Brain Tumor Segmentation Using Enhanced 3D U-Net Model

AAI-521 Computer Vision - Group 14

by

Hani Jandali

University of San Diego

December 9th, 2024

Abstract

Glioma represents some of the most common and aggressive malignant primary tumors of the central nervous system present in adults. Forming from glial cells within the brain, their difficulty in diagnosis is compounded by their seemingly generic and common symptoms, such as headaches, cognitive changes, weakness in limbs, and vision or speech difficulties which only present in late stages of formation (National Brain Tumor Society, n.d.). Yet, even when symptoms present and MRIs are taken gliomas are misdiagnosed due to their extreme intrinsic heterogeneity in appearance, shape, histology, and inter-observer variability (National Brain Tumor Society, n.d.). As a result of poor methods in diagnosis and prognosis, 48% of all glioma cancers caught and recognized are likewise categorized as glioblastomas, or late stage glioma, requiring immediate and aggressive care through a simultaneous concoction of surgery, radiotherapy, and chemotherapy (Glioblastoma Research Organization, n.d.). However, even then outlook remains bleak, with a median survival period of 12 to 15 months after diagnosis with treatment, and a five year survival rate of only 6.9% (National Brain Tumor Society, n.d.).

Many of the obstacles that persist in diagnosing glioblastoma, particularly the characteristics of heterogeneity in appearance, shape, and histology, as well as inter-observer variability, provide a challenge to radiologists and neurologists alike, but are prime for deep learning models that leverage computer vision attributes. By utilizing the mpMRI images present in the 2021 BraTS challenge (Brain Tumor Segmentation Challenge, n.d.), we develop a U-Net architecture geared towards the classification, segmentation, and masking of glioblastoma in order to enable early detection and diagnosis. In the end, the results provide a proof of concept, returning an accuracy of 99.52%, F-Score of 79.08%, an IOU-Score of 73.80%, and loss value of 0.4524. Subsequently, the model demonstrates the potential of computer vision in areas of

critical care needs where normal methodologies currently employed are relatively subpar in comparison.

Dataset

The first challenge to approaching the issue of identifying and classifying glioma was data. Often, data within healthcare if too sensitive for public use and availability. However, the data composing the foundation of our approach originates from the 2021 BraTS challenge, a multi-institutional dataset composed of multi-parametric magnetic resonance imaging (mpMRI) from patients with late stage glioblastoma (Brain Tumor Segmentation Challenge, n.d.). Developed by the Radiological Society of North America (RSNA), in conjunction with the American Society of Neuroradiology (ASNR) and the Medical Image Computing and Computer Assisted Interventions (MICCAI) society, our truncated version of the data is composed of 1,253 MRI instances, each instance containing five types of MRIs broken into four modalities and a ground truth segmentation annotated by experts. The modalities define different details captured by means of MRI, such as T1 (or T1-weighted) providing anatomical reference, T1CE (or T1-contrast enhanced) highlighting and enhancing tumor regions, T2 (or T2-weighted) providing structural detail, and Fluid Attenuated Inversion Recovery (or FLAIR) which suppresses fluid signals during MRI capture to emphasize edema (Brain Tumor Segmentation Challenge, n.d. & Figure 1). Segmentation images likewise have multiple subclasses within each image, highlighting the necrotic core of dead tissue at the center of the tumor, swelling around the tumor, and actively growing portions of the tumor. The depth of information for each MRI, along with the standardized size of 240 by 240 by 155 voxels per MRI saved in a NIfTI format dataset totaling 13.4 GB (Schettler, D., 2021) with no missing datafiles across all instances.

Preprocessing

Considering the challenge has occurred annually since 2017, in tandem with the delicate, precise, and important nature of the data necessitating comprehensive scrutiny from the RSNA, ASNR, and MICCA societies, the data required little preprocessing. Across all modalities, there were an equal number of files numbering 1,253 instances with all MRI images previously standardized to a three dimensional 240 by 240 by 155 pixelation or voxel. Likewise, across all images present, there are is no additional background noise, unrelated images, objects, blurs, or shifts in the MRIs, with all similarly formatted in NIfTI formatting. As a result, preprocessing was simplified, requiring no manipulation of the data itself or creation of classes or subclasses within the data. Subsequently, the data was instead prepared for the model, with the T1, T1CE, T2, and FLAIR modalities stacked as channels along the last axis to form a single input of 240 by 240 by 155 by 4 across each instance. Considering the dimensionality reduction within the model across multiple blocks, the z-axis figure was padded to 160 to ensure smooth division by a factor of 2, resulting in a new instance size of 240 by 240 by 160 by 4.

With the refined file size per instance, we independently normalized each modality in our images to a range of [0,1] for consistent input in scaling, followed by a conversion of the segmentation masks into categorized arrays via one-hot encoding for future multi-class identification. In order to refine and better generalize our model to both validation and additional real world data, we augmented some files to simulate variations in scan orientations and image intensities while retaining the key structure of each MRI. Downstream, this aided our models robustness to variability and elevated its resulting metric scores.

Although the data was relatively clean and required little processing, the large nature of the dataset itself posed difficulty in ensuring the informational quality of all images across all instances. There remained the possibility of image captures without brain tissue present in the

instance, presenting a form of noise for our model that may result in negative adaptations in testing. To ensure the informational quality of all inputs, a filter was applied within the DataGenerator class to calculate the proportion of non-zero voxels in a volume relative to total volume of voxels in the image (Jandali, 2024). As a result, below a set threshold of 3.0%, low-quality or empty MRI images were removed and excluded from training and validation datasets. Therefore, the resulting images were ensured, after preprocessing, to be relevant to predicting segmentations in the brain and classes of cancerous tumors.

UNet Model

After preprocessing, the data was split into both a training and validation set, totaling 1,002 instances and 251 instances respectively. The model employed here is a variation on convolutional neural networks (CNN) developed precisely for computer vision, specifically three-dimensional image segmentation, called a 3D U-Net (Ronneberger, O, 2015). The choice behind the model was rather simple, as it provided the greatest ability to generate precise, voxel-level segmentation maps for three-dimensional images without a large resource constraint as does occur with a traditional CNN while also avoiding the feature abstraction pitfalls present in other tested models such as ResNet or YOLO (Ronneberger, O, 2015). Computational and resource strain was a consistent, motivating factor behind many choices within this project and somewhat alleviated by the efficient architecture of 3D U-Net models.

The deep learning model itself took in our processed input of 240 by 240 by 160 by 4 three-dimensional MRI images with stacked modalities and performed voxel-wise segmentation at each “slice” of the image. The architecture is simplified into three different portions: the encoder, itself composed of three blocks, the bottleneck block, and the decoder composed of an additional three blocks. The first layer of the initial block applied a 3D convolution with 32

kernels of 3 by 3 by 3 sizing before feeding into a batch normalization layer and continuing into the second convolution of the first block, in which another 32 kernel, 3 by 3 by 3 dimension convolution was applied to capture high-level features and abstract them for segmentation. The systematic pattern of the first block was replicated in the second and third blocks of the encoder, with each subsequent block downsampling by a factor of two (as reflected in the kernel values of 64 and 128 respectively). The fourth block of the U-Net model represented the "bottleneck" layer, or the deepest portion of the network, with 256 units per layer. It is at this stage in which the highest-level and deepest feature extraction occurs, in which our network captured abstract semantic features at the lowest resolution, providing the foundation for reconstructing the predicted segmentations during the decoder (upsampling) portion of the model. From there, the instance moves forward to the last three blocks composing the decoder, in which the sample is gradually upsampled and spatially reconstructed by the same factor as the encoder blocks. The final sample is then outputted by a 1 by 1 by 1 convolution reducing the feature maps to five channels (one channel per class prediction) with a softmax activation function outputting the likely class probabilities of each individual voxel. All layers utilized ReLU activation functions due to their computational efficiency and mitigation of vanishing gradients as explored in the *Constraints* section.

Overall, the model itself totaled at 5,608,069 parameters, providing the best balance between compromise and deep feature abstraction considering the complexity of the MRI images and limited resources and similar obstacles detailed in the *Constraints* section. Still, the model required approximately 3.75 hours to train one epoch at 38.1 GB of constant GPU usage (Figure 2), requiring a total of approximately 40 hours for the ten epoch model presented in our findings.

Constraints

As mentioned above, many constraints encountered across the project restricted the true potential of our model regardless of our satisfactory and promising results. The primary difficulty was resource allocation. Initial models developed to abstract segmentation patterns and features from the MRI instances resulted in U-Net models totaling over 370,000,000 trainable parameters, which likewise required intense GPU, RAM, and system storage demands. Initial reductions of the model reduced the trainable parameters to nearly 90,000,000, in which cloud computing with larger resource allocation was pursued. However, requests for Amazon high GPU EC2 instances and Google Cloud high GPU instances were rejected, and Lambda Labs and Paperspace cloud resources providing too little an increase in available resources compared to Google Cloud A100 High-Ram runtimes. However, a highly reduced U-Net model was required in order to train within the boundaries of available GPU resources.

In conjunction with a limited U-Net model, an effort was simultaneously made to reduce the NIftI files to .npy files to improve data loading, data handling, and utilize parallelization between loading and processing, albeit at the expense of potential loss of metadata. Unfortunately, the data transformation became a restraint in itself, devaluing important features within MRI scans that further reflected in poor metric and visual evaluations. An alternative experimentation was cropping the MRI images from 240 by 240 by 155 to 192 by 192 by 128 in order to reduce memory usage and improve computational speed. Given that the borders of the MRI data are simply black pixels of no value (Figure 1), elimination of the empty space would reduce memory and improve training times. However, in order to provide an appropriate, processible input for our model, the dimensions of the image needed to be divisible by 64, thus resulting in the aforementioned 192 by 192 by 128 dimensionality but also a partial cropping and cutting off of brains in some MRI scans. The tradeoff was productive in training, reducing the

size of the image from 35,712,000 voxels to 18,874,368 voxels, however still inadequate given the few hours still required per epoch. Likewise, it also presented with minor drops in intersection-over-union (IoU) and F-score metrics.

Further reduction to both image dimension and model complexity was attempted, resulting in a 388,000 trainable parameters model with MRI images of 128 by 128 by 128 by 4, with an achieved epoch time of fifty two minutes. However, after twenty epochs, the model proved too shallow and simple to abstract any useful information for our segmentation and classification goals, reflecting a steep drop-off in IoU and F-scores. The best compromise for feature abstraction with promising results was a relatively deep, 5,608,069 trainable parameter model on the original, uncropped MRI images, requiring a full utilization of GPU resources at approximately 3.75 hours per epoch in training. Training in multiple segmented batches across multiple days, we achieved ten epochs.

Apart from the model, an unexpected constraint was an initially vanishing gradient, for which we changed activation functions to “ReLU” to eliminate without severe computational expense. Subsequently, gradient explosion occurred while training, requiring batch normalization at each layer within the model to stabilize training. Further still, a set learning rate proved to be a limiting factor, with an adjustable reduceLROnPlateau function applied to reduce learning rate when validation loss stopped improving to a minimum rate of 1e-4. This method proved the most versatile and useful, as reflected below.

Metrics and Results

Considering the segmentation task, traditional metrics like accuracy were insufficient in translating the effectiveness of our model. Instead, we used a combination of Dice Loss and Categorical Crossentropy to measure loss epoch over epoch, reflecting the overlap between

predicted segmentation and ground truth segmentations as well as the voxel-wise classification accuracy of the segmentation (Jandali, H. 2024). Likewise, Intersection-over-Union (IoU) was used to evaluate the overlap between predicted segmentation and ground truth as a ratio of their intersection relative to their union to assess how accurately segmentation boundaries were drawn. F-Score was used to assess actual model performance at a threshold of 0.5. After ten epochs of training, the results were relatively promising, with a peak F-score of 79.08%, IoU score of 73.80%, and loss value of 0.4524, indicating that our model was reasonably good at identifying tumor regions without falsely misidentifying other regions of the brain as cancerous (Figure 3). Similarly, the IoU score reflects the same conclusion, demonstrating great accuracy at capturing tumor regions but room for improvement in refining the boundaries of the tumor itself. The loss value of 0.4524 indicates that the models predictions are relatively close to the ground truth segmentation with some slight room for improvement. These results were further validated through visual confirmation, with our model predicting the necrotic core of dead tissue at the center of the glioblastoma (red), swelling around the tumor (green), and actively growing portion of the glioma (yellow) (Figure 5).

Conclusion

Overall, the model was relatively successful. Despite the resource restraints limiting the depth and complexity of the U-Net model to abstract features from the MRI images, the scores remained promising and further reflect the serious potential of computer vision models predicting and isolating developing glioblastomas before onset symptoms present. In essence, we succesfully segmented and classified different types of glioma tumors from MRI images that may have otherwise been missed entirely or misdiagnosed between multiple doctors.

References

Aboud Glioblastoma. (n.d.). *Nationa Brain Tumor Society.* https://braintumor.org/events/glioblastoma-awareness-day/about-glioblastoma/

Brain Tumor Segmentation. (n.d.). *Brain tumor segmentation*. http://braintumorsegmentation.org/

Glioblastoma: Diagnosis and treatment. (June 20th, 2024). *Mayo Clinic.* https://www.mayoclinic.org/diseases-conditions/glioblastoma/diagnosis-treatment/drc-20569078

Jandali, H. "Brain-Tumor-Segmentation." *Github*, 2024. https://github.com/H-Jan/Brain-Tumor-Segmentation.git

Kihira, S., Mei, X., Mahmoudi, K., Liu, Z., Dogra, S., Belani, P., Nael, K. (2022). U-Net based segmentation and characterization of gliomas. *Cancers*, *14*(18), 4457. https://doi.org/10.3390/cancers14184457

Liyen, K., Drummond, K., Hunn, H., Williams, D., O'Brien, T., Monif, M. (2020). - Potential biomarkers and challenges in glioma diagnosis, therapy and prognosis. *BMJ Neurology Open.* https://doi.org/10.1136/bmjno-2020-000069

McKinnon, C. (2021). Glioblastoma: clinical presentation, diagnosis, and management. *BMJ, 364().* https://doi.org/10.1136/bmj.n1560

Rafi, A., Ali, J., Akram, T., Fiaz, K., Raza Shahid, A., Raza, B., & Mustafa Madni, T. (2020). U-Net based glioblastoma segmentation with patient's overall survival prediction. In *Intelligent Computing Systems: Third International Symposium, ISICS 2020, 3()*, 22-32. https://doi.org/10.1007/978-3-030-43364-2_3

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, 3*(18), 234-241. https://doi.org/10.48550/arXiv.1505.04597

Schettler, D. (2021). BraTS 2021 Task 1. *Kaggle.* https://www.kaggle.com/datasets/dschettler8845/brats-2021-task1/data

What is the Average Glioblastoma Survival Rate? (n.d.). *Glioblastoma Research Organization.* https://www.gbmresearch.org/blog/glioblastoma-survival-rate

Appendix

Figure 1: Multi-modal MRI sequences of glioblastoma in BraTS dataset. (Jandali, 2024)

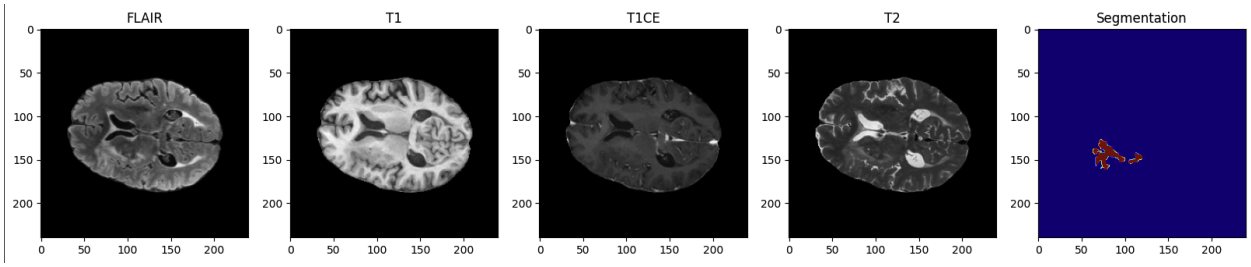

Figure 2: Resource management for RAM, GPU, and local storage. (Jandali, 2024)

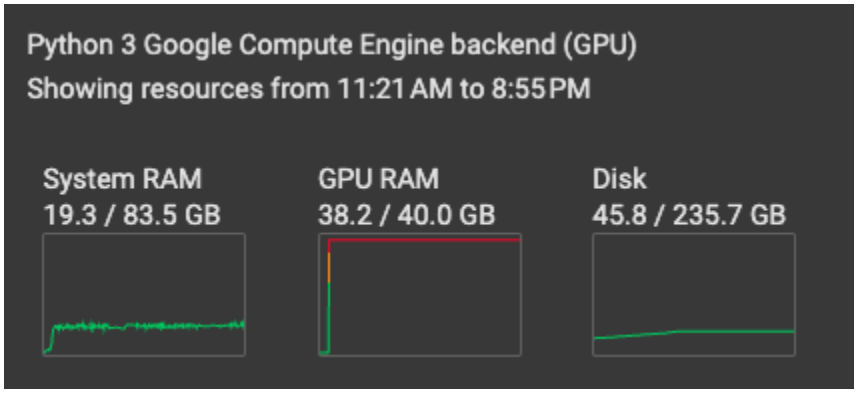

Figure 3: Metric values across final three epochs of training. (Jandali, 2024)

```
Accuracy Values Over Epochs:
Training accuracy: [0.995024561882019, 0.9954317212104797, 0.9956684708595276, 0.9957475662231445]
Validation accuracy: [0.9947934150695801, 0.9944134950637817, 0.9940317869186401, 0.9951434135437012]

F1-score Values Over Epochs:
Training f1-score: [0.7934303283691406, 0.8099790811538696, 0.8210851550102234, 0.8239758014678955]
Validation f1-score: [0.732033908367157, 0.7907635569572449, 0.7503693699836731, 0.7402228713035583]

Iou_score Values Over Epochs:
Training iou_score: [0.7045858502388, 0.7227997779846191, 0.7368743419647217, 0.7399448752403259]
Validation iou_score: [0.6501575112342834, 0.7027507424354553, 0.6625991463661194, 0.6520891785621643]

Loss Values Over Epochs:
Training loss: [0.4384726285934448, 0.41648975014686584, 0.4067210853099823, 0.3995061218738556]
Validation loss: [0.5007746815681458, 0.4524010419845581, 0.4907603859901428, 0.4559890329837799]

Learning_rate Values Over Epochs:
Training learning_rate: [0.0010000000474974513, 0.0010000000474974513, 0.0010000000474974513, 0.0010000000474974513]
```

Figure 4: Predicted segmentation of glioblastoma from 3D U-Net model. (Jandali, 2024).

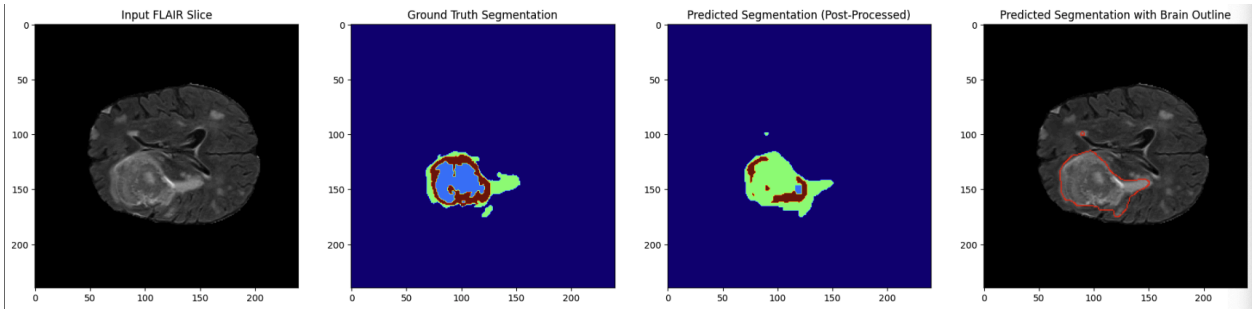

Figure 5: Predicted segmentation of glioblastoma with tissue classification. (Jandali, 2024).