Trialing U-Net Training Modifications for Segmenting Gliomas Using Open Source Deep Learning Framework

David G. Ellis(✉) and Michele R. Aizenberg

Department of Neurosurgery, University of Nebraska Medical Center, Omaha, NE, USA
david.ellis@unmc.edu

Abstract. Automatic brain segmentation has the potential to save time and resources for researchers and clinicians. We aimed to improve upon previously proposed methods by implementing the U-Net model and trialing various modifications to the training and inference strategies. The trials were performed and tested on the Multimodal Brain Tumor Segmentation dataset that provides MR images of brain tumors along with manual segmentations for hundreds of subjects. The U-Net models were trained on a training set of MR images from 369 subjects and then tested against a validation set of images from 125 subjects. The proposed modifications included predicting the labeled region contours, permutations of the input data via rotation and reflection, grouping labels together, as well as creating an ensemble of models. The ensemble of models provided the best results compared to any of the other methods, but the other modifications did not demonstrate improvement. Future work will look at reducing the level of the training augmentation so that the models are better able to generalize to the validation set. Overall, our open source deep learning framework allowed us to quickly implement and test multiple U-Net training modifications. The code for this project is available at https://github.com/ellisdg/3DUnetCNN.

Keywords: Deep learning · Brain tumor segmentation · U-Net

1 Introduction

The automatic segmentation of brain tumors from MR imaging has the potential to save clinicians and researchers time by providing accurate labeling of tumor regions without manual editing. The Multimodal Brain Tumor Segmentation (BraTS) challenge evaluates automatic methods for segmenting glioma type brain tumors by hosting an annual event that invites participants to train and test their segmentation methods on the BraTS dataset [1]. This dataset consists of MR images from hundreds of subjects along with images containing manually labeled tumor regions for those same subjects.

Previous BraTS challenges have shown deep learning to be the most accurate method to segment tumor regions [2–5]. The U-Net convolutional neural network model is a common approach chosen by many of the top-performing teams [3–6]. This model architecture employs convolution layers that encode the information stored in the images

A. Crimi and S. Bakas (Eds.): BrainLes 2020, LNCS 12659, pp. 40–49, 2021.
https://doi.org/10.1007/978-3-030-72087-2_4

at progressively smaller resolutions and then decodes the output of those layers to create a segmentation map at the original resolution. Inspired by previous challenges, our project implemented the U-Net model and trialed various modifications to the training and inference strategy.

2 Methods

2.1 Data

As a part of the challenge, BraTS provided sets of training images from 369 subjects with T1w, contrast-enhanced T1w, T2w, and T2w-FLAIR images along with manually labeled tumor segmentation maps (Fig. 1) [1, 2, 7–9]. The segmentation maps were labeled 1 for the necrotic center and non-enhancing tumor, 2 for edema, and 4 for enhancing tumor. BraTS also provided 125 subjects with sets of images without the segmentation maps as a validation group. We evaluated the performance on the validation set through submissions of segmentation maps to the BraTS challenge online portal. We cropped and resampled all input images to a size of 160 × 160 × 160 voxels.

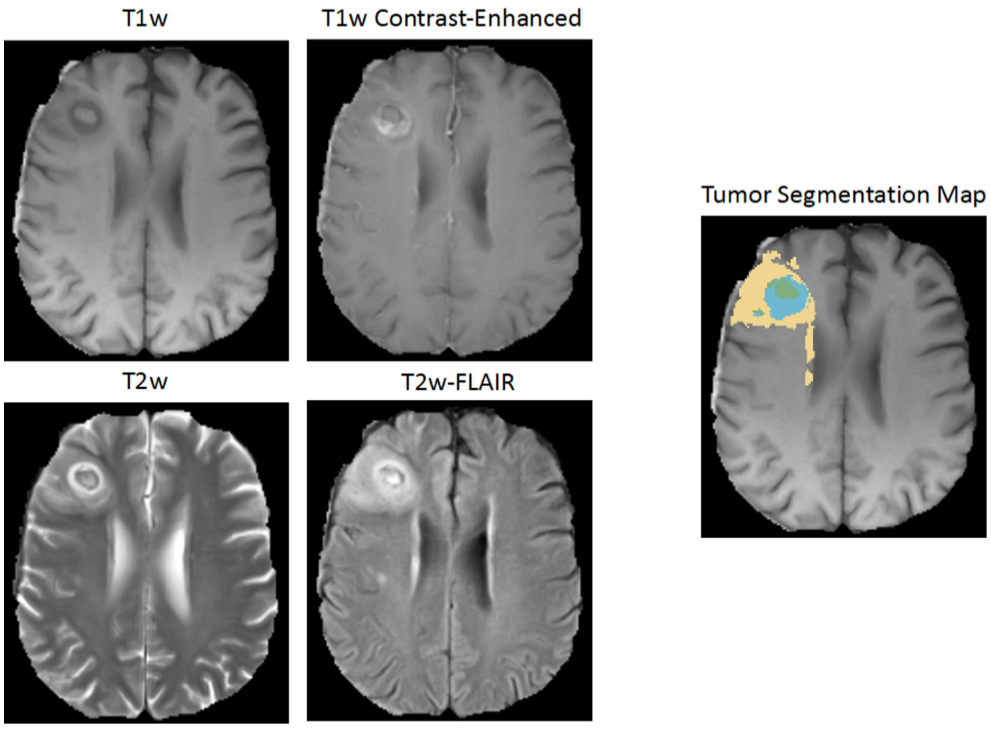

Fig. 1. Example subject part of the BraTS training dataset. Shown are the T1w, T1w with gadolinium contrast-enhancing agent, T2w, and T2w-FLAIR images used as the inputs for training and the manually segmented label map of the tumor used as the ground truth.

2.2 Model Architecture

We trained a U-Net style convolutional neural network with five layers, two ResNet style blocks per encoding and decoding layer, a base width of 32 channels, 20% channel dropout between the two encoding residual blocks of the first layer, and a receptive field of $160 \times 160 \times 160$ voxels [4, 10–12]. Each residual block consisted of two convolutional blocks performing group normalization, followed by rectified linear unit activation, and a $3 \times 3 \times 3$ convolution. The outputs of each residual block were summed with the outputs of a $1 \times 1 \times 1$ convolution of the inputs to that block. Between each encoding layer the images were downsampled by a factor of two using a strided convolution. The number of channels was doubled at each consecutive level of the model. The decoding layers concatenated the outputs of the encoding layer of the same depth to the upsampled output of the previous decoding layer. A final $1 \times 1 \times 1$ convolution linearly resampled the outputs from the 32 channels of the last decoding layer, and a sigmoid activation function was applied to predict the target segmentation maps.

2.3 Augmentation

We employed six different types of augmentation at training time, as detailed below.

Noise. Random Gaussian noise with a mean of zero and a standard deviation of 0.1 multiplied by the standard deviation of the input images was added to the input images randomly with a 50% probability per training iteration.

Blurring. The input images were blurred using a Gaussian kernel randomly with a 50% probability per training iteration. The full-width half-maximum (FWHM) of the kernel was randomly generated independently for each direction according to a normal distribution with a mean of 1.5 mm and a standard deviation of 0.5 mm.

Left–right Mirroring. The input images were mirrored so that the left and right sides of the images were randomly flipped on 50% of the training iterations.

Scale Distortion. The scale of the input images was randomly distorted for each axis independently, with a standard deviation of 0.1. The scale distortion was applied randomly on 50% of the training iterations.

Translation. The input images were randomly translated with a standard deviation of 0.05 times the extent of the cropped images. This translation was performed independently for each direction. This augmentation was applied randomly to 50% of the training iterations.

2.4 Training

Two Nvidia V100 GPUs with 32 gigabytes of memory each were used for training. One minus the average of the dice score per channel was used as the loss function during training. The initial learning rate was 10^{-4} and was decreased by a factor of 0.5 every time the validation loss plateaued for 20 epochs. Each model was trained for seven days (due to the allocated time limit on computational resources) or until the validation loss plateaued for 50 epochs.

3 Experiments

3.1 Effects of Thresholding

To perform an initial test on thresholding methods, we trained a single model on half of the training data. We then examined the effect of summing the label activations before thresholding and experimented with various threshold values.

3.2 Contours, Permutations, and Grouped Labels

In order to test different modifications to the network training, we trained a set of 4 models on the full training set with various modifications as detailed below.

Contours. The per-channel dice loss weighs all tumor voxels of the same label equally, but tumor segmentation can intuitively be considered an estimation of the boundaries of the labeled tumor regions. A high-quality segmentation will accurately estimate the boundaries between the brain and the labeled tumor, the tumor core and the edema, and the enhancing core and the necrotic center voxels. Segmentation methods often accurately segment the center of the labeled tumor regions accurately but fail to segment the boundary between two labels correctly. We tested the effect of teaching the model to focus on the region boundaries by adding the estimated contour of each label as a separate channel during training. Contours of the segmentation maps were generated by performing a binary erosion on each labeled region and then subtracting the eroded labeled region from the original labeled region.

Permutations. Combinations of rotations and reflections allow for 48 unique lossless transformations of the input feature input volumes. These permutations were performed randomly with a 50% probability for each training input. After training a model to predict the tumor labels for all permutations, we also tested the effect of permuting the validation images and then averaging the predictions over all 48 permutations.

Grouped Labels. The BraTS competition scores segmentation maps not based on the accuracy of the individual labels but rather the accuracy of the following label groups: whole tumor (WT), tumor core (TC), and enhancing tumor (ET). The WT group includes labels 1, 2, and 4, the TC group includes labels 1 and 4, and the TC group includes only label 4. We trained a model to predict the grouped labels rather than the individual labels themselves to test if this would result in higher scoring predictions [3].

3.3 Ensembles

Ten models were trained via 10-fold cross-validation on the training set. The models were trained with weighted contours and randomly permutated input volumes on every iteration. The cross-validation dice loss was measured for each fold, and the models performing the best were saved along with the models that completed the full training. The fully trained models were then used to predict the validation set, and the predictions were averaged. This ensemble of predictions was compared to those from the ensemble of the best performing cross-validation models. The fully trained ensemble predictions were also compared to the averaged predictions of both the best cross-validation models and the fully trained models.

3.4 Test Set

The ensemble of the ten fully trained models was used to predict the BraTS 2020 test set consisting of 166 sets of images provided without segmentation maps. The predicted segmentation maps were submitted through BraTS challenge online portal and the results were reported back to the authors after the completion of the challenge.

4 Results

4.1 Effects of Thresholding

The summation of the label activations prior to thresholding with a 0.5 threshold resulted in improved dice scores for all three regions as well as shorter Hausdorff95 distances for the WT and TC regions as compared to no summation before thresholding (Table 1). Interestingly, summing the activations and then thresholding at 0.9 resulted in the highest ET region scores.

Table 1. Mean Dice and Hausdorff95 measurements of the predictions resulting from various thresholding of the sigmoid label activations on the validation set. T refers to the threshold that differentiates between unlabeled and labeled voxels. Sum (Y/N) refers to whether or not the responses for each label were summed before thresholding.

		Dice			Hausdorff95		
T	Sum	ET	WT	TC	ET	WT	TC
0.3	N	0.7149	0.8938	0.8089	36.5840	4.9753	**7.9247**
0.5	N	0.7238	0.8943	0.8071	33.5743	5.3918	10.0016
0.7	N	0.7294	0.8581	0.7726	30.3896	8.7654	12.5598
0.9	N	0.6535	0.7536	0.6768	36.5133	9.6495	13.2661
0.5	Y	0.7231	**0.9002**	**0.8090**	33.6038	**4.8244**	9.8996
0.7	Y	0.7307	0.8949	0.8080	30.6517	5.2314	10.0038
0.9	Y	**0.7322**	0.8659	0.8015	**27.6895**	6.1277	10.2072

4.2 Contours, Permutations, and Grouped Labels

Weighting the contours of the labeled regions slightly improved the whole tumor dice score, but resulted in lower dice scores for the other regions compared to the baseline model trained without the weighted contours as shown in Table 2. Adding random permutations during training did not increase the dice scores but did result in lower Hausdorff95 distances for the whole tumor and tumor core regions. Permuting the data and averaging the predictions for all possible permutations did not improve results. Predicting the grouped labels rather than the individual labels resulted in better ET and TC dice scores but worse TC Hausdorff95 distance. Overall, none of the proposed alterations demonstrated enhanced validation scores over any of the other methods for every evaluation metric.

Table 2. Mean Dice and Hausdorff95 measurements of the predictions from models with the proposed modifications on the BraTS validation dataset. The proposed modifications were models that predicted the contours and as well as the labels, augmentation via permutations, averaging the predictions from each permutation, and predicting the WT, TC, and ET label groups.

	Dice			Hausdorff95		
Trial	ET	WT	TC	ET	WT	TC
Baseline	**0.7412**	0.8988	0.8086	**28.2891**	5.0059	13.9990
w/contours	0.7296	**0.9032**	0.8046	32.3023	5.8643	13.8081
w/contours & permutations	0.7263	0.8995	0.8083	36.3551	**4.7713**	**7.2312**
w/contours, permutations, and permuted predictions	0.7278	0.9014	0.8057	36.2750	4.8782	8.7185
w/contours, permutations, & grouped labels	0.7392	0.9003	**0.8136**	35.9019	4.9854	10.3612

4.3 Ensembles

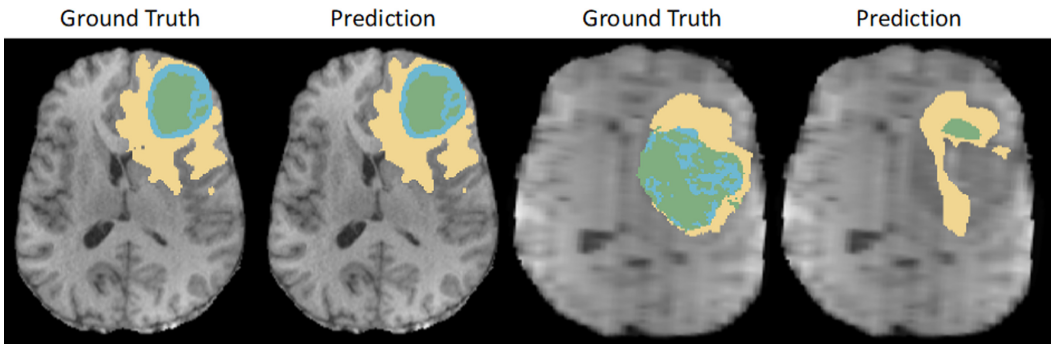

Fig. 2. Example predictions from a single of the cross-validated model on the training set compared to the ground truth segmentation maps overlayed onto the T1w image from two separate subjects. The segmentation maps are colored yellow for edema, blue for enhancing tumor, and green for the necrotic center. The predicted map on the left demonstrates a highly accurate predicted segmentation. In contrast, the predicted map on the right shows a very poor predicted segmentation, likely due to the poor-quality imaging. (Color figure online)

Overall, the baseline models that trained for the full seven days performed the best on the validation set as shown in Table 3. Figure 2 shows example predictions from a single model alongside the ground truth segmentations on the cross-validated training data. Surprisingly, the models that performed the best on the held-out cross-validation data during training had worse scores on the validation set. Also, the combined predictions from all the models scored worse than the baseline models on the validation set.

4.4 Test Set Results

The baseline ensemble consisting of the ten fully trained models was used to predict the test set cases. The results are listed in Table 4 and example predictions are shown in Fig. 3.

Table 3. Mean Dice and Hausdorff95 measurements of the predictions for the validation set from ensembles of cross-validated models. The baseline models were trained for seven days. During training, the models that performed the best on the cross-validation hold-out data were saved. The predictions from both the baseline and best models were also averaged and evaluated.

	Dice			Hausdorff95		
Trial	ET	WT	TC	ET	WT	TC
Baseline (10 models)	**0.7530**	**0.9071**	0.8292	**32.6782**	**4.5609**	9.2570
Best models (10 models)	0.7451	0.9067	**0.8318**	35.6361	4.5914	**9.1908**
Combined (20 models)	0.7444	0.9070	0.8303	35.6515	4.5695	9.2499

Table 4. Results of the ten model U-Net ensemble on the BraTS 2020 test set.

	Dice			Hausdorff95		
	ET	WT	TC	ET	WT	TC
Mean	0.8162	0.8779	0.8475	11.2353	11.0939	19.2312
Std Dev	0.1863	0.1512	0.2372	57.1383	49.5001	74.6450
Median	0.8528	0.9217	0.9248	1.4142	3.0000	2.2361
25^{th} quantile	0.8003	0.8839	0.8727	1.0000	1.7321	1.4142
75^{th} quantile	0.9170	0.9493	0.9595	2.4495	5.3852	3.9354

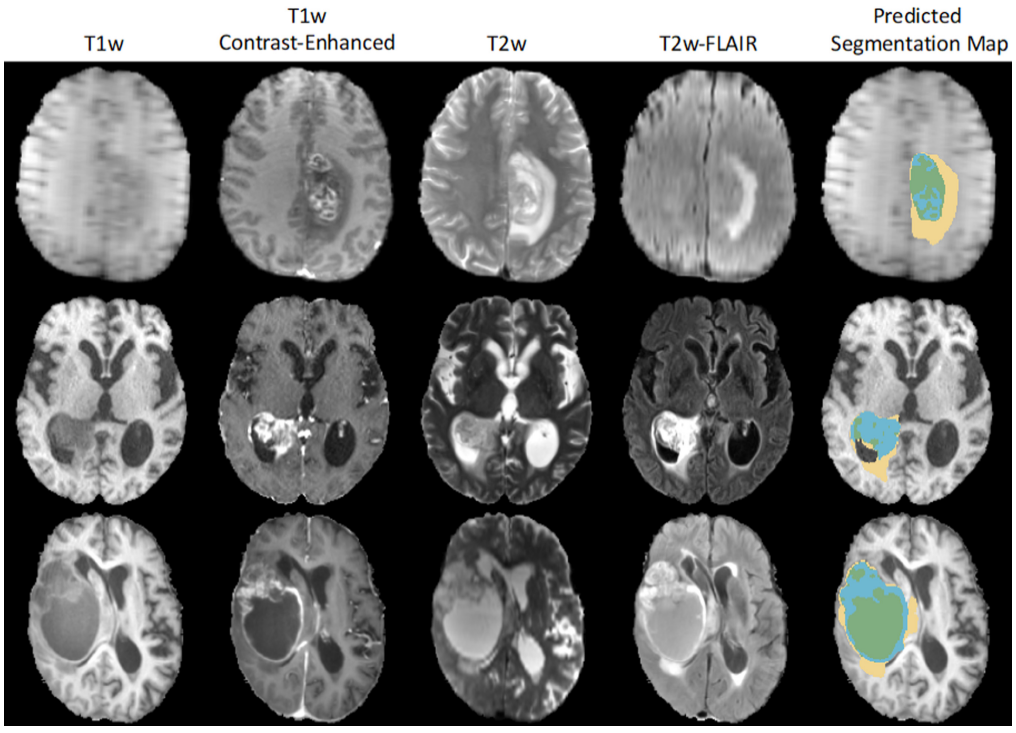

Fig. 3. Example images and predictions from the testing set. Predicted segmentation maps are shown on the far right overlayed on top of the T1 image. Predicted segmentation of the edema is shown in yellow, enhancing tumor in blue, and necrotic center in green. (Color figure online)

5 Discussion

Using an ensemble of models trained via cross-validation offered the best predictions out of all the modifications proposed. This result aligns with results from previous challenges showing that ensembles of models often out-perform the individual models [2, 3]. However, combining more models saved during training did not further increase the scores on the validation set, indicating that bigger ensembles do not always yield better results. It is likely that the models that scored the best on the cross-validation were slightly overfitted to the training set and thus decreased the scores of the fully trained models when ensembled together. Furthermore, creating ensembles with different style of models and approaches is likely to yield more diverse predictions across models in the ensemble. This diversity of predictions may produce better results than the ensembles used in this project that consisted only of identically trained U-Net models.

The median scores of the model were much higher than the average scores, especially for the Hausdorff95 distances. This indicates that the average scores are being heavily skewed by poor predictions of outlying cases. It is possible that post-processing the predictions to account for these outlier cases as performed in [3] may provide enhanced results.

Though much of the segmentation errors occur at the boundaries between labeled tumor regions, weighting the contours of the regions did not improve the results on the validation set. Training with the weighted contours produced predictions with worse dice scores for the ET and TC regions than a model trained without the weighted contours.

Applying random permutations to the input data produced mixed results. The scores on the model trained with the permuted data were not sizably better in any of the scores except for the TC Hausdorff 95 distance, which was much lower than the same model trained without permutations. Surprisingly, averaging the predictions from all possible permutations did not improve the accuracy of the combined predictions on the validation set. It is possible that the rotation of the volumes hurts the model's ability to learn information that is specific to a given axis. For example, slices in the z-axis are typically thicker than the slices in the x and y axes, prior to resampling. Furthermore, the BraTS protocol describes annotations being performed on the axial slices [2]. By rotating the volumes during training, the model is forced to treat all axes equally and cannot learn to mimic the patterns that may exist due to the manual annotation methodology. This can be avoided by only doing mirroring rotation that flip axes without rotating the volumes.

Another advantage of not performing rotation permutations is that receptive field is not required to be the shape of a cube. We used a receptive field that was 160 $\times$ 160 $\times$ 160 voxels. This allowed any one dimension to be rotated and switched with another dimension. If no rotations are used, then the receptive field can be changed to better match the data. Brain images tend to be longer in the x and y dimensions than in the z dimension. Therefore, a smaller number of voxels along the z axis and a larger number of voxels along the y axis may produce better results than our cube shaped receptive field.

Overall, the augmentation strategy appears to have been too aggressive. A training set size of 369 subjects is likely big enough to train generalizable models without such heavy augmentation. When using datasets this big, permutating the input data via rotations

should likely be avoided as the model may be missing valuable information encoded only in a given direction.

6 Conclusion

We trialed several different modifications to U-Net model training, including weighting the region contours, grouping the labels, permuting the input data, and combining models into ensembles. Overall, creating the ensembles of models was the only modification that demonstrated a clear improvement over other methods. Our open source code for this project is available at https://github.com/ellisdg/3DUnetCNN.

Acknowledgements. This work was completed utilizing the Holland Computing Center of the University of Nebraska, which receives support from the Nebraska Research Initiative.

References

1. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Trans. Med. Imaging **34**(10), 1993–2024 (2014)
2. Bakas, S., 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)
3. Jiang, Z., Ding, C., Liu, M., Tao, D.: Two-stage cascaded U-Net: 1st place solution to BraTS challenge 2019 segmentation Task. In: Crimi, A., Bakas, S. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part I, pp. 231–241. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-46640-4_22
4. Myronenko, A.: 3D MRI brain tumor segmentation using autoencoder regularization. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, pp. 311–320. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_28
5. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: No new-net. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 4th International Workshop, BrainLes 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Revised Selected Papers, Part II, pp. 234–244. Springer International Publishing, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_21
6. Isensee, F., Kickingereder, P., Wick, W., Bendszus, M., Maier-Hein, K.H.: Brain tumor segmentation and radiomics survival prediction: contribution to the brats 2017 challenge. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, pp. 287–297. Springer International Publishing, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_25
7. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. Cancer Imaging Arch. **286** (2017)

8. Bakas, S., et al.: 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)
9. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
10. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015. Lecture Notes in Computer Science, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
11. He, K., et al.: Deep residual learning for image recognition. CoRR abs/1512.03385 (2015)
12. Ellis, D.G., Aizenberg, M.R.: Structural brain imaging predicts individual-level task activation maps using deep learning. bioRxiv, p. 2020.10.05.306951 (2020)