Brain Tumor Segmentation and Radiomics Survival Prediction: Contribution to the BRATS 2017 Challenge

Fabian Isensee[1], Philipp Kickingereder[2], Wolfgang Wick[3], Martin Bendszus[2], and Klaus H. Maier-Hein[1]

[1] Medical Image Computing, German Cancer Research Center (DKFZ),
Heidelberg, Germany
[2] Department of Neuroradiology, University of Heidelberg Medical Center,
Heidelberg, Germany
[3] Neurology Clinic, University of Heidelberg Medical Center,
Heidelberg, Germany

Abstract. Quantitative analysis of brain tumors is critical for clinical decision making. While manual segmentation is tedious, time consuming and subjective, this task is at the same time very challenging to solve for automatic segmentation methods. In this paper we present our most recent effort on developing a robust segmentation algorithm in the form of a convolutional neural network. Our network architecture was inspired by the U-Net and has been carefully modified to maximize brain tumor segmentation performance. We use a dice loss function to cope with class imbalances and use extensive data augmentation to successfully prevent overfitting. Our method beats the current state of the art on BraTS 2015 and shows promising results on the BraTS 2017 validation set (dice scores of 0.896, 0.797 and 0.732 for whole tumor, tumor core and enhancing tumor, respectively). We furthermore take part in the survival prediction subchallenge by training an ensemble of a random forest regressor and a multilayer perceptron ensemble on shape features describing the tumor subregions. Our ensemble achieves 335.08 root mean squared error (232.76 mean absolute error) in a five fold cross-validation over the 163 training cases.

Keywords: CNN, Brain Tumor, Glioblastoma, Deep Learning

1 Introduction

Quantitative assessment of brain tumors provides valuable information and therefore constitutes an essential part of diagnostic procedures. Automatic segmentation is attractive in this context, as it allows for faster, more objective and potentially more accurate description of relevant tumor parameters, such as the volume of its subregions. Due to the irregular nature of tumors, however, the development of algorithms capable of automatic segmentation remains challenging.

The brain tumor segmentation challenge (BraTS) [1] aims at encouraging the development of state of the art methods for tumor segmentation by providing a large dataset of annotated low grade gliomas (LGG) and high grade glioblastomas (HGG). Unlike the previous years, the BraTS 2017 training dataset, which consists of 210 HGG and 75 LGG cases, was annotated manually by one to four raters and all segmentations were approved by expert raters [2–4]. For each patient a T1 weighted, a post-contrast T1-weighted, a T2-weighted and a FLAIR MRI was provided. The MRI originate from 19 institutions and were acquired with different protocols, magnetic field strengths and MRI scanners. Each tumor was segmented into edema, necrosis and non-enhancing tumor and active/enhancing tumor. The segmentation performance of participating algorithms is measured based on the DICE coefficient, sensitivity, specificity and Hausdorff distance. Additionally to the segmentation challenge, BraTS 2017 also required participants to develop an algorithm for survival prediction. For this purpose the survival (in days) of 163 training cases was provided as well.

Inspired by the recent success of convolutional neural networks, an increasing number of deep learning based automatic segmentation algorithms have been proposed. Havaei et al. [5] use a multi-scale architecture by combining features from pathways with different filter sizes. They furthermore improve their results by cascading their models. Kamnitsas et al. [6] proposed a fully connected multi-scale CNN that was among the first to employ 3D convolutions. It comprises a high resolution and a low resolution pathway that are recombined to form the final segmentation output. For their submission to the brain tumor segmentation challenge in 2016 [7], they enhanced their architecture through the addition of residual connections for improved segmentation performance. They addressed the class imbalance problem through a sophisticated training data sampling strategy. Kayalibay et al. [8] developed very successful adaptation of the popular U-Net architecture [9] and achieved state of the art results for the BraTS 2015 dataset. Notably, they employed a Jaccard loss function that intrinsically handles class imbalances. They make use of the large receptive field of their architecture to process entire patients at once, at the cost of being able to train with only one patient per batch. Here we propose our contribution to the BraTS 2017 challenge that is also based on the popular U-Net architecture [9]. Our network possesses twice as many filters than [8] while being trained with a slightly smaller input patch size and a larger batch size. We furthermore employ a multiclass adaptation of the dice loss [10] and make extensive use of data augmentation.

Image based tumor phenotyping and derived clinically relevant parameters such as predicted survival is typically done by means of radiomics. Intensity, shape and texture features are thereby computed from segmentation masks of the tumor subregions and subsequently used to train a machine learning algorithm. These features may also be complemented by other measures handcrafted to the problem at hand, such as the distance of the tumor to the ventricles [11]. Although our main focus was put on the segmentation part of the challenge, we developed a simple radiomics based approach combined with a random forest regressor and a multilayer perceptron ensemble for survival prediction.

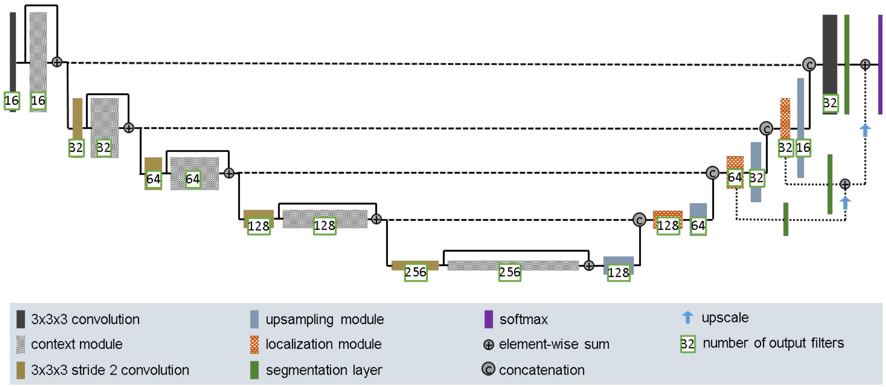

Fig. 1. Network architecture. Our architecture is inspired by the UNet [9]. The context pathway (left) aggregates high level information that is subsequently localized precisely in the localization pathway (right). Inspired by [8] we inject gradient signals deep into the network through deep supervision.

2 Methods

2.1 Segmentation

Data preprocessing With MRI intensity values being non standardized, normalization is critical to allow for data from different institutes, scanners and acquired with varying protocols to be processed by one single algorithm. This is particularly true for neural networks where imaging modalities are typically treated as color channels. Here we need to ensure that the value ranges match not only between patients but between the modalities as well in order to avoid initial biases of the network. We found the following simple workflow to work surprisingly well. First, we normalize each modality of each patient independently by subtracting the mean and dividing by the standard deviation of the brain region. We then clip the resulting images at $[-5, 5]$ to remove outliers and subsequently rescale to $[0, 1]$, with the non-brain region being set to 0.

Network architecture Our network is inspired by the U-Net architecture [9]. We designed the network to process large 3D input blocks of 128x128x128 voxels. In contrast to many previous approaches who manually combined different input resolutions or pathways with varying filter sizes, the U-Net based approach allows the network to intrinsically recombine different scales throughout the entire network. Just like the U-Net, our architecture comprises a context aggregation pathway that encodes increasingly abstract representations of the input as we progress deeper into the network, followed by a localization pathway that recombines these representations with shallower features to precisely localize the structures of interest. We refer to the vertical depth (the depth in the U shape) as level, with higher levels being lower spatial resolution, but higher dimensional feature representations. The activations in the context pathway are computed by *context modules*. Likewise, we call the processing blocks in the localization pathway *localization modules*. Each context module is in fact a pre-activation residual block [12] with two 3x3x3 convolutional layers and a dropout layer ($p_{\text{drop}} = 0.3$) in between. Context modules are connected by stride 2 3x3x3 convolutions. We increase the feature map resolution in the localization pathway by means of upscaling (size 2, stride 2) followed by a 3x3x3 convolution that halves the number of feature maps (*upsampling module*). Following the upsampling, feature maps from the localization pathway are concatenated with feature maps from the context pathway and subsequently passed to a localization module. A localization module consists of a 3x3x3 convolution followed by a 1x1x1 convolution and halves the number of feature maps. Inspired by [8] we employ deep supervision in the localization pathway by integrating segmentation layers at different levels of the network and combining them via elementwise summation to form the final network output. Throughout the network we use leaky ReLu nonlinearities for all feature map computing convolutions. We furthermore replace the traditional batch with instance normalization [13] since we found that the stochasticity induced by small batch sizes destabilizes batch normalization.

Training Procedure Our network architecture is trained with randomly sampled patches of size 128x128x128 voxels and batch size 2. We refer to an epoch as an iteration over 100 batches and train for a total of 300 epochs. Training is done using the ADAM optimizer with an initial learning rate $\text{lr}_{\text{init}} = 5 \cdot 10^{-4}$, the following learning rate schedule: $\text{lr}_{\text{init}} \cdot 0.985^{\text{epoch}}$ and a l2 weight decay of 10^{-5}.

One challenge in medical image segmentation is the class imbalance in the data that hampers the training when using the conventional categorical crossentropy loss. In the BraTS 2017 training data for example, there is 166 times as much background (label 0) as there is enhancing tumor (label 4). We approach this issue by formulating a multiclass Dice loss function that is differentiable and can be easily integrated into deep learning frameworks:

$$\mathcal{L}_{\text{dc}} = -\frac{2}{|K|} \sum_{k \in K} \frac{\sum_i u_i^k v_i^k}{\sum_i u_i^k + \sum_i v_i^k} \tag{1}$$

where u is the softmax output of the network and v is a one hot encoding of the ground truth segmentation map. Both u and v have shape i by c with i being the number of pixels in the training patch and $k \in K$ being the classes.

When training large neural networks from limited training data, special care has to be taken to prevent overfitting. We address this problem by utilizing a large variety of data augmentation techniques. Whenever possible, we initialize these techniques using aggressive parameters that we subsequently attenuate over the course of the training. The following augmentation techniques were applied on the fly during training: random rotations, random scaling, random elastic deformations, gamma correction augmentation and mirroring.

The fully convolutional nature of our network allows to process arbitrarily sized inputs. At test time we therefore segment an entire patient at once, alleviating problems that may arise when computing the segmentation in tiles with a network that has padded convolutions. We furthermore use test time data augmentation by mirroring the images and averaging the softmax outputs.

2.2 Survival Prediction

The task of survival prediction underpins the clinical relevance of the BraTS challenge, but at the same time is very challenging, particularly due to the absence of treatment information. For this subchallenge, only the image information and the age of the patients was provided. Our approach to survival prediction is based on radiomics. We characterize the tumors using image based features that are computed on the segmentation masks. We compute shape features (13 features), first order statistics (19 features) and gray level co-occurence matrix features (28 features) with the pyradiomics package [14]. The tumor regions for which we computed the features were the edema (ede), enhancing tumor (enh), necrosis (nec), tumor core (core) and whole tumor (whole). We computed only shape features for edema and the whole tumor, shape and first order features for core and the entire feature set for necrosis and enhancing. With the image features being computed for all modalities, we extracted a total of 517 features.

These features are then used for training a regression ensemble for survival prediction. Random forests are well established in the radiomics community for performing well, especially when many features but only few training data are available. These properties make random forest regressors the prime choice for the scenario at hand (518 features, 163 training cases). We train a random forest regressor (RFR) with 1000 trees and the mean squared error as split criterion. Additionally, we designed an ensemble of multilayer perceptrons (MLP) to complement the output of the regression forest. The ensemble consists of 15 MLPs, each with 3 hidden layers, 64 units per layer and trained with a mean squared error loss function. We use batch normalization, dropout ($p_{\text{drop}} = 0.5$) and add gaussian noise ($\mu = 0, \sigma = 0.1$) in each hidden layer. The outputs of the RFR and the MLP ensemble are averaged to obtain our final prediction.

3 Results

Segmentation We trained and evaluated our network on the BraTS 2017 and 2015 training datasets via five fold cross-validation. No external data was used and the network was trained from scratch. Furthermore, we used the five networks obtained by the corresponding cross-validation as an ensemble to predict the respective validation(BraTS 2017) and test (BraTS 2015) set. Both the training set and validation/test set results were evaluated using the online evaluation platforms to ensure comparability with other participants.

Table 1 compares the performance of our algorithm to other state of the art methods on the BraTS 2015 test set. Our method compares favorably to other

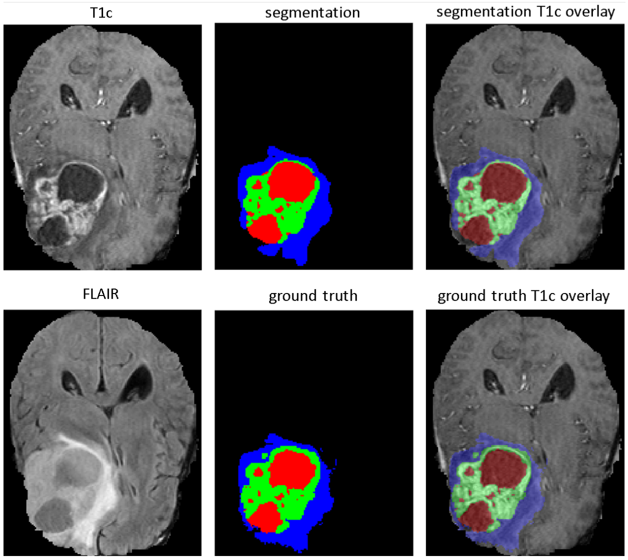

Fig. 2. Qualitative segmentation result. Our approach is capable of segmenting large as well as fine grained regions accurately.

	Dice			Sensitivity			PPV		
	whole	core	enh.	whole	core	enh.	whole	core	enh.
Kamnitsas et al. [6]	**0.85**	0.67	0.63	0.88	0.60	0.67	**0.85**	**0.86**	**0.63**
Kayalibay et al. [8]	**0.85**	0.72	0.61	**0.91**	**0.73**	0.67	0.82	0.77	0.61
ours	**0.85**	**0.74**	**0.64**	**0.91**	**0.73**	**0.72**	0.83	0.80	**0.63**

Table 1. BraTS 2015 test set results.

state of the art neural networks and is currently ranked first in the BraTS 2015 test set online leaderboard. In Table 2 we show an overview over the segmentation performance of our model on the BraTS 2017 dataset. A qualitative segmentation result (Brats17_TCIA_469_1) is shown in Figure 2. Notably, we achieve dice scores of 0.896, 0.797 and 0.732 for whole, core and enhancing, respectively, on the BraTS 2017 validation set. This result places us among the best performing methods according to the online validation leaderboard.

Survival Prediction We extensively evaluated the components of our regression ensemble as well as different feature sets with the aim of minimizing the mean squared error by running 5-fold cross-validations on the 163 provided training cases. A summary of our findings for both the ground truth and our segmentations is shown in Table 3. We observed that the random forest regressor performs very well across all feature sets while the MLP ensemble is much less stable. The overall best results were obtained by averaging the MLP ensem-

Dataset	Dice			Sensitivity			Specificity			Hausdorff Dist.		
	whole	core	enh.	whole	core	enh.	whole	core	enh.	whole	core	enh.
BraTS 2017 Train	0.895	0.828	0.707	0.890	0.831	0.800	0.995	0.997	0.998	6.04	6.95	6.24
BraTS 2017 Val	0.896	0.797	0.732	0.896	0.781	0.790	0.996	0.999	0.998	6.97	9.48	4.55

Table 2. Results for the BraTS 2017 dataset. Train: 5 fold cross-validation on the training data (285 cases). Val: Result on the validation dataset (46 cases).

Features	Ground Truth Segmentation			Our Segmentation		
	RFR	MLP ens	combined	RFR	MLP ens	combined
shape, age (66)	334.89	**352.00**	**339.61**	353.12	**343.19**	**335.08**
glcm, age (225)	348.14	462.16	381.25	**350.78**	388.99	357.41
first order, age (229)	358.69	388.44	362.20	354.66	381.42	355.89
shape, glcm, age (290)	**344.86**	431.96	367.14	346.40	378.73	349.13
shape, first order, age (294)	352.64	372.59	350.62	351.56	360.24	342.46
glcm, first order, age (453)	353.18	443.64	378.83	354.30	383.82	356.25
all (518)	350.40	385.66	354.86	352.95	372.04	348.55

Table 3. Survival prediction experiments. We trained a random forest regressor (RFR) and a MLP ensemble (MLP ens). Averaging RFR and MLP ensemble yields the *combined* result. The best root mean squared error is achieved when using RFR and MLP ensemble together with only shape features and the patients age.

ble output with the one from the random forest regressor (column *combined*) and using only shape features and the age of a patient. Interestingly, while the random forest performance is almost identical between ground truth and our segmentations, the MLP ensemble performs better on our segmentations for all feature sets, which is also reflected by the *combined* results. The best root mean squared error we achieved was 335.08 (mean absolute error 232.76).

4 Discussion

In this paper we presented contribution to the BraTS 2017 challenge. For the segmentation part of the challenge we developed a deep convolutional neural network architecture which was trained using extensive data augmentation and a dice loss formulation. We achieve state of the art results on BraTS 2015 and presented promising scores on the BraTS 2017 validation set. Training time was of about five days per network. Due to time restrictions we were limited in the number of architectural variants and data augmentation methods we could explore, yet we expect to find even better performing constellations for our final test set submission in the near future. Careful architecture optimizations already allowed us to train with large 128x128x128 patches and a batch size of 2 with 16 filters in the highest level, which is significantly more than in [8]. Training with larger batch sizes and more convolutional filters in a multi-GPU setup should yield further improvements, especially provided that we did not observe significant overfitting in our experiments. While most of our effort was concentrated on the segmentation part of the challenge, we also proposed an ensemble of a random forest regressor and a multilayer perceptron ensemble for the survival prediction subchallenge. By using only shape based features, we achieved a root mean squared error of 335.08 and a mean absolute error of 232.76 in a five fold cross-validation on the training data and using our segmentations.

References

1. B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest *et al.*, "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE TMI*, vol. 34, no. 10, pp. 1993–2024, 2015.
2. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features," *Nature Scientific Data*, 2017 (In Press).
3. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-GBM collection," *TCIA*, 2017.
4. S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection," *TCIA*, 2017.
5. M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, "Brain tumor segmentation with deep neural networks," *MIA*, vol. 35, pp. 18–31, 2017.
6. K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker, "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *MIA*, vol. 36, pp. 61–78, 2017.
7. K. Kamnitsas, E. Ferrante, S. Parisot, C. Ledig, A. V. Nori, A. Criminisi, D. Rueckert, and B. Glocker, "DeepMedic for brain tumor segmentation," in *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries.* Springer, 2016, pp. 138–149.
8. B. Kayalibay, G. Jensen, and P. van der Smagt, "CNN-based segmentation of medical imaging data," *arXiv preprint arXiv:1701.03056*, 2017.
9. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI.* Springer, 2015, pp. 234–241.
10. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *International Conference on 3D Vision.* IEEE, 2016, pp. 565–571.
11. L. Macyszyn, H. Akbari, J. M. Pisapia, X. Da, M. Attiah, V. Pigrish, Y. Bi, S. Pal, R. V. Davuluri, L. Roccograndi *et al.*, "Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques," *Neuro-oncology*, vol. 18, no. 3, pp. 417–425, 2015.
12. K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *ECCV.* Springer, 2016, pp. 630–645.
13. D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv preprint arXiv:1607.08022*, 2016.
14. J. J. M. van Griethuysen, A. Fedorov, C. Parmar, A. Hosny, N. Aucoin, V. Narayan, R. G. H. Beets-Tan, J.-C. Fillion-Robin, S. Pieper, and H. J. W. L. Aerts, "Computational radiomics system to decode the radiographic phenotype," *Cancer Research (Accepted)*, 2017.