CS230

Deep Learning Segmentation of High and Low Grade Gliomas in 3D Brain Tumor Scans with U-Net Ensemble Methods

Abigail VanderPloeg*
Department of Computer Science
Stanford University
abigailv@stanford.edu

Abstract

At the Stanford Cancer Center, radiologists currently rely principally upon by-hand methods in order to analyze MRI cranial scans of patients with brain tumors. With the ongoing development of the application of deep learning to the task of image segmentation, these methods can be applied to brain tumor imaging data in order to expedite the process of identifying and analyzing patient tumors. These brain tumors (the most common of which is called a glioma) can be of an enhancing, high grade type (HGG), which form a clearer outline in tumor imaging, or a non-enhancing, low grade type (LGG). This paper proposes an ensemble model architecture which combines the output predictions of sub-models trained specifically to segment HGG and LGG tumor images in order to segment on new tumor image data. The paper found that the best-performing ensemble method to segment brain tumor images pulled from the benchmark Brain Tumor Segmentation (BraTS) dataset [1] utilizes a a simplified U-Net model architecture (with only one CNN layer per each "depth level" of the contracting then expanding computations) and a max confidence ensemble layer to transform the output predictions of the sub-models

1 Introduction

Gliomas are the most common type of brain tumor, and can vary greatly in their aggressiveness and shape. Neuro-radiologists at the Stanford Cancer Center currently rely upon hand-labeling in order to determine the location of and differentiate between the sub-regions of a glioma. Thus, the ability to locate and segment gliomas within 3D medical scans with computational models is valuable in order to aid in oncologists' diagnoses as well as to predict patient survival outcomes. Within gliomas, there is a distinction between high grade, rapidly growing gliomas, and low grade, slow-growing gliomas, which are non-enhancing (do not show a clear outline) in imaging [2]. A principle challenge is to train a model that achieves high performance when segmenting both enhacing and non-enhacing glioma scan regions. Because low-grade gliomas are non-enhacing (do not form a clear outline in imaging scans), it is more difficult for segmentation models to distinguish between the tumor area and the surrounding healthy brain tissue [2]. Comparatively, the enhancing tumor core of high grade gliomas can be more easily distinguished against regions of healthy brain tissue. The challenge of creating accurate segmentations for both types of glioma is proliferated within traditional models that are trained to segment both enhancing (high grade) and non-enhancing (low grade) gliomas.

*In collaboration with Dr. Hakura Itakura in the Stanford Cancer Center.

CS230: Deep Learning, Spring 2021, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

Models that have garnered top performance on tumor segmentation benchmarks have shown to be disproportionately well-suited to segment only particular sub-regions of the glioma (such as the non-enhancing core, or enhancing core) rather than exhibiting top performance across segmentations of all sub-regions [3]. This papers aims to implement and improve upon high-performing deep learning segmentation models to train, using 3D brain tumor MRI scans, an ensemble model constructed from U-Net submodels for use on Stanford Cancer Center brain scans. The specific goal of the model is to effectively segment both enhancing and non-enhacing regions of gliomas through using ensemble methods.

2 Related Work

The benchmark Brain Tumor Segmentation (BraTS) Dataset [1], published by the University of Pennsylvania Center for Biomedical Image Computer and Analytics (later discussed in the Dataset section) issues an annual challenge that surfaces the highest-performing models on the data. One of the highest-performing neural network model architectures for tumor segmentation is a U-Net model architecture [4]. This U-Net architecture includes first contracting CNN layers to capture context (important features such as boundaries between glioma regions), then expanding CNN layers in order to localize predictions to the entire image scan. Recently in 2020, Feng et. al. tested the use of ensemble methods constructed across multiple U-Net models, and showed that the combined predictions outperformed the predictions of a single U-Net model. [5] However, because the sub-models the group created are trained to fit to both enhancing and non-enhancing tumors concurrently, each sub-model cannot specifically focus on segmenting a specific region of the glioma.

3 Dataset and Features

The (Brain Tumor Image Segmentation Benchmark) BraTS dataset is a collection of 65 pre-operative MR brain scans from patients with both low and high-grade gliomas [1]. It includes manual segmentations (produced by one to four raters) that annotate the sub-regions of the glioma: the necrosis, GD-enhancing tumor (not present for low-grade gliomas), the peritumoral edema, and the non-enhancing tumor core. The images are composite 3-dimensional visualizations of the tumor, and the images are divided into four subfile dataset: T1, T2, Flair, and T1Ce sequences (all of which are different cross-sectionals of the brain and are composed of differnt volumes). Each volume has 155 slices, and 210 volumes are of high grade gliomas while 75 are of low grade. For the purpose of our segmentation, we will focus on training models to segment on the T1Ce (constrast-enhancing) scans. The radiologist on staff with our Research team states that this scan (which will enhance the contrast of the tumor against the remaining section of the brain) is the most informative.

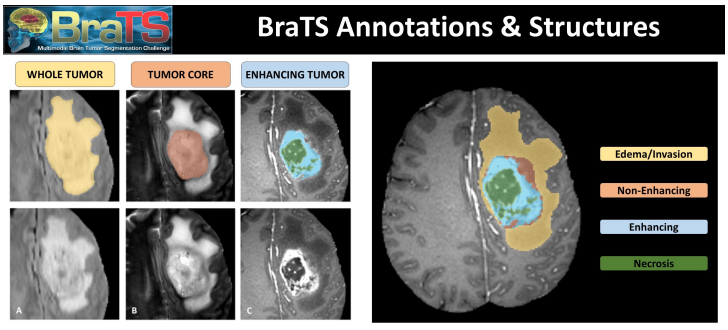

Figure 1: An example predicted and ground truth segmentation of regions of both a high grade and low grade glioma. The low grade glioma does not have a red enhancing tumor core as the high grade glioma does. [1]

4 Methods

Because previous models have not shown to be high-performing across all sub-regions of gliomas (both enhancing and non-enhancing) [3], one promising method is to combine different models that are trained separately to be well-suited to segment a specific sub-region of a glioma. Ensemble methods is a technique used to combine the predictions from distinct models into a (hopefully) optimal prediction [6]. The first step in establishing a deep learning ensemble approach for the image segmentation was to first create individual models that could predict well on specifically non-enhancing (low grade) or enhancing (high grade) image scans.

As discussed, models employing a U-Net model architecture [4] have exhibited high performance on the task of brain tumor segmentation. Thus, for the sub-models within the larger ensemble algorithm, I went forward with implementing U-Net architecture models. The model architecture outlined by Isenee et. al [4] and a slightly modified model architecture are shown below in Figure 2. During training, the original model architecture was shown to have high validation loss, and the modified model architecture, which simplifies each "step" of the model in order to reduce complexity, is able to achieve lower validation loss. Thus, this modified model architecture is used to generate the LGG and HGG-specific sub-models in the final ensemble model. In the training of each model a Dice score loss is utilized (as done by the official BraTS dataset challenge), which compares the areas of the model's segmentation and the expert-annotated segmentation, and normalizes the number of true positives (per voxel, or 3D pixel) to the average size of the segment areas.

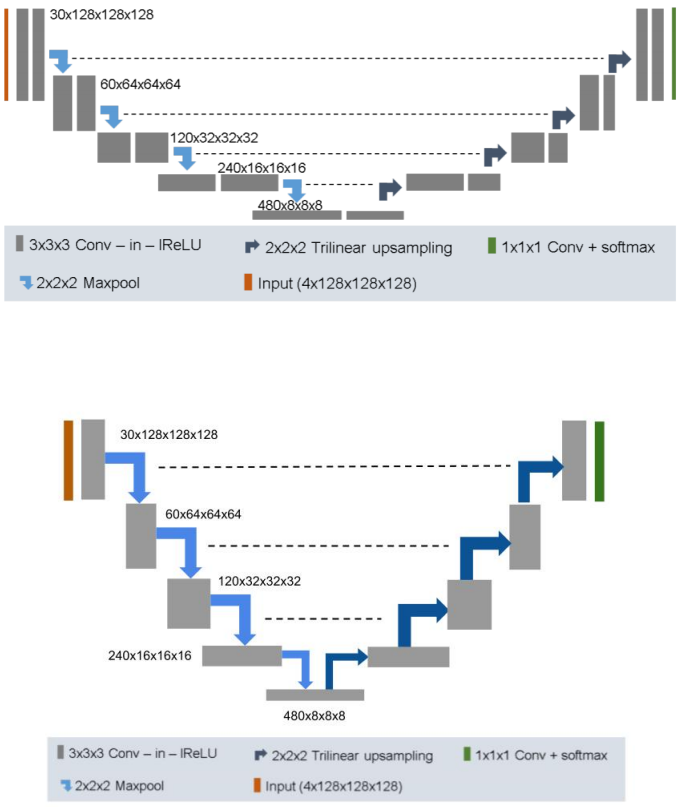

Figure 2: The model architecture diagrams for different iterations of the the sub-model U-Nets. The right-hand version, which reduces complexity and overfitting by "decreasing the width" of each depth level of the U-Net architecture, is utilized in the final ensemble model.

In order to create the layer of the ensemble model which will combine the output predictions of the U-Net sub-models into final predictions, there were three separate methods tested: an weighted averaging of the output predictions, a "maximum-confidence" prediction, and a final method which

implements a 3D CNN binary classifier in order to determine which sub-model to feed the test sample to for prediction.

The first two methods utilize relatively straightforward transformations across the output predictions of each sub-model. While the weighted average of the output predictions provides a good baseline for the, the maximum confidence prediction (which takes the highest-probability prediction from among the sub-model predictions) is able to selectively utilize different submodels where they are most relevant. This is especially applicable when selecting a prediction between sub-models that are trained to perform well on specific sub-regions of a tumor.

For the binary classification of a new image samples as a HGG or LGG tumor, a 3D CNN model architecture [7] is trained and predicts on images of standardized sizes from the original MRI scans (shrinking from a $255 \times 255 \times 155$ size to a $128 \times 128 \times 64$ size in pre-processing). The model architecture diagram for the 3D CNN is displayed in Figure 3. The training and validation of the binary classification CNN utilizes a binary cross-entropy loss. Given ground truth class labels y and predicted class labels $p(y)$ over N examples, the logarithm of the binary cross-entropy loss is calculated as $-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(p(y_i)) + (1 - y_i) \cdot \log(1 - p(y_i))$.

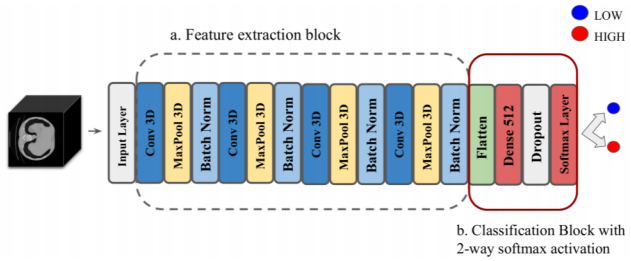

Figure 3: The model architecture diagram for the 3D CNN binary classifier, implemented as described by Zunair, et. al [7].

5 Experiments/Results/Discussion

The first experiments related to the choice of the sub-model architecture to train separately on the LGG and HGG tumor data. The graphs in Figure 4 demonstrate the training and validation dice score loss progressions of the two U-Net model architecture variations displayed in Figure 2 on both the LGG and HGG data. All models utilize the following hyperparameters: samples = 210, a (X, Y, Z) patch size of (80, 80, 16), and a batch size of 2 over 30 epochs. The optimizer used is an Adam optimizer with a learning rate of 0.0001.

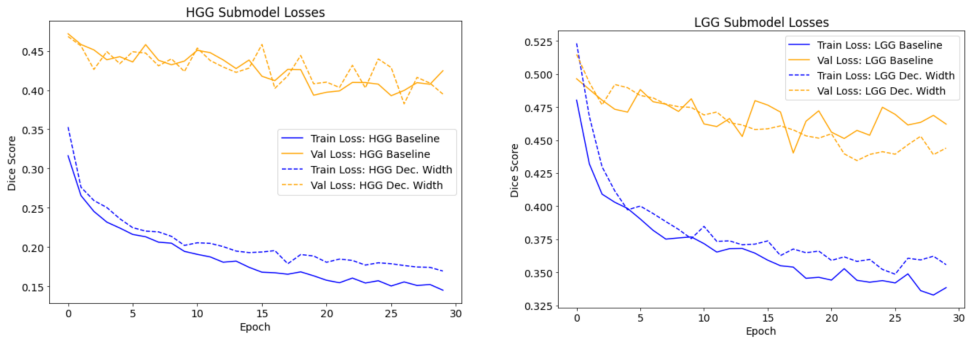

Figure 4: Training loss curves for U-Net submodels trained on LGG (non-enhancing) and HGG (enhancing) tumor cores. The LGG loss may be generally higher due to the lack of a clear out-line or enhacing region, thus making the tumor generally more difficult to segment.

The "decreased width", less complex model architecture is able to, despite higher training loss, achieve lower validation loss on the LGG submodels and slightly lower validation loss on the HGG sub-models. Thus, for the final LGG and HGG sub-models, the models constructed with this "decreased width" modification a U-NET model architecture are utilized.

Upon the selection of the optimal sub-models, the next experiments required analysis of various methods that could be utilized to combined the prediction results of the LGG and HGG-specific submodels. The validation dice score results from the different methods of transformations on sub-model output predictions are shown below. They are shown alongside a final validation loss on a version of the U-Net submodel trained to classify on both LGG and HGG data, which ended training (on the same hyperparameter set as the U-Net submodels). The baseline weighted average method uses an equal weighting of both sub-models and the max-confidence method requires no hyperparameters. The 3D CNN classifier (which then feeds the segmentation class to the respective sub-model) utilizes a binary cross-entropy loss, a learning rate of 0.0001, and the Adam optimizer over 100 epochs.

Validation (Dice Score) Loss on Ensemble Methods and Baseline Model			
Weighted Average	Max Confidence	3D CNN Classifier	Baseline Combined Model
0.44	**0.42**	0.47	0.45

These results are promising when compared to the results obtain from the baseline predictions with a version of the U-Net submodel trained to classify on both LGG and HGG data. The 3D CNN Classifier continued to overfit to predict the more common HGG label even after hyperparameter tuning, and got a final validation loss closer to that of the HGG-specific sub-model as a result. Ultimately the max confidence ensemble layer provided the best estimates.

6 Conclusion and Future Work

In conclusion, the paper found that the best-performing ensemble method to segment brain tumor images pulled from the benchmark Brain Tumor Segmentation (BraTS) dataset [1] utilizes a a simplified U-Net model architecture (with only one CNN layer per each "depth level" of the contracting then expanding computations) and a max confidence ensemble layer to transform the output predictions of the sub-models. This model framework was able to get very close results to a model trained on both LGG and HGG data concurrently, and shows promise especially given future work that can be done to both refine the U-Net submodels through hyperparameter tuning and further architecture modifications and testing of different ensemble methods, especially a tuned version of the 3D CNN Classifier. Ultimately, upon showing the results to the radiologist in contact with our research team, they had mentioned that, although the loss scores remain somewhat high on certain samples, the overall results were informative and promising.

7 Contributions

As I worked on the project individually, I completed all model implementations and experiments myself. However, this would not have been possible without the mentorship of Dr. Itakura, who provided invaluable guidance on the nature of the problem being solved and our specific research goals.

8 Code

The following code files for this project can be found in the code section of my Gradescope Project Milestone submission:

- `unet_model.ipynb`
  This file generates a U-Net Architecture sub-model to be fed into an ensemble model. The user can specify the type of tumor regions which the model should be trained on (HGG LGG, both).

- `classifier_model.ipynb`
  This file generates a 3D CNN Architecture sub-model that can predict, for a given brain tumor scan, whether the glioma is low-grade (LGG) or low grad (HGG).
- `loss_plots.ipynb`
  This file plots the training and validation loss from different sub-models for comparison.
- `ensemble.ipynb`
  This file loads in multiple pre-trained sub-models, and utilizes the individual output predictions to compute a final prediction. This can either be done by performing a transformation (max confidence, averaging) on output predictions across sub-models, or by letting the classifier model determine which submodel (LGG or HGG-specific) should be used to compute the output.

References

[1] B Menze, A Jakab, S Bauer, J Kalpathy-Cramer, K Farahani, J Kirby, et al. Multimodal brain tumor image segmentation. benchmark: change detection. *Proceedings of MICCAI-BRATS 2016*, 2016.

[2] Jeanine T Grier and Tracy Batchelor. Low-grade gliomas in adults. *The oncologist*, 11(6):681–693, 2006.

[3] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin 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.

[4] Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H Maier-Hein. No new-net. In *International MICCAI Brainlesion Workshop*, pages 234–244. Springer, 2018.

[5] Xue Feng, Nicholas J Tustison, Sohil H Patel, and Craig H Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in computational neuroscience*, 14:25, 2020.

[6] Thomas G Dietterich. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*, pages 1–15. Springer, 2000.

[7] Hasib Zunair, Aimon Rahman, Nabeel Mohammed, and Joseph Paul Cohen. Uniformizing techniques to process ct scans with 3d cnns for tuberculosis prediction. In *International Workshop on Predictive Intelligence In Medicine*, pages 156–168. Springer, 2020.