electrical & Computer Engineering
University of California, Santa Barbara
shailja@ucsb.edu

***Abstract*—Segmenting different regions in glioma is a challenging task. Current *state-of-art method* uses UNet and DeepMedic architecture and attains the maximum accuracy of about 88%. In this paper, we propose a method that adds an average ensembling layer on top of the existing neural network model which has the best known performance. We show that this improves the accuracy of tumor segmentation to 90%. Also, we show results from different methods by varying the ensembling layer and its classifiers that suggest an improvement in the accuracy. We show that leveraging the uniqueness of each model is an important step towards building predictive models. We use the BraTS2018 dataset [1]- [4] for all our experiments.**

I. Introduction

Glioma is a common type of brain tumors that originates in the glial cells that surround and support neurons. It has different degrees of aggressiveness, variable prognosis and various heterogeneous histological sub-regions. This intrinsic heterogeneity of gliomas is also portrayed in their appearance and shape, as their sub-regions are described by varying intensity profiles disseminated across multi-modal MRI scans, reflecting varying tumor biological properties. Quantitative evaluations of *state-of-the-art* tumor segmentation algorithms revealed considerable disagreement between the human raters in segmenting various tumor sub-regions (dice scores [1] in the range $74\% - 85\%$), illustrating the difficulty of this task. We found that different algorithms worked best for different sub-regions (reaching performance comparable to human inter-rater variability), but that no single algorithm ranked in the top for all sub-regions simultaneously.

II. Literature Review

There is a significant amount of recent work on brain tumor segmentation and survival prediction. In [5], seven different 3D neural network models with different parameters are integerated to obtain the final brain tumor mask. Moreover, a hierarchical pipeline to segment the different types of tumor compartments using anisotropic convolutional neural networks was designed in [6]. The network architecture in [7] is derived from a 3D U-Net with additional residual connections on context pathway and additional multi-scale aggregation on localization pathways, using the Dice loss in the training phase to circumvent class imbalance. In this paper, we have used a methodology to integrate multiple DeepMedics [9] and patch-based 3D U-Nets [10] with different parameters and different training strategies in order to get a robust brain tumor segmentation from multi-modal structural MR images. This segmentation approach is based on the recent work in [8]. We have used the results of the probability map from [8] to construct our approach. In Table I, we have presented the accuracy results of different methods measured using a metric called dice score for final voxel level classification.

Models	WT	TC	ET
DeepMedic_ce_4	0.8847	0.7813	0.7057
DeepMedic_ce_25	0.8868	0.7953	0.7177
DeepMedic_ce_24	0.8845	0.7854	0.7184
U3DNet_dice_c4	0.8719	0.7715	0.6912
U3DNet_dice_c25	0.8757	0.7817	0.6965
U3DNet_ce_c4	0.8847	0.7775	0.6857

TABLE I
Performance (dice score) of robust brain tumor segmentation

III. Objective

The objective of the project is to produce segmentation labels of the different glioma sub-regions. The sub-regions considered are: 1) the "enhancing tumor" (ET), 2) the "tumor core" (TC), and 3) the "whole tumor" (WT) [Figure 1].

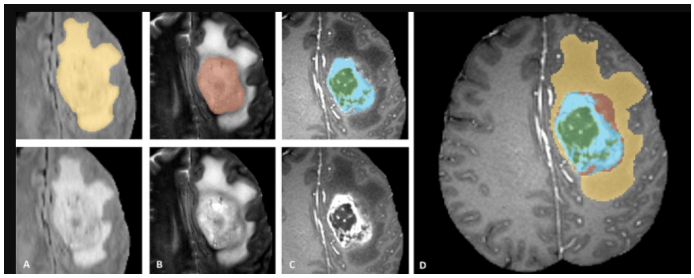

Fig. 1. Manual annotation through expert raters. Shown are image patches with the tumor structures that are annotated in the different modalities (top left) and the final labels for the whole dataset (right). Image patches show from left to right: the whole tumor visible in FLAIR (a), the tumor core visible in t2 (b), the enhancing tumor structures visible in t1c (blue), surrounding the cystic/necrotic components of the core (green) (c). Segmentations are combined to generate the final labels of the tumor structures (d): edema (yellow), non-enhancing solid core (red), necrotic/cystic core (green), enhancing core(blue).

[1]Intersection over union metric to evaluate similarity between images.

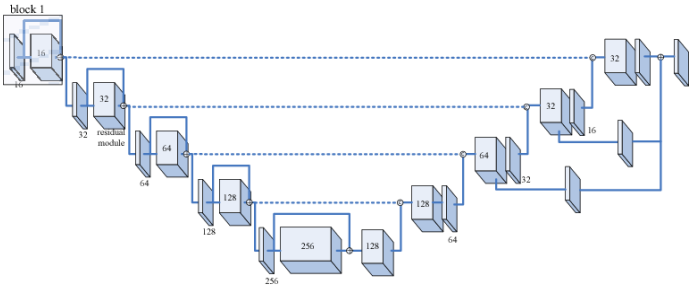

Fig. 2. 3D UNet Architecture

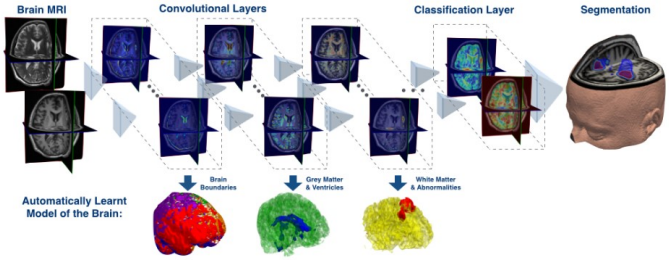

Fig. 3. DeepMedic Architecture

IV. Experiments

We performed the following experiments with the aim to improve the accuracy of segmentation. The description of the experiments along with the results are given in the following subsections.

A. Average Ensemble

We used the probability maps obtained after applying robust brain tumor segmentation in [8] for this experiment. The final segmentation mask of the brain tumor is calculated by taking the average of the output probability maps from each model in our ensemble. Table II shows the comparison of the results obtained from UNet and DeepMedic models with Average Ensemble of five DeepMedic and five UNet models.

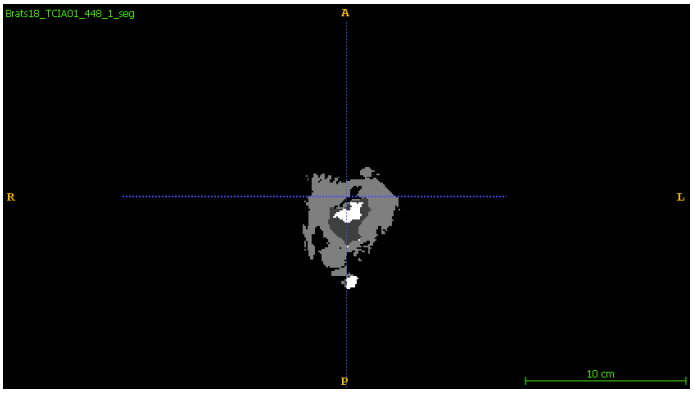

Fig. 4. Output from Average Ensemble, labels are background(0), edema(1), necrosis(2) or tumor(4)

Models	WT	TC	ET
DeepMedic	0.88	0.79	0.71
UNet	0.87	0.78	0.69
Average Ensemble	0.90	0.81	0.73

TABLE II
Average Ensemble

B. XGB, LGBM and Random Forest Last Layer

Average ensembling method performs better and improves the accuracy. It also proves that there is uniqueness among different models which can be exploited even more. With this in mind, we added the last layer of classifier to our integrated model of one DeepMedic and one UNet model but it performed poorly. One possible reason could be that multi-class classification performed poorly in general. However, to circumvent this class imbalance, we classified the whole tumor region. We used the light GBM and XGB framework that uses tree-based algorithms for faster training speed and higher efficiency for this experiment. The results show that this method improves the accuracy by about 0.1%. Results from this experiment are summarized in Table III.

Fig. 5. Inputs from one DeepMedic and one UNet model to the classifier layer

Classifier	Dice score improvement (%)	Training speed	Subjects
LGBM	0.089	1	180
Random Forest	0.008	3	10
XGB	0.061	2	30

TABLE III
Improvement of Dice score (w.r.t. average ensemble) after addition of classifier layer

C. Training with balanced data

Since the classifier layer is using tree-based algorithms, we could improve the results by modifying the training data to form a balanced tree. For this purpose, we utilized the analysis of the Average Ensemble result. According to the Table IV, we took different value of x and eliminated $x\%$ of the true negative[2] voxels to form a relatively more balanced tree. Table VI shows the results for different values of x. In contrast to our intuition, this performs poorly because our model is trained to detect foreground when the background has $x\%$ less background but we are testing on an image with a large percentage of background voxels. This results in

[2]A true positive test result is one that detects the tumor. A true negative test result is one that detects background.

poor performance in detecting the background which decreses the accuracy as is evident from the results shown for this experiment.

True positive (1.6%)	False negative (0.15%)
False positive (0.05%)	True negative (98.2%)

TABLE IV
ANALYSIS OF AVERAGE ENSEMBLE RESULT

True positive (100%)	False negative (100%)
False positive (100%)	True negative (x%)

TABLE V
TRAINING DATA BASED ON AVERAGE ENSEMBLING

Classifier	x=82%	x=61%	x=50%	x=20%
LGBM	0.50	0.63	-	0.89
Random Forest	0.85	-	-	-
XGB	0.72	0.86	0.88	-

TABLE VI
FOR 200 SUBJECTS, X% TRUE NEGATIVE ELIMINATION

D. StackNet Ensembling

StackNet is a computational, scalable and analytical, meta-modeling framework implemented in Python that resembles a feedforward neural network and uses Wolperts stacked generalization on multiple levels to improve accuracy in machine learning predictive problems. In this paper, we experimented with two stacked layers of lgbm and xgb in the second last layer and random forest in the last layer [Figure 6]. We trained for 25 subjects on five DeepMedic and five UNet models to achieve the accuracy of 85%. Due to limitation in time and resources, the experiment is still in progress.

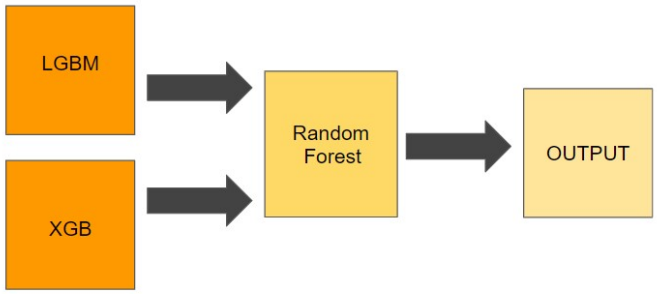

Fig. 6. StackNet models

V. CONCLUSION

Our results indicate that the algorithm in [8] can be enhanced to reach Dice scores of over 90% for whole tumor segmentation. Based on the experiments performed, we can conclude that our approach of average ensembling results in a more robust tumor segmentation. Our work also showed that there is slight improvement in dice score of about 0.1% w.r.t average method after addition of lgbm and xgb classifiers. In addition to pushing the limits of individual tumor segmentation algorithms, we can infer from our results that future gains may also be obtained by investigating how to implement and fuse several different algorithms using by ensembling strategies.

VI. FUTURE WORK

Many interesting future research directions are possible in line with the work presented in this paper. To improve the accuracy of tumor segmentation for brain MRI images, it would be interesting to explore the integration of algorithms to fuse different known brain tumor segmentation methods. For example majority vote, stacking, ensembling, meta-learning and other fusion strategies could be used together to improve the performance. Furthermore, a useful next step in the method that we have used would be to fix the percentage of background voxels eliminated to get the highest dice score coefficient for the test subjects. We expect that using those parameters would improve the accuracy of segmentation. We can also experiment further with different stacknet models to improve the results.

ACKNOWLEDGEMENTS

I would like to thank Po-Yu Kao for the many insightful discussions and help with the approach & experiments. I would also like to thank the teaching assistants of ECE 194N and Prof. B.S. Manjunath for their guidance in this course project.

REFERENCES

[1] Menze BH et al. "The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS)", IEEE Transactions on Medical Imaging 34(10), 1993-2024 (2015) DOI: 10.1109/TMI.2014.2377694
[2] Bakas S et al. "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features", Nature Scientific Data, 4:170117 (2017) DOI: 10.1038/sdata.2017.117
[3] Bakas S et al. "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-GBM collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.KLXWJJ1Q
[4] Bakas S et al. "Segmentation Labels and Radiomic Features for the Pre-operative Scans of the TCGA-LGG collection", The Cancer Imaging Archive, 2017. DOI: 10.7937/K9/TCIA.2017.GJQ7R0EF
[5] Kamnitsas, Konstantinos, et al. "Ensembles of multiple models and architectures for robust brain tumour segmentation." International MICCAI Brainlesion Workshop. Springer, Cham, 2017.
[6] Wang, Guotai, et al. "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks." International MICCAI Brainlesion Workshop. Springer, Cham, 2017.
[7] Isensee, Fabian, et al. "Brain tumor segmentation and radiomics survival prediction: contribution to the BRATS 2017 challenge." International MICCAI Brainlesion Workshop. Springer, Cham, 2017.
[8] Kao, Po-Yu, et al. "Brain tumor segmentation and tractographic feature extraction from structural mr images for overall survival prediction." International MICCAI Brainlesion Workshop. Springer, Cham, 2018.
[9] Kamnitsas, Konstantinos, et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation." Medical image analysis 36 (2017): 61-78.
[10] Cicek, Ozgun, et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2016.