BraTS Challenge Manuscripts

MICCAI 2014

Harvard Medical School

Boston, Massachusetts

September 14, 2014

S-17 cluster MICCAI Computational Decision Support in Brain Cancer Cluster of Events

S-W17 Computational Clinical Decision Support and Precision Medicine in Brain Cancer-BrainTumor

Chairs: K. Farahani (NCI), L. Clarke (NCI), C. Jaffe (Boston U)

This special offering of NCI-MICCIA Computational Decision Support in Brain Cancer Cluster of Events consists of a morning workshop and two afternoon image processing challenges.

The purpose of the workshop is to consider basic requirements and current resources for open science development of systems in support of computational precision medicine in brain tumor diagnosis and treatment planning.

7:00 Registration and Coffee

8:00 K. Farahani (NCI)-Welcome Introduction

8:05 D.Gutman (Emory)-The Cancer Digital Slide Archive: An Online Resource for Integrative Digital Pathology

8:25 A. Tannenbaum (Stony Brook)-Tumor Margin Delineation

8:45 R. Colen (MD Anderson)-Imaging Genomics, Imaging-omics and Big Data

9:05 H. Aerts (Harvard)-Radiomics: Getting Our ‘Omics from Imaging

9:25 W. Wells (Brigham&Women’s, Harvard)-Uncertainty Management in Image Processing

9:45 Panel Discussion- Moderator: J. Saltz (Stony Brook)

10:05 Coffee

10:30 T. Syeda-Mahmood (IBM)-Advances in tumor segmentation for disease assessment

10:50 S. Mercer (Microsoft Research)-Using CodaLab for Algorithmic Competition and Experimentation

11:10 J. Freymann, J. Kirby (NCI/Leidos) The Cancer Imaging Archive

11:30 Panel Discussion – Moderator: B. Menze (UTM)

11:50 C. Jaffe(Boston U) – Conclusion

S-17 cluster MICCAI Computational Decision Support in Brain Cancer Cluster of Events

S-C17A Brain Tumor Image Segmentation Challenge-BRATS

Chairs: M. Reyes (U. Bern), J. Kalpathy-Cramer (MGH, Harvard), B. Menze (TUM)

1:15 Presentation by Chairs: Multi-modal Image Segmentation and Classification of Brain Tumors

1:45 Presentations by top 3 challenge winners (10 min each)

2:15 General Discussion

3:00 Coffee

S-C17B Digital Pathology Classification and Segmentation Challenge

Chairs: T. Kurc (Stony Brook), J. Davis (Stony Brook), J. Saltz (Stony Brook)

3:30 Presentation(s) by Chairs: Multi-scale, Multi-modal Imaging in Cancer Research and Imaging Quantification

4:00 Presentations by top 3 challenge winners (10 min each)

4:30 General Discussion and Wrap-up

Brain Tumor Segmentation with Deep Neural Networks

Axel Davy[1], Mohammad Havaei[2], David Warde-Farley[3], Antoine Biard[4], Lam Tran[5], Pierre-Marc Jodoin[2], Aaron Courville[3], Hugo Larochelle[2], Chris Pal[3,6], and Yoshua Bengio[3]

[1] École normale supérieure, Paris, France
[2] Université de Sherbrooke, Sherbrooke, Canada
[3] Université de Montréal, Montréal, Canada
[4] École polytechnique, Palaiseau, France
[5] University of Rochester, New York, USA
[6] École Polytechnique de Montréal, Canada

Abstract. Deep Neural Networks (DNNs) are often successful in problems needing to extract information from complexe, high-dimensional inputs, for which useful features are not obvious to design. This paper presents our work on applying DNNs to brain tumor segmentation for the BRATS challenge. We are currently experimenting with several several DNN architectures, leveraging the recent advances in the field such as convolutional layers, max pooling, Maxout units and Dropout regularization. We present preliminary results, for our best performing network on the BRATS2013 training set, leaderboard dataset and challenge dataset.

The results are obtained from the evaluation tool available on the Virtual Skeleton database. While we do not beat the best results of BRATS2013 participants with our current architecture, our results are promising.

1 Introduction

Deep Neural Networks (DNNs) have recently attracted more attention due to their state-of-the-art performance on several datasets such as ImageNet [7] and CIFAR-10 [5]. DNNs have also been applied successfully to segmentation problems [2, 6], the type of task considered here. However, to the best of our knowledge, there is no existing work on DNNs applied to brain tumor segmentation.

We are currently experimenting with several architectural variations of DNNs, for tackling brain segmentation. Our best architecture, which we briefly describe here, is based on convolutional layers, Maxout [5] and Dropout [9]. We also describe future variations we'd like to investigate before the end of the challenge.

The data used here is the one available for the BRATS2013 challenge, whose training set is composed of 20 brains of High Grade (HG) patients and 10 brains of Low Grade (LG) patients. There are 5 segmentation labels: Non-tumor, Necrosis, Edema, Non-enhancing tumor and Enhancing Tumor. While the BRATS2014 challenge introduces two new optional tasks (Longitudinal Lesion Segmentation and Diagnostic Image Classification), we do not plan to participate to those.

2 Methods

We start by defining some of the building blocks that we are investigating and using in our DNN architectures. Specifically, these building blocks allow us to form different types of Convolutional Neural Networks (CNNs). CNNs are a very efficient and effective class of models for computer vision, and they have been shown to learn and extract visual features able to generalize well across many tasks [3].

We attack the problem of brain tumor segmentation by solving it slice by slice from the axial view. Thus, the input x of our model corresponds to 2D image (slice), where each pixel is associated with multiple channels, each corresponding to a different image modality.

Convolutional layer CNN features are modeled by a set of kernels convolved over the input image x, followed by an optional element-wise non-linearity (e.g. a sigmoidal non-linearity). The result of the convolution of each kernel is referred to as a feature map. The size (width, height) of the kernels are hyper-parameters that must be specified by the user. However the kernel itself is learned during training. By treating the different feature maps as channels, resulting output of a convolutional layer can again be interpreted as an image, allowing for the stacking of multiple such layers.

From the neural network perspective, feature maps correspond to a layer of several hidden artificial neurons. Specifically, each coordinate within a feature map corresponds to an individual neuron, for which the size of its receptive field corresponds to the kernel's size. A kernel's value also represents the weights of the connections between the layer's neurons and the neurons in the previous layer. It is often found in practice that the learned kernels resemble edge detectors, each kernel being tuned to a different spatial frequency, scale and orientation, as is appropriate for the statistics of the training data.

Maxout convolutional layer This is a variant of a convolutional layer. In this case, each feature map is instead associated with 2 kernels. The feature map is computed by convolving both kernels and taking the pair-wise maximum value between both convolutions. See [5] for more details.

Max pooling layer In order to introduce invariance to local deformations such as translation, it has been found beneficial to subsample feature maps by taking the maximum feature (neuron) value over sub-windows, within each feature map. Such an operation is known as max pooling.

Fully connected layer Neurons in a convolutional layer have limited receptive field, meaning that each neuron only depends on a small local patch within the image. Moreover, within a feature map, neurons share the same set of weights for their connections with the previous layer. Fully connected layer do without these constraints: each hidden unit in the layer is connected to all units in the previous layer, and the weights of these connections are specific to each neuron. The size of the hidden layer must be specified and is considered as a hyper-parameter.

Fully connected Maxout layer This is simply the fully connected version of the Convolutional Maxout layer. In practice we use 5 set of weights for this layer instead of 2 as opposed to the convolutional Maxout layer.

Softmax layer This is a special case of fully connected layer, where the activation function is the softmax function: $softmax(\mathbf{a}) = \exp(\mathbf{a})/Z$ where Z is a normalization constant. In words, this function converts real valued vectors into a vector with positive entries that sum to one, and thus that can be interpreted as a probability distribution. Such a layer is usually used for the last (output) layer, to obtain a distribution over segmentation labels.

Dropout Dropout is a regularization method that stochastically adds noise in the computation of the hidden layers of a DNN. This is done by multiplying each hidden or input unit by 0 (i.e. masking) with a certain probability (e.g. 0.5), independently for each unit. This encourages the neural network to learn features that are useful ”on their own” since each unit cannot assume that other units in the layer won't be masked. At test time, units are instead multiplied by one minus the probability of being masked. For more details, see [9].

The above building blocks open the door to several architectural choices in designing a DNN model. We are currently exploring several such variations. In this paper, we focus on the architecture that has been working best so far.

2.1 Preprocessing

In an attempt to test the ability of DNNs to learn useful features from scratch, we employed only minimal preprocessing. We removed the 1% highest and lowest intensities, as done in [8].

We then applied the N4ITK filter on the T1 and T1c modalities. We did not applied it to T2 and FLAIR, because the intensity of the tumor can get attenuated by the filter when the tumor region is large, especially at the center of the tumor. These choices were found to work best in our experiments. To apply N4ITK, we used ANTS [1]. We then normalized the data within each input channel, by subtracting channel's mean and dividing by the channel's standard deviation.

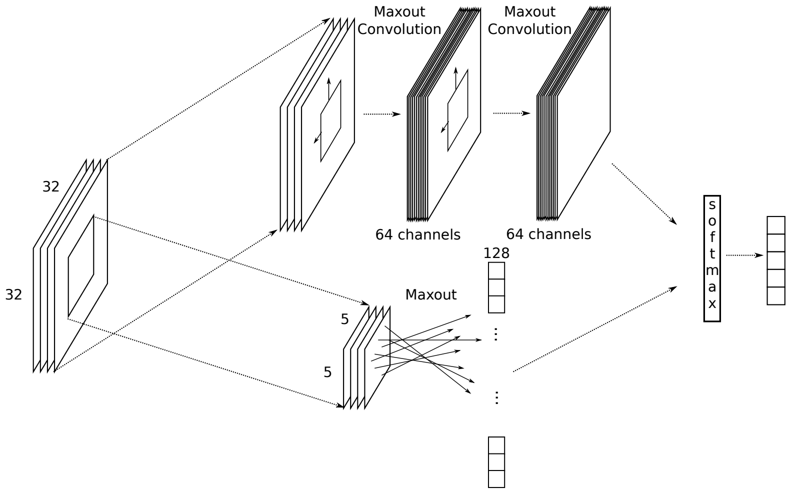


Fig. 1: Our best architecture has two paths: one concentrates on a small region around the pixel to classify, while the other looks at a wider region. The smaller path uses a fully connected Maxout layer, while the larger path is composed of two Maxout convolutional layers. The two paths' outputs are merged into a fully-connected softmax layer, which is used as our model for the segmentation's label distribution.

2.2 Current best architecture

Our best architecture is illustrated in Figure 1. It is a DNN trained on patches taken from 2D slices of the brains. Specifically, it is trained on 32x32 patches of 2D slices to predict the label of the pixel at the center of the patch.

The network has two pathways: The first is a *convolutional pathway*, connected to the entire 32x32 patch, while the second is *full-connected* to a smaller 5x5 sub-window at the center of the patch and has fewer layers. The motivation for this architectural choice is that we want the decision on the label of a pixel to be influenced by two aspects: the visual details of the region around that pixel and its larger "context" (are we near the skull, etc.). The full-connected pathway serves the first purpose while the convolutional pathway serves the latter. In our experiments, we find that the full-connected pathway is not as vital to get good performance, but helps get better contours (Figure 2).

3 Implementation details

Our implementation is based on the Pylearn2 library [4]. Pylearn2 is an open-source machine learning library specializing in deep learning algorithms. It also supports the use of GPUs, which can greatly accelerate the execution of deep learning algorithms.

To train the network, we use stochastic gradient descent with Dropout. The loss is the negative log of the probability of the correct label, where probability is read out of the output softmax layer. We first train on inputs chosen randomly, but such that all labels are equiprobable. Then, we re-train the softmax layer with a more representative distribution of the labels. We found that regularisation is very important in obtaining good results. On all the layers, we bound the absolute value of the weights and on the softmax layer we apply both L1 and L2 regularization

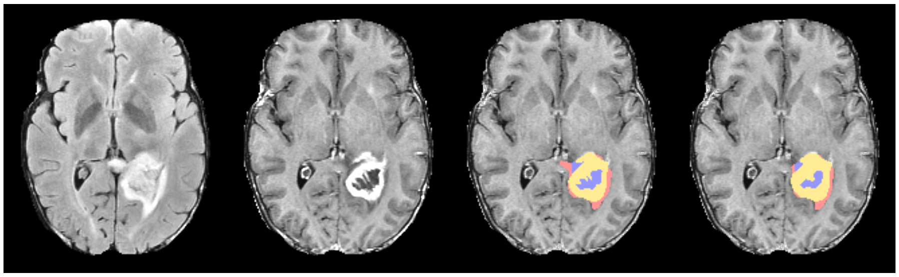


Fig. 2: The FLAIR and T1C of brain HG_0310, slice 77, followed by the segmentation produced by a network with our best architecture and by the segmentation from a similar network but without the full connected pathway. We see that the full connected pathway allows the network to more finely detail the boundary between different labels.

Name	Dice score			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
HG_0301	0.83	0.80	0.72	0.80	0.75	0.64	0.87	0.85	0.83
HG_0302	0.85	0.68	0.76	0.77	0.83	0.74	0.94	0.58	0.78
HG_0303	0.84	0.86	0.70	0.84	0.82	0.61	0.84	0.90	0.82
HG_0304	0.83	0.79	0.62	0.85	0.76	0.53	0.81	0.82	0.73
HG_0305	0.85	0.70	0.64	0.80	0.72	0.50	0.90	0.69	0.88
HG_0306	0.87	0.77	0.70	0.93	0.85	0.73	0.82	0.70	0.67
HG_0307	0.87	0.36	0.42	0.86	0.24	0.40	0.88	0.70	0.44
HG_0308	0.90	0.87	0.67	0.88	0.91	0.59	0.92	0.84	0.78
HG_0309	0.81	0.73	0.79	0.96	0.68	0.73	0.70	0.78	0.86
HG_0310	0.83	0.88	0.81	0.85	0.83	0.72	0.81	0.93	0.92
Total	0.85	0.74	0.68	0.85	0.74	0.62	0.85	0.78	0.77

Table 1: Results per brain and on the total for the 2013 Challenge dataset.

to prevent overfitting. We've also found that adding additional layers to the network doesn't give any performance improvement.

At test time, when segmenting an entire brain, we have to compute predictions one pixel at a time, which takes around 20 minutes per brain (using a GPU and including preprocessing). Faster predictions could be made by implementing the computation of both pathways as with convolutions over the entire brain. This is due to the nature of convolutions, where the weights are shared along different spatial positions.

4 Results

Table 1 shows our results on the 2013 Challenge dataset for our best architecture. We didn't have time to train and test our network on the 2014 dataset. In Table 2 are also presented our results on the Training and Leaderboard datasets.

With the current version of the architecture without any post processing, we are ranked 10^{th} on the Challenge, 8^{th} on the Training and 5^{th} on the Leaderboard datasets.

Given the minimal preprocessing we have used, these results are quite good. Additional preprocessing, such as the identification of white/gray matter and the cerebro-spinal fluid (CSF) would surely help the network have fewer false positives and increase its performance. Postprocessing could also help us remove some false positives. The network tends to have more false positives near the skull and at the top and the bottom of the brain.

4.1 Other architectural variations tested

We tried variations of the architecture to incorporate 3D information from the data. However, the results were not satisfying. A variation we tried was to give 3 adjacent patches along the third dimension as input, instead of a single slice. However it made no difference in the performance of the model, suggesting a single slice contains sufficiently enough information.

Name	Dice score			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
Train HG/LG	0.79	0.68	0.57	0.81	0.75	0.54	0.79	0.67	0.63
Leaderboad	0.72	0.63	0.56	0.69	0.64	0.50	0.82	0.68	0.68

Table 2: Results for the 2013 Training and Leaderboard datasets.

Also, we tried giving 3 orthogonal patches around the pixel to classify, by taking slides along the 3 possible directions. Due to differences in resolutions of the MRI data, we found this architecture to overfit on the training data and not generalize well.

5 Future work

We intend to further investigate architectural variations before the challenge's deadline.

Instead of training based on the prediction of an individual (center) pixel, we wish to design architectures that can jointly predict several neighbouring labels. This would allow us to more directly model the expected dependencies between the labels of nearby pixels. We mention in Section 3 that predictions at multiple locations could be obtained by implementing both the convolutional and fully-connected pathways as convolutions over larger regions than the current 32x32 input patches. We have already implemented a simpler version of this approach (without the full-connected pathway path), for which predictions on an entire brain takes around 1 minute (using a GPU). We are thus in a good position to start exploring models making structured predictions of the labels.

One approach we will investigate is to incorporate the DNN's outputs within a Conditional Random Field (CRF) model of the distribution over the labels. The CRF would incorporate pair-wise potentials between adjacent pixel positions. Another approach would be to design an architecture with cascaded predictions, where predictions further down the cascade would use as inputs the predictions computed earlier in the cascade.

6 Conclusion

In this paper we have proposed a way to do brain tumor segmentation with deep neural networks. We described our current best architecture and identified certain modeling choices that we've found important to obtain good performances. The time needed to segment an entire brain is around 20 minutes with a GPU accelerated implementation and we are confident we can decrease this to just a few minutes. We are optimistic that better results will be obtained with the not-yet-implemeted architecture using a CRF output model and improved preprocessing/postprocessing.

References

1. Brian B Avants, Nick Tustison, and Gang Song. Advanced normalization tools (ants). *Insight J*, 2009.
2. Ross Girshick Bharath Hariharan1, Pablo Arbel and Jitendra Malik. Simultaneous detection and segmentation. *arXiv preprint arXiv:1407.1808*, 2014.
3. Jeff Donahue, Yangqing Jia, Oriol Vinyals, Judy Hoffman, Ning Zhang, Eric Tzeng, and Trevor Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. In *ICML*. 2014.
4. Ian J. Goodfellow, David Warde-Farley, Pascal Lamblin, Vincent Dumoulin, Mehdi Mirza, Razvan Pascanu, James Bergstra, Frédéric Bastien, and Yoshua Bengio. Pylearn2: a machine learning research library. *arXiv preprint arXiv:1308.4214*, 2013.
5. Ian J. Goodfellow, David Warde-Farley, Mehdi Mirza, Aaron Courville, and Yoshua Bengio. Maxout networks. In *ICML*, 2013.
6. Gary B. Huang and Viren Jain. Deep and wide multiscale recursive networks for robust image labeling. *arXiv preprint arXiv:1310.0354*, 2013.
7. A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet classification with deep convolutional neural networks. In *NIPS*. 2012.
8. Chris Durst Nick Tustison, Max Wintermark and Brian Avants. Ants and árboles. In *NCI-MICCAI BRATS*, 2013.
9. Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

Extremely randomized trees based brain tumor segmentation

Michael Goetz[1], Christian Weber[1,2], Josiah Bloecher[1], Bram Stieltjes[2], Hans-Peter Meinzer[1], and Klaus Maier-Hein[1]

[1]Medical and Biological Informatics, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Quantitative Image-based Disease Characterization, DKFZ, Heidelberg, Germany

Abstract. Random Decision Forest-based approaches have previously shown promising performance in the domain of brain tumor segmentation. We extend this idea by using an ExtraTree-classifier. Several features are calculated based on normalized T1, T2, T1 with contrast agent and T2 Flair MR-images. With these features an ExtraTree-classifier is trained and used to predict different tissue classes on voxel level. The results are compared to other state-of-the-art approaches by participating at the BraTS 2013 challenge.

1 Introduction

The segmentation of brain tumors is an important prerequisite in different scenarios related to treatment controlling, radiotherapy planning and longitudinal studies. Manual segmentation is not only time-consuming and prone to errors, but additionally complicated by the fact that the necessary information is distributed over different MR-contrasts. Therefore a lot of research has been done to improve the segmentation process and create automatic segmentation methods based on multimodal MR images.

A promising approach is the use of Random Decision Forests like as done in the works of Reza et al. [1], Tustison et al. [1] and Zikic et al. [2], which learn the appearance of tumorous and healthy tissue using this method.

While the proposed solution is similar to those mentioned before it differs mainly in the used classifier. Instead of Random Decision Forests [4] we use Extremely randomized Trees (ExtraTrees) [3] which are similar to Random Decision Forests but introduce more randomness during the training phase. It has previously been shown that this often improves the variance / bias trade-off and gives slightly better results than Random Decision Forests do [3].

2 Method

2.1 Preprocessing

The preprocessing pipeline for our experiments consisted of two steps. First the N4 bias field correction algorithm [5] was used to correct nonuniformity within each MR-file. In a second step the histogram was normalized. This is especially challenging in the case of brain tumor MR images. In addition to the usual MR-artefacts which cause bright areas in parts of the image, the large variability of brain tumors has a massive influence on the histogram. Figure 1 shows some exemplary non-normalized histograms. It can be clearly seen that they differ not only in range of values but also in shape. Normalizing these histograms to match a template histogram as it is done by the pice-wise linear normalization [6] can lead to a wrong result if the shapes are too different.

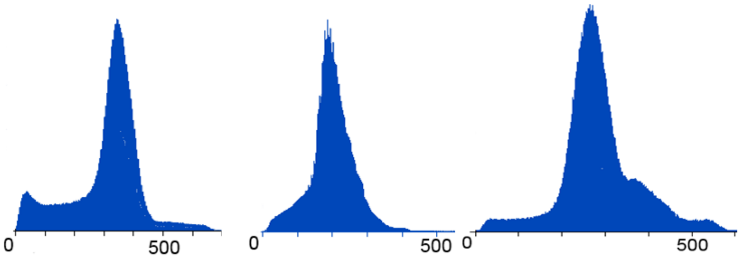


Fig. 1. Exemplary histograms of 3 non-normalized MR-Flair-images out of the BraTS-dataset. The histogram is over the complete non-zero image.

To overcome these problems a simple normalization to the image mode, e.g. the gray-value of the highest histogram bin, was used. This was done by subtracting the mode from each gray-value and then normalizing the standard derivation to 1.

2.2 Features

54 features were calculated for each voxel and each modality. The features of all modalities were then combined into the final feature vector.

Gray Value: The gray value of each voxel was used as a feature. The images were also filtered with gaussian filters with a sigma of 3 and 7 voxel-lengths and the corresponding gray values were used as features.

Local Histogram: A local histogram was calculated within a radius of 5 voxels; each of the 11 bins were used as features.

First order statistics: Within a radius of 3 voxels the mean, variance, skewness, kurtosis, minimum and maximum of all gray values were added as features.

Second order statistics: A co-occurrence matrix [7] filled with all values within a radius of 3 was used to calculate the second order statistics for the three main directions. The features extracted from the co-occurrence matrix were energy, entropy, correlation, inertia, clustershade, clusterprominence, harralick feature, and the difference of moments.

Histogram based segmentations: The output segmentation of some widely used parameter-less automatic threshold methods implemented in ITK [8] were used as features, namely Huang, Intermode, Isodata, Kittler, Li, Entropy, Moments and Otsu. For all except the Otsu-threshold a two-class problem is assumed. For the Otsu, a two-, a three- and a four-class problem were assumed.

2.3 Classifier

An Extremely Randomized Trees (ExtraTrees) [3] classifier was used. This classifier is similar to Random Decision Forests but differs in how the randomness is introduced during the training. To train an ExtraTrees-classifier multiple trees are trained, each tree is trained on all training data. Similar to the Random Decision Forest the best split at a node is found by analyzing a subset of all available features. Instead of searching for the best threshold for each feature a single threshold for each feature is selected at random. From these random splits the one that leads to the highest increase in the used score is then selected. The higher grade of randomness during the training yields more independent trees and thus further decreases the variance [3]. Due to that ExtraTrees tend to give slightly better results than Random Decision Forests.

For the training of the classifier 5% of the training data were randomly sampled to reduce the training time. The classifier is then trained combining the features described above to a 208-dimensional feature vector.

2.4 Experiments

The results were evaluated by participating in the BraTS 2013 challenge. A classifier is trained on the 20 training datasets using all available modalities, namely T2 Flair, T1, T1 with contrast agent and T2. With the so-trained classifier the 10 high-grade glioma evaluation datasets are labeled and the results are evaluated by the provided online tool.

For the evaluation the overlap with 3 labels is measured using the DICE-score. The first label, the *complete tumor*, includes necrosis, edema and both enhancing and non-enhancing tumor. The second label, *tumor core*, is the same as the complete tumor but without edema. Finally, the label *enhancing tumor* is evaluated.

3 Results

Table 1 provides the DICE-scores for the test cases. Figure 2 and Figure 3 depict exemplary slices of the original images and the retrieved segmentations.

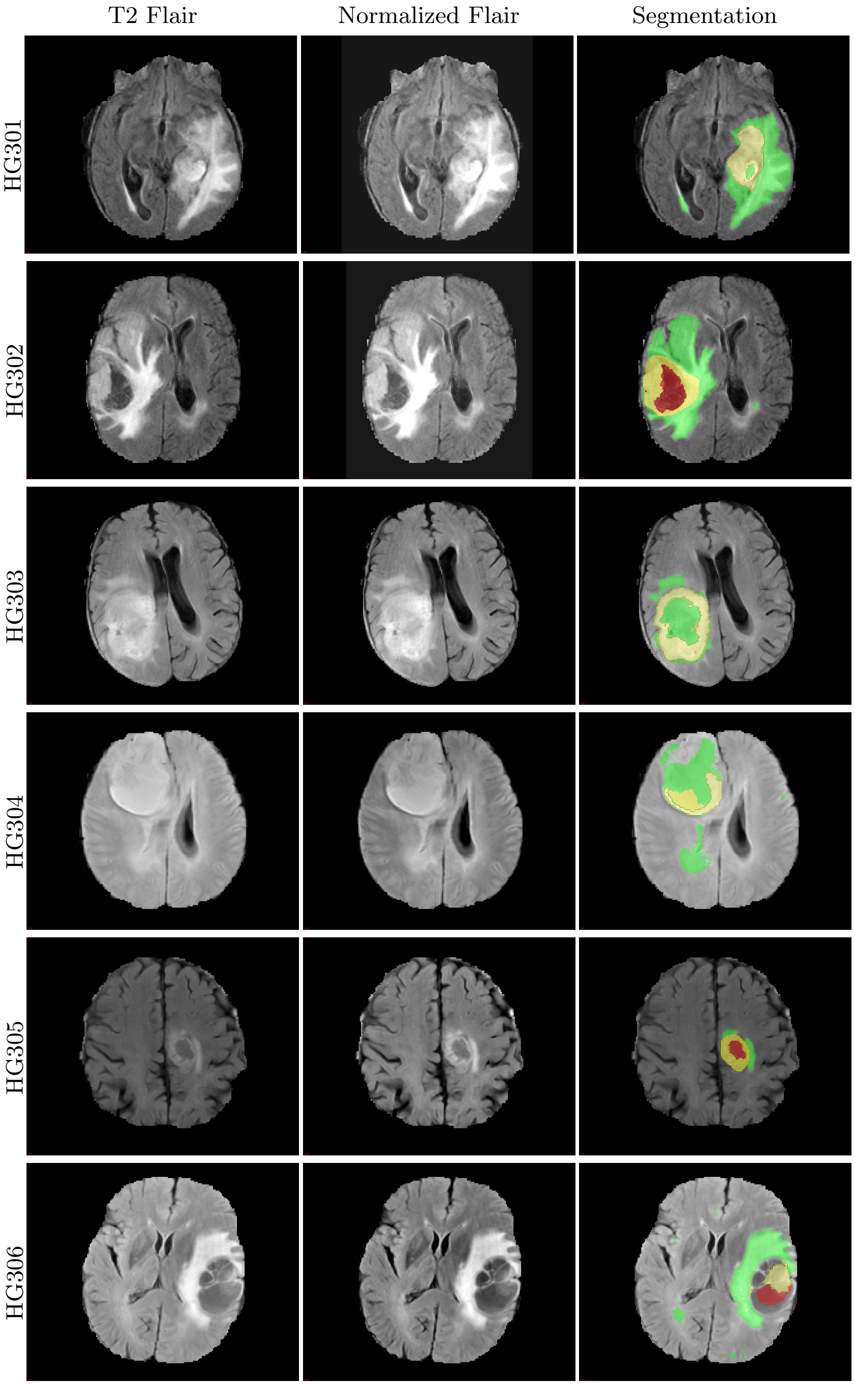


Fig. 2. Example slices from patients HG0301 to HG0306. The first column shows the original Flair image, the second the Flair image normalized with N4-Bias-Field correction and Mode-normalization. The last column shows the ally received results. The color coding is: green: 'edema', yellow: 'active tumor', red: 'necrosis'

Table 1. DICE score for the single test data sets.

Dataset	Complete tumor	Tumor core	Enhancing tumor
HG0301	0.85	0.87	0.79
HG0302	0.83	0.74	0.85
HG0303	0.86	0.78	0.74
HG0304	0.75	0.63	0.53
HG0305	0.88	0.73	0.69
HG0306	0.82	0.58	0.63
HG0307	0.81	0.47	0.48
HG0308	0.89	0.89	0.66
HG0309	0.75	0.50	0.68
HG0310	0.88	0.86	0.80
mean:	0.83±0.048	0.71±0.144	0.68±0.113

4 Discussion

We present a new approach for multi-modal brain tumor segmentation using ExtraTrees instead of Random Decision Forests and tested it using the BraTS 2013 test data. The performance of the approach is comparable to the quality of other state-of-the-art algorithms which had been tested against the same dataset. This shows that ExtraTrees are well suited for the classification of tumorous brain tissue. In the future, it will be interesting to find out whether other approaches can be improved by simply replacing Random Decision Forest classifiers with ExtraTrees.

4.1 Acknowledgments

This work was carried out with the support of the German Research Foundation (DFG) within project I04, SFB/TRR 125 Cognition-Guided Surgery.

References

[1] Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., ... and Shotton, J.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). 2014

[2] Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., Thomas, O.M, Das, T., Jena, R and Price, S.J.: Decision Forests for Tissue-Specific Segmentation of High-Grade Gliomas in Multi-channel MR. In: *Proceedings of MICCAI 2012*

[3] Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees. In: *Machine Learning*, 2006

[4] Breiman, L.: Random Forest. In: *Machine learning*, 2001

[5] Tustison N.J., Avants B.B., Cook P.A., Zheng Y., Egan A., Yushkevich P.A. and Gee J.C.: N4ITK: improved N3 bias correction. In: *IEEE Trans Med Imaging.*, 2010

[6] Nyl L.G., Udupa J.K. and Zhang, X. New variants of a method of MRI scale standardization. In: *IEEE Transaction on Medical Imaging*, 2000

[7] Haralick, R.M. Statistical and Structural Approaches to Texture In: *Proceedings of the IEEE*, 1979

[8] Beare R. Histogram-based Thresholding
In: *http://www.kitware.com/source/home/post/54*, 2012

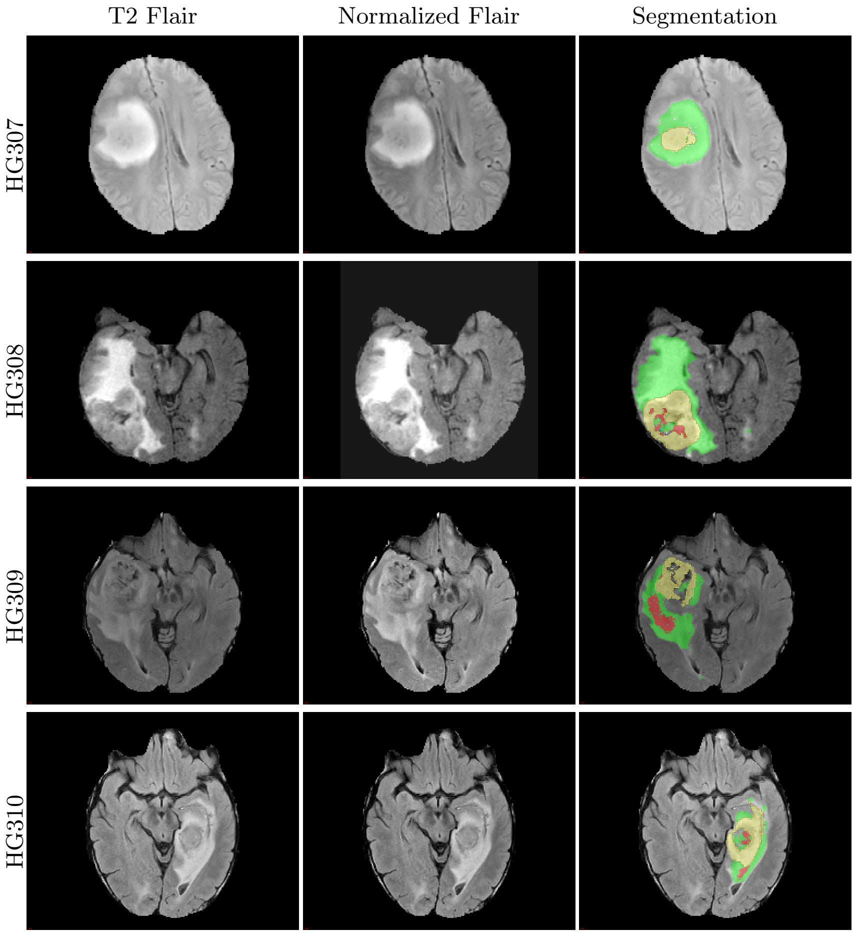


Fig. 3. Example slices from patients HG0307 to HG0310. The first column shows the original Flair image, the second the Flair image normalized with N4-Bias-Field correction and Mode-normalization. The last column shows the ally received results. The color coding is: green: 'edema', yellow: 'active tumor', red: 'necrosis'

ilastik for Multi-modal Brain Tumor Segmentation

Jens Kleesiek[1,2,3], Armin Biller[1,3], Gregor Urban[2], Ullrich Köthe[2], Martin Bendszus[1], and Fred Hamprecht[2]

[1] Division of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany
[2] Heidelberg University HCI/IWR, Heidelberg, Germany
[3] Division of Radiology, German Cancer Research Center, Heidelberg, Germany
kleesiek@uni-heidelberg.de

Abstract. We present the application of *ilastik*, the open source interactive learning and segmentation toolkit, for brain tumor segmentation in multi-modal magnetic resonance images. Even without utilizing the interactive nature of the toolkit, we are able to achieve Dice scores comparable to human inter-rater variability and are ranked in the top-5 results for the BraTS 2013 challenge data set, where no ground truth is publicly available. As careful intensity calibration is crucial for discriminative models, we propose a cerebrospinal fluid (CSF) normalization technique for pre-processing, which appears to be robust and effective. Further, we evaluate different post-processing methods for the random forest (RF) predictions obtained with *ilastik*.

Keywords: Multi-modal MRI, Brain tumor segmentation, BraTS challenge

1 Introduction

Segmenting brain tumors from multi-modal imaging data is a very challenging medical image analysis task due to the fact that magnetic resonance imaging (MRI) is usually not quantitative and lesion areas are mostly defined through intensity changes relative to surrounding normal tissue. Furthermore, the task is complicated by partial volume effects and various artifacts, e.g. due to the inhomogeneities of the magnetic field or motion of the patient during the examination. Hence, it is not surprising that even manual segmentations by experts exhibit significant intra- and inter-rater variability, which is estimated to be up to 20 % and 28 %, respectively [8].

The state-of-the-art brain tumor segmentation methods can roughly be divided in discriminative and generative approaches. For a comprehensive recent overview please see Menze et al. [9]. In general, the task of a discriminative method is to perform a tissue classification of unseen data, based on the raw data and voxel-wise or regionally extracted features. For training, supervised approaches usually rely on labels that were assigned by human expert raters and are considered to resemble ground truth. In the current study, we mostly follow this canonical approach, but introduce important variations during pre- and post-processing (see Sec. 2). The core of the proposed segmentation pipeline is *ilastik*[4] that allows predictions in close to real time [10]. The generic framework of *ilastik* has been used successfully in different domains, e.g [6, 7]. Instead of exploiting the intended usage of *ilastik*, i.e. interactive machine learning via a convenient graphical user interface, we non-interactively generate project files with random labels drawn from the annotated training data and then use the pixel classification workflow in batch prediction mode for training and prediction. The pixel classification workflow is based on a random forest (RF) classifier [3]. Although possible, user interaction beyond pre-recorded groundtruth- and CSF-labeling (see below) is not required. The proposed pipeline achieves accuracies comparable to human raters and, at the time of writing, is ranked in the top-5 of all submitted results for the BraTS 2013 challenge data set.

In this workshop paper we elucidate the proposed method in detail (Sec. 2), report (Sec. 3) and discuss (Sec. 4) the results achieved for the BraTS 2013 training and challenge data set [9].

2 Materials and Methods

2.1 Data

We use the BraTS 2013 training and challenge data set provided via the Virtual Skeleton Database (VSD) [5]. The synthetic data was excluded, because it i) was not evaluated in the 2013 challenge and ii) the synthetic data sets "are less variable in intensity and less artifact-loaded than real images" [9].

The data stems from MR scanners of different vendors and with different field strengths. It comprises co-registered native and contrast enhanced T1-weighted images, as well as T2-weighted and T2-FLAIR images. The images contain low grade (LG) and high grade (HG) tumors. For a detailed description please see Menze et al. [9].

2.2 Pre-processing

The pre-processing comprises two steps. First we employ histogram normalization as implemented by the *HistogramMatching* routine of 3D-Slicer[5]. As reference images we used the four different modalities of an arbitrary data set (HG0001). To exclude the background during matching, all voxels whose grayscale values were smaller than the mean grayscale value were excluded. Next, we normalized each individual modality with the mean value of the CSF. To obtain these values we interactively trained *ilastik* with ten randomly chosen data sets from the training set. This two class classification (CSF vs. rest) is a fairly easy task, because CSF exhibits an unambiguous combination of intensity values in the multi-modal images (dark in T1, T1c and FLAIR but bright in T2). The effect of this proposed two-step normalization technique can be seen in Fig. 1.

[4] `https://github.com/ilastik`

[5] `http://www.slicer.org`

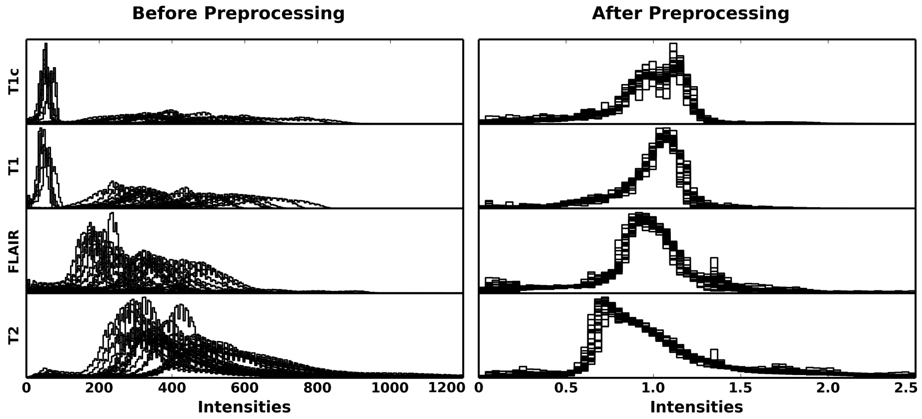


Fig. 1. Effect of the proposed two-step normalization technique. On the left side histograms of the raw intensity values of the BraTS 2013 training set (LG and HG, $N = 30$) are plotted separately for each modality. The right side shows the histograms after normalization with CSF.

After normalization we augmented the four base sequences by subtracting each modality from every other. In combination with the original four images this yields a stack of ten volumes that consecutively are used for voxel-wise feature computation. For each channel we calculated the Laplacian of Gaussian (scale 1.0), the structure tensor eigenvalues (scale 1.6) and the Hessian of Gaussian eigenvalues (scale 1.6), as implemented in the *ilastik* feature selection applet.

2.3 Pixel Classification

The *ilastik* project consists of three core software libraries: *volumina*, *lazyflow* and *ilastik*. *Lazyflow* provides threading utilities for distributing concurrent workloads across multiple cores. To achieve close to real time computations in interactive mode, this library ensures, that only computations are preformed that are strictly required to produce an output for the actually displayed data. Visualization of the multi-dimensional data, that possibly can be larger than RAM, is realized with *volumina*. These two frameworks are then orchestrated to an integrated software tool via the *ilastik* library.

Pixel classification is one of the available workflows. It relies on ten random forests with 10 trees each that are trained in parallel and eventually are merged into a single forest. Gini impurity is used as a split criterion and the number of randomly chosen features at each split is proportional to the square root of the total number of features.

To use *ilastik* in an automatic fashion, we created project files off-line. For each of the four tumor classes (edema, enhancing, non-enhancing and necrosis) up to 200 training samples, i.e. multi-dimensional feature vectors, were randomly chosen from the provided ground truth labels of every training data set. Another

1000 random samples were taken from the normal tissue of each training data set. Further, we introduced 'air' as an additional class that was granted an additional 20 labels. Different classifiers were trained for LG and HG tumors.

2.4 Post-processing

For post-processing we evaluated different strategies with increasing computational costs. In the simplest case we use simple Gaussian smoothing to clean-up the RF predictions. A more sophisticated approach relies on a guided filter as proposed by He et al. [4]. This is an edge-preserving filter that does not suffer from gradient reversal artifacts as for instance a bilateral filter and it can be computed in linear time. We also employ graph-cut optimization via the α-expansion algorithm [2] to adjust the labels. For this purpose we transformed the pseudo-probabilities P of the RF into unary potentials:

$$U(\mathbf{x}) = -\log(P(\mathbf{x})) \; . \tag{1}$$

If the labels of two variables differ we assign a cost of $c = 0.4$. The computations are realized with the *OpenGM* library [1].

A common downstream processing of the labels consists of identifying connected components (CC) and discarding all those that are < 3000 voxels. This is realized with the VIGRA library[6].

2.5 Evaluation of the Results

For comparison of the predicted segmentations we computed different standard measures, with an emphasize on the Dice coefficient as suggested in Menze et al. [9]. This metric characterizes the voxel-wise overlap of two segmented regions, by normalizing the number of true positives with the average size of the two regions. To evaluate the performance on the BraTS 2013 training data we performed leave-one-out cross-validation (LOO-CV) and used the Comparison and Validation of Image Computing (COVALIC) toolkit[7] to obtain the comparison metrics. This toolkit is also used by the challenge organizers for the evaluation. The challenge data, for which no ground truth is publicly available, was evaluated through the challenge website[8].

3 Results

Results for the LOO-CV of the training data are summarized in Tab. 1, for the challenge data in Tab. 2. For a description of the different post-processing methods please see Sec. 2.4.

[6] `https://github.com/ukoethe/vigra`

[7] `https://github.com/InsightSoftwareConsortium/covalic`

[8] `http://www.virtualskeleton.ch`

Table 1. Dice scores for BratTS 2013 training data with LOO-CV

Method	whole	LG/HG	core	LG/HG	active
Human Rater [9]	85	84/88	75	67/93	74
ilastik	75	73/76	60	58/61	65
ilastik + CC	80	78/81	64	60/66	69
ilastik + Gaussian Smoothing + CC	84	82/84	68	61/71	72
ilastik + Guided Filter + CC	83	81/84	68	61/72	71
ilastik + OpenGM + CC	83	81/84	67	61/70	72

Table 2. Dice scores for BratTS 2013 challenge data (only HG)

Method	whole	core	active
Best 2013	87	78	74
Current Best	**92**	**79**	**76**
ilastik + OpenGM + CC	87	76	74

4 Discussion

Our results (Tab. 2) on the 2013 challenge data set are comparable to the inter-rater variability reported for the BraTS data [9]. At the time of writing they are ranked in the top-5 of all submitted results. On the training data we perform slightly worse (rank 7). This might be explained by the fact that we omitted the synthetic data, for which higher Dice scores were reached as for similar real data [9].

In contrast to most methods reported in Menze et al. [9], we do not perform a bias field correction with N4ITK [11] during pre-processing, because it did not improve our result on the training data. Instead, we propose to perform intensity normalization with the mean CSF value, which proved to be a robust and effective technique (Fig. 1).

The evaluation of the different post-processing methods on the training set with LOO-CV (Tab. 2) shows the added value of "cleaning-up" the RF predictions. The three different methods used, exhibit a similar performance but come at different computational costs. Especially, simple Gaussian smoothing is a fast and effective method.

Looking at our segmentations in detail, we noticed the presence of 'holes', which –according to our predictions– correspond to islands of healthy neuronal tissue. From a neuro-oncological point of view this is plausible and can not be ruled out per se. However, due to the labeling instructions for the experts [9], it is not very likely that those kind of islands occur in the ground truth data. Primarily aiming at an interactive clinical workflow, we decided not to fill these holes with a computational method, which supposedly would improve our challenge results further.

Future work aims at integrating the insights obtained during the challenge into an *ilastik* workflow that can be easily deployed in clinical routine and for clinical trials.

Acknowledgments. We thank Thorsten Beier, Christoph Decker, Markus Döring, Burçin Eröcal, Carsten Haubold and the entire ilastik team for technical help and valuable comments. This work was supported by a postdoctoral fellowship from the Medical Faculty of the University of Heidelberg.

References

1. Andres, B., Beier, T., Kappes, J.H.: OpenGM: A C++ library for discrete graphical models. ArXiv e-prints (2012), `http://arxiv.org/abs/1206.0111`
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. IEEE Trans. Pattern Anal. Mach. Intell. 23(11), 1222–1239 (Nov 2001), `http://dx.doi.org/10.1109/34.969114`
3. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
4. He, K., Sun, J., Tang, X.: Guided image filtering. IEEE Trans Pattern Anal Mach Intell 35(6), 1397–409 (Jun 2013)
5. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P.: The virtual skeleton database: an open access repository for biomedical research and collaboration. J Med Internet Res 15(11), e245 (2013)
6. Kreshuk, A., Koethe, U., Pax, E., Bock, D.D., Hamprecht, F.A.: Automated detection of synapses in serial section transmission electron microscopy image stacks. PLoS One 9(2), e87351 (2014)
7. Kroeger, T., Mikula, S., Denk, W., Koethe, U., Hamprecht, F.A.: Learning to segment neurons with non-local quality measures. Med Image Comput Comput Assist Interv 16(Pt 2), 419–27 (2013)
8. Mazzara, G.P., Velthuizen, R.P., Pearlman, J.L., Greenberg, H.M., Wagner, H.: Brain tumor target volume determination for radiation treatment planning through automated mri segmentation. Int J Radiat Oncol Biol Phys 59(1), 300–12 (May 2004)
9. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), submitted to IEEE Transactions on Medical Imaging
10. Sommer, C., Straehle, C., Koethe, U., Hamprecht, F.A.: "ilastik: Interactive learning and segmentation toolkit". In: 8th IEEE International Symposium on Biomedical Imaging (ISBI) (2011)
11. Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., Gee, J.C.: N4itk: improved n3 bias correction. IEEE Trans Med Imaging 29(6), 1310–20 (Jun 2010)

Multimodal Brain Tumor Image Segmentation Using GLISTR

Dongjin Kwon, Hamed Akbari, Xiao Da, Bilwaj Gaonkar,
and Christos Davatzikos

Center for Biomedical Image Computing and Analytics, University of Pennsylvania

Abstract. In this paper, we summarize our approach to the brain tumor segmentation challenge (BRATS). Our method, called GLISTR, is a joint segmentation and registration method for brain tumors. Using this method, we simultaneously segment brain scans and register these scans to a normal atlas. We grow tumors from tumor seed points using a tumor growth model and modify a normal atlas into on with tumors and edema. We then estimate the mapping between the modified atlas and the scans, posteriors for each tissue labels, and the tumor growth model parameters via an EM framework. We apply GLISTR to the BRATS 2013 data set to evaluate segmentation performances.

1 Introduction

Segmenting brain tumors is a challenging problem due to the complex shapes of the pathology and their heterogenous textures. Also, multifocal masses of such tumor make this problem even more difficult. We solve this problem by our GLioma Image SegmenTation and Registration method (GLISTR), firstly introduced in [1] and later conceptually improved in [3]. Using GLISTR, we could segment multifocal tumors using multiple tumor growths and estimate complex appearances of tumors using tumor shape priors. As we label the entire brain region using registered tissue priors, the segmentation of pathological regions is complemented by that of healthy regions.

2 Methods

Our method generates a patient-specific atlas by embedding tumors on a normal atlas using a tumor growth model [2]. For multifocal tumors, we use multiple tumor seeds and grow a tumor on each seed, and then combine grown tumors into the single tumor probability map. The normal atlas is modified into on with tumors and edema using this tumor probability map. We also generate a tumor shape prior using the random walk with restart which uses multiple tumor seeds as initial labels. We incorporate the tumor shape prior into an EM framework via empirical Bayes model. Using this framework, we simultaneously estimate the mapping between the patient-specific atlas and input scans, posteriors for each tissue labels, and the tumor growth model parameters. More detailed procedures are described in [3].

Table 1. BRATS 2013 Results.

	Dice			PPV			Sensitivity		
Data Set	whole	core	active	whole	core	active	whole	core	active
Leaderboard	0.86	0.79	0.59	0.88	0.84	0.60	0.86	0.81	0.63
Challenge	0.88	0.83	0.72	0.92	0.90	0.74	0.84	0.78	0.72

3 Results

Our method requires minimal user initializations including seed points and radius for each tumor and one sample point for each tissue class. Users could use the visual interface of GLISTR to easily mark each point. For preprocessing, we coregistered all four modalities (T1, T1-CE, T2, and FLAIR), corrected MR field inhomogeneity, and scaled intensities to fit [0, 255]. We tested our method to the BRATS 2013 data set via the BRATS online tools [4]. The leaderboard data set consists of 21 high-grade and 4 low-grade glioma subjects and the challenge data set consists of 10 high-grade glioma subjects. The results are shown in Table 1. The performance measures include Dice scores, positive predictive value (PPV), and sensitivity for three interest regions: whole(complete abnormal regions including tumor and edema), core (tumor regions), and active (enhancing regions of tumor). Our method showed the top performances among participants and especially performed well on estimating tumor core regions. The average running time of our method was 85 min on an Intel Core i7 3.4 GHz machine with Windows operating system.

References

1. Gooya, A., Pohl, K.M., Billelo, M., Cirillo, L., Biros, G., Melhem, E.R., Davatzikos, C.: GLISTR: Glioma Image Segmentation and Registration. IEEE Trans. Med. Imaging 31(10), 1941–1954 (2012)
2. Hogea, C., Davatzikos, C., Biros, G.: An image-driven parameter estimation problem for a reaction-diffusion glioma growth model with mass effects. J. Math. Biol. 56(6), 793–825 (2008)
3. Kwon, D., Shinohara, R.T., Akbari, H., Davatzikos, C.: Combining Generative Models for Multifocal Glioma Segmentation and Registration. In: Med. Image Comput. Comput. Assist. Interv. (MICCAI). pp. 763–770 (2014)
4. Menze, B.H., et al.: The BRATS Online Tools - Multimodal Brain Tumor Segmentation (BRATS 2013). http://www.virtualskeleton.ch/BRATS/Start2013

Appearance- and Context-sensitive Features for Brain Tumor Segmentation

Raphael Meier[1], Stefan Bauer[1,2], Johannes Slotboom[2], Roland Wiest[2], and Mauricio Reyes[1]

[1] Institute for Surgical Technologies and Biomechanics, University of Bern
[2] Inselspital, Bern University Hospital, Switzerland
raphael.meier@istb.unibe.ch

Abstract. The proposed method for fully-automatic brain tumor segmentation builds upon the combined information from image appearance and image context. We employ a variety of different feature types to capture this information. Based on these features, a decision forest performs voxel-wise tissue classification followed by a spatial regularization via a conditional random field. Our method was evaluated on two data sets of the BRATS 2013 challenge achieving high performance within a reasonable average computation time of 5 minutes per subject.

1 Introduction

Current clinical guidelines (e.g. RANO/AVAGlio [3]) rely on manual, bidimensional measures for response assessment of malignant gliomas. In a recent publication [12], it was shown that such measurements are highly sensitive to MRI head placement. As a more reliable alternative 3D tumor volumetry was proposed. Manual tumor segmentation is time-consuming and subject to observer bias [5]. Hence, fully-automatic brain tumor segmentation methods are desired, reducing these issues.

A majority of the current best performing methods rely on techniques from machine learning [1, 4]. A major insight we obtained through our participation in previous segmentation challenges is that the representation of the input data, generally referred to as *features*, plays a crucial role in machine learning-based segmentation models. Thus, our present approach is driven by an extensive set of different features capturing different aspects of the input data.

2 Preliminaries

Structural MRI. Our approach relies on four different MRI sequences that are routinely used in clinical acquisiton protocols, namely T_1-, T_{1c}- (post-contrast), T_2-, $FLAIR$-weighted images. We regard the entire four MR sequences as a multi-sequence image Ω.

Classification. We pose the problem of brain tumor segmentation as a voxel-wise classification problem. Thus, we seek a hypothesis h that relates a voxel, represented by its feature vector $\mathbf{x}$, to a corresponding tissue (class) label y (i.e. $h(\mathbf{x}) : \mathbf{x} \to y$). We consider seven possible tissue classes: three unaffected (gray matter, white matter, csf) and four tumor tissues (necrosis, edema, enhancing and non-enhancing tumor). Based on a given fully-labeled training set $\mathcal{S} = \left\{\left(\mathbf{x}^{(i)}, y^{(i)}\right) : i = 1, ..., |\mathcal{S}|\right\}$ we estimate h (supervised learning).

3 Methods

The present method builds on the insights and developments of two previously published approaches [2, 8]. In [2] the original formulation of the algorithm that is still valid was proposed. In [8] it was extended to a generative-discriminative hybrid model. The present method abandons the generative part and instead relies on an enhanced feature set leading to an increased performance with reduced computation time (compared to [8]). The pipeline is depicted in figure 1. After preprocessing (smoothing, intensity normalization, bias-field correction) of an image $\Omega^{(j)}$, we extract appearance- and context-sensitive features. A classification forest is employed to provide a voxel-wise tissue classification ($\tilde{y}$) that is subsequently refined by a spatial regularization.

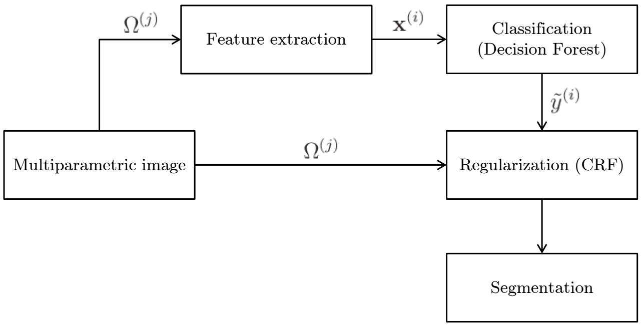


Fig. 1: Segmentation pipeline. After the multi-sequence image has been preprocessed, voxel-wise features are extracted, followed by classification and subsequent spatial regularization.

3.1 Appearance-sensitive features

Appearance-sensitive features try to capture contrast information. These features profit directly from the usage of multiple different MR sequences and encompass the voxel-wise intensity values, first-order texture features and gradient features. The first-order texture information is contained in the histogram of an image (or image region). We extract them over box-shaped Moore neighborhoods varying in size (containing either 3^3, 5^3 or 7^3 voxels). In addition, we generate gradient magnitude images of each respective MR sequence image and extract local mean and variance over the same neighborhoods.

Furthermore, we investigated the use of second-order texture features (extracted from intensity-based co-occurence matrix). Since their usage did not lead to any improvement, we discarded them from the final feature set. At this point, one could argue to also include features that characterize the shape of a tumor. However, given the enormous variability (especially when considering the tumoral subcompartments) of this aspect, we decided to not include any notion of shape as a feature.

3.2 Context-sensitive features

Gliomas can occur everywhere in the brain. Nevertheless, it is unlikely that they arise in the cerebellum or brainstem, i.e. the infratentorial part of the brain. We target to capture this cue with the help of an atlas image. We register the T_{1c}-weighted patient image to the atlas image employing an affine transformation. Prior to this step, all other MR sequences have been rigidly registered to the T_{1c}-weighted image. After registration of the patient image to the atlas, we obtain for every voxel i in the patient image its corresponding (physical) coordinates in the atlas image $\{x_i, y_i, z_i\}$, which we refer to as atlas-normalized coordinates. We use the term "normalized" since all training and testing images are transformed into the same atlas coordinate system. Since we are only interested in a rough estimate of the respective location in the atlas (e.g. is the position of the voxel supra- or infratentorial?), we smooth the final atlas-normalized coordinates using a Gaussian kernel ($\sigma = 1.5$).

The spatial arrangement of different tumor subcompartments in case of gliomas (especially Glioblastomas) is characterized through a more or less well-defined order of layers (at least if we are working with the present definition of four tumor subcompartments). If we consider the T_1- and $FLAIR$-weighted images in figure 2, we can recognize that in the T_1 certain parts (e.g. necrotic core) are hypointense, whereas in the $FLAIR$ they appear hyperintense. Thus, the dynamic range of intensity values given both modalities is in general larger than for healthy tissue. Our basic idea is to capture this information with the following procedure:

1. For a voxel i send out four (in-plane) rays of length d with an angle α, where $d \in \{10, 20\}$ (in voxels) and $\alpha \in \{0°, 90°, 180°, 270°\}$.
2. For every ray construct the histogram H using intensity values from T_1 *and* $FLAIR$ images.
3. Compute the range of the histogram: $r = H_{max} - H_{min}$, where H_{max} and H_{min} are the maximum and minimum (occupied) intensity bins of the histogram.
4. Compute the mean range $\bar{r}$ of the four rays.

The mean range $\bar{r}$ is then used as final feature which we simply call *ray feature*. By working with histograms our features are invariant against small shifts. The reason why we restricted ourselves to rays casted in-plane and not out-of-plane is that the slice thickness can vary greatly. In initial experiments, we observed that especially the classification of the necrotic core improves when proposed ray features are used. This makes sense since the necrotic part of the tumor appears hypointense in T_1-weighted images and is typically surrounded by active tumor which is hyperintense in $FLAIR$ images.

Finally, we employ symmetric intensity differences which capture asymmetries across the brain hemispheres induced by the tumor. The axis of symmetry is defined as the midsagittal plane of the previously registered atlas. For increasing the robustness of the symmetric features, we smooth the images with a Gaussian kernel ($\sigma = 3.0$) before extracting them.

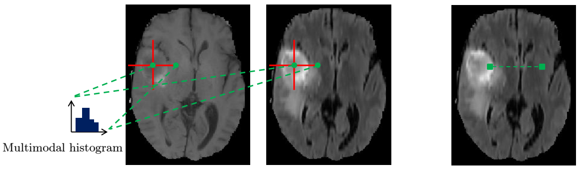


Fig. 2: Ray feature (left) and symmetry feature (right).

Besides the previously described features, we investigated the use of two other feature types: Context-rich features [6] and Local Binary Patterns [10]. However, we did not observe a statistically significant improvement when employing these features. Consequently, we discarded them from our final feature set. In the end, we obtain a 237-dimensional feature vector $\mathbf{x}$.

3.3 Classification Forest

For classification, we employ a decision forest (which we used extensively in other work [2, 8, 9]). The classification forest is trained on the fully-labeled training set $\mathcal{S}$. Important to notice is that we rely on axis-aligned weak learners as split functions and simple class-histograms as prediction models (stored in leafs). The predicted class label is defined according to the MAP-rule: $\tilde{y} = \arg\max_y p(y|\mathbf{x})$ (which corresponds to h), where the probability is generated via the class-histograms stored in the respective leaf of the decision trees.

3.4 Spatial Regularization

The spatial regularization is identical to our hierarchical approach from [2], where it is formulated as an energy minimization problem of a conditional random field (CRF) defined on a grid-graph that corresponds to the image volume. For more details, we refer the reader to [2].

4 Results

We evaluated our method on two datasets. First, the BRATS2013 training set which encompasses 30 patient images (including both high-grade and low-grade gliomas). Second, the BRATS2013 challenge data set which consists of 10 patient images bearing high-grade gliomas. Prior to the evaluation, the sequence images were rigidly registered to the T_{1c}-image and skullstripped. The model is trained either on high- or low-grade cases only. Consequently, we performed a 5-fold cross validation for the high-grade cases and a leave-one-out cross validation for the low-grade gliomas of the training set. We trained on the 20 high-grade cases of the training set to segment the challenge set. Quantitative evaluation of the segmentation results was conducted online on the Virtual Skeleton Database (VSD)[3] and is listed in table 1. The decision forest was implemented using the Sherwood library [13]. The average computation time per patient image is around 5 minutes.

Region	Dice	Jaccard	PPV	Sensitivity
Complete tumor (HGG)	0.84 ± 0.03	0.72 ± 0.04	0.8 ± 0.06	0.89 ± 0.07
Tumor core (HGG)	0.73 ± 0.14	0.59 ± 0.15	0.8 ± 0.12	0.7 ± 0.19
Enhancing tumor (HGG)	0.68 ± 0.11	0.53 ± 0.12	0.72 ± 0.11	0.7 ± 0.19
Complete tumor (HGG&LGG)	0.83 ± 0.1	0.72 ± 0.14	0.85 ± 0.09	0.83 ± 0.15
Tumor core (HGG&LGG)	0.66 ± 0.24	0.59 ± 0.24	0.74 ± 0.25	0.66 ± 0.27
Enhancing tumor (HGG&LGG)	0.58 ± 0.34	0.47 ± 0.3	0.66 ± 0.36	0.54 ± 0.35

Table 1: Results of online evaluation for cases of BRATS2013 challenge (top) and training (bottom) data set. Performance measures are given as mean values ± standard deviation.

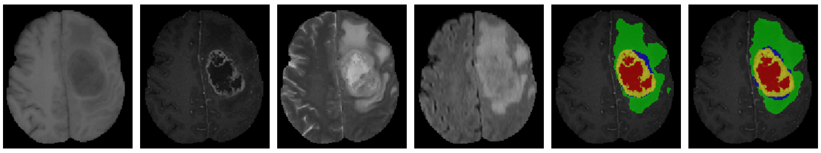


Fig. 3: Segmentation result for case HG0011. From left to right: T_1-, T_{1c}-, T_2-, $FLAIR$-weighted image, overlayed ground truth on T_{1c} image (necrotic = red, enhancing tumor = yellow, non-enhancing tumor = blue, edema = green), overlayed segmentation result of our method.

[3] https://www.virtualskeleton.ch/

5 Discussion and Conclusion

We propose a fully-automatic, machine learning-based method that builds upon the combined information from image appearance as well as context. This method is an integral part of the BraTumIA software suite, which is a clinically validated [11] tool for radiologists to perform brain tumor image analysis[4]. Clearly, the use of different features improves the performance of our method. However, we experienced that the introduction of a new type of feature does not necessarily lead to an improvement (this applies especially in the situation when the number of features is already large and their nature diverse). We think that further improvements can be obtained by a more effective use of the available training data (as e.g. proposed in [7]) rather than more advanced features.

Acknowledgments. This project has received funding from the European Union's Seventh Programme for research, technological development and demonstration under grant agreement No [600841], from the Swiss Cancer League and the Swiss National Science Foundation.

References

1. S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes. A survey of MRI-based medical image analysis for brain tumor studies. *PMB*, 58(13), 2013.
2. S. Bauer, T. Fejes, J. Slotboom, R. Wiest, L.-P. Nolte, and M. Reyes. Segmentation of Brain Tumor Images Based on Integrated Hierarchical Classification and Regularization. In *Proceedings of MICCAI-BRATS 2012*, 2012.
3. O. L. Chinot, D. R. Macdonald, L. E. Abrey, G. Zahlmann, Y. Kerloëguen, and T. F. Cloughesy. Response assessment criteria for glioblastoma: practical adaptation and implementation in clinical trials of antiangiogenic therapy. *Current Neurology and Neuroscience Reports*, 13(5), 2013.
4. B. Menze, et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). Submitted 2014.
5. M. A. Deeley, A. Chen, R. Datteri, J. H. Noble, a. J. Cmelak, E. F. Donnelly, A. W. Malcolm, L. Moretti, J. Jaboin, et al. Comparison of manual and automatic segmentation methods for brain structures in the presence of space-occupying lesions: a multi-expert study. *PMB*, 56(14), 2011.
6. E. Geremia, O. Clatz, B. H. Menze, E. Konukoglu, A. Criminisi, and N. Ayache. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. *Neuroimage*, 57(2), 2011.
7. H. Lombaert, D. Zikic, A. Criminisi, and N. Ayache. Laplacian Forests : Semantic Image Segmentation by Guided Bagging. In *Proceedings of MICCAI 2014 (In Press)*, 2014. R. Wiest, and M. Reyes
8. R. Meier, S. Bauer, J. Slotboom, et al. A Hybrid Model for Multimodal Brain Tumor Segmentation. In *Proceedings of MICCAI-BRATS 2013*, 2013.
9. R. Meier, S. Bauer, J. Slotboom, R. Wiest, and M. Reyes. Patient-specific Semi-Supervised Learning for Postoperative Brain Tumor Segmentation. In *Proceedings of MICCAI 2014 (In Press)*, 2014.

[4] http://www.nitrc.org/projects/bratumia

10. T. Ojala. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1), 1996.
11. N. Porz, S. Bauer, A. Pica, P. Schucht, J. Beck, R. K. Verma, J. Slotboom, M. Reyes, and R. Wiest. Multi-modal glioblastoma segmentation: man versus machine. *PloS one*, 9(5), 2014.
12. M. Reuter, E. R. Gerstner, O. Rapalino, et al. Impact of MRI head placement on glioma response assessment. *Journal of Neuro-oncology*, 118(1), 2014.
13. A. Criminisi and J. Shotton. Decision Forests for Computer Vision and Medical Image Analysis. Springer, 2013.

Improved Brain Tumor Tissue Segmentation Using Texture Features

S. Reza and K. M. Iftekharuddin
{sreza002, kiftekha}@odu.edu

Vision Lab, Department of Electrical and Computer Engineering,
Old Dominion University, Norfolk, VA 23529, USA.

Abstract

In this work, we obtain improved automatic brain tumor segmentation (BTS) performance based on our prior methods [1] [2]. We also statistically validate the efficacy of our improved tumor tissue segmentation methods. Despite excellent ranking of our BTS methods in BRATS-2013 challenge [4], few misclassifications in the tumor core region appeared to have compromised the overall performance. In order to lower these misclassifications, this work develops morphological filtering for post-processing of segmented tissues. Preliminary results from both BRATS-2013 and BRATS-2014 training dataset suggest that further BTS improvement may be achieved with the additional morphological filter. We further plan to obtain cross validated results using BRATS-2014 data for the final submission.

Keywords: Tumor Segmentation, Texture feature, Morphological filter, BRATS, MR.

Methods

The proposed segmentation method is an improvement over our BTS prior works [1] [2]. The improvement is obtained by carefully devising a morphological post processing technique. The overall flow diagram of the proposed method is shown in Fig. 1.

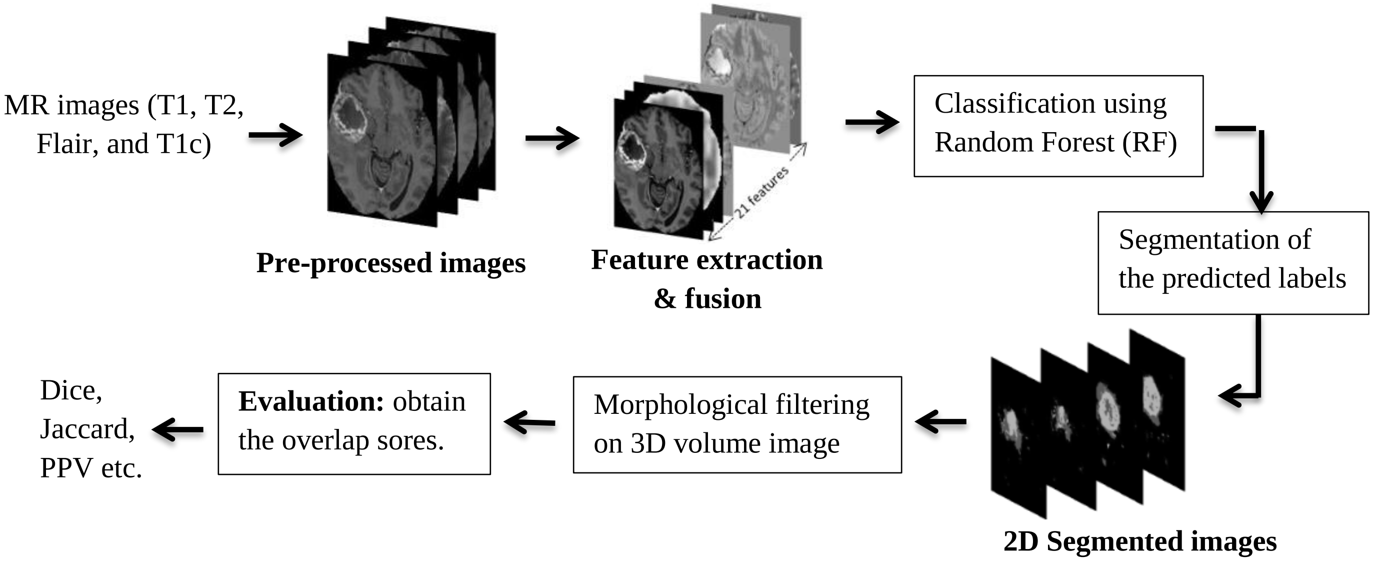

MR images (T1, T2, Flair, and T1c)
Pre-processed images
21 features
Feature extraction & fusion
Classification using Random Forest (RF)
Segmentation of the predicted labels
2D Segmented images
Morphological filtering on 3D volume image
Evaluation: obtain the overlap sores.
Dice, Jaccard, PPV etc.

Figure 1: Simplified flow diagram of the proposed method

The detail description of first three steps in Fig. 1 can be found in [1] [2]. Here is a brief overview of the complete steps:

i. Pre-processing to include bias correction [5] and MR intensity inhomogeneity correction [6].
ii. Feature extraction to include two types of features:
   a. Global: MR intensities and intensity differences among the modalities.
   b. Texture features: fractal PTPSA [7], mBm[8] , textons [9].
iii. Pixel level classification and prediction using Random Forest [10] classifier.
iv. Generating 2D segmented images from the predicted labels and then 3D volume image.
v. Post processing using two stage binary morphological filter:
   a. Stage-1: Based on the connected component, the filter keeps only the larger objects and removes the smaller objects from the 3D volume. Example of some small objects is shown with green circle in Figure-2 (b).
   b. Stage-2: Holes in the tumor core region is detected. Based on the neighbor intensities, labels are assigned in the holes region.
vi. Evaluation using final output MR volume.

Dataset

Two dataset of glioma tumors have been used

- BRATS-2013 training dataset. 20 High grade (HG) and 10 Low grade (LG)
- BRATS-2014 training dataset. 190 HG, 26 LG

Results and Discussions

From the predicted labels of each pixel, we obtain the 2D segmented images. These 2D segmented images are used for post processing using morphological filtering to obtain better pixel wise labeling. Finally, the refined images are stacked to generate the 3D volume images. Example tissue segment and improvement using the morphological filtering are shown in Figure-2.

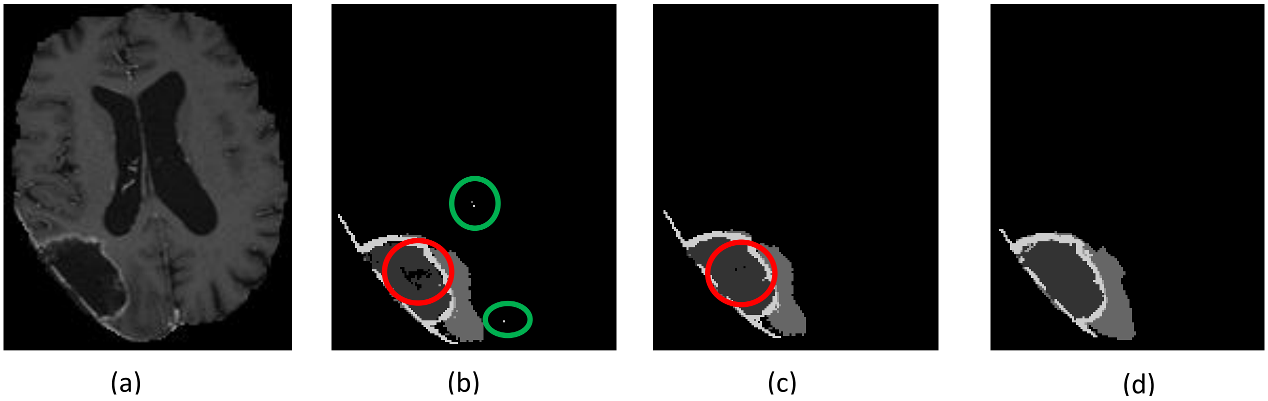
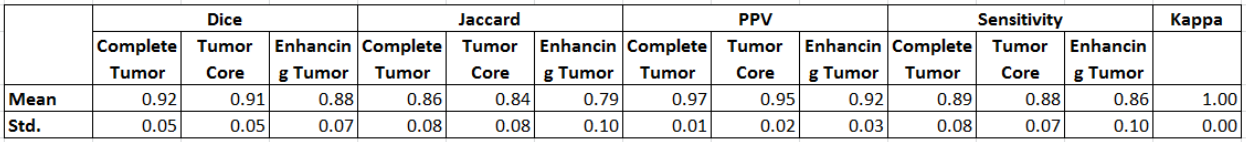


Figure 2: Segmented tissues with corresponding input and ground-truth images. (a) Corresponding T1c, (b) previous result /without filtering (c) current result with filtering, red circle shows the region of improvement (d) ground-truth. Labels in the ground-truth: 1-necrosis, 2- edema, 3-non-enhancing

Quantitative evaluation of segmented results:

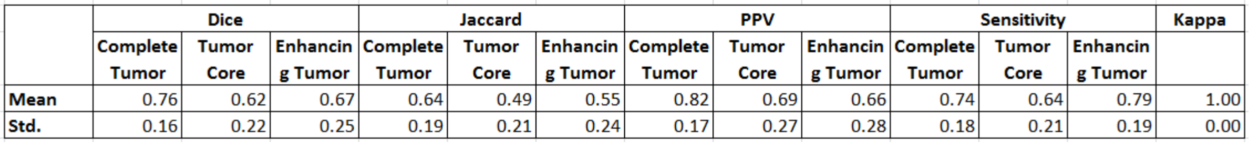
All our segmented results are evaluated according to the three different categories set up by BRATS-2013. The details on these three categories are as follows: Complete Tumor: (1-necrosis, 2-Edema, 3-non-enhancing tumor, 4-enhance tumor); Tumor Core: (3-non-enhance tumor, 4-enhance tumor); and Enhance tumor: (4-enhacne tumor). We perform 3-fold cross validation on 30 training patients of BRATS-2013, and the results are also reported in our previous submission [1]. The average scores of the 3-fold cross validated results are in Table 1. In summary, Dice overlap metric of our segmentation rate varies between 88% to 92% for enhanced tumor, tumor core, and complete tumor respectively.



	Dice			Jaccard			PPV			Sensitivity			Kappa
	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	
Mean	0.92	0.91	0.88	0.86	0.84	0.79	0.97	0.95	0.92	0.89	0.88	0.86	1.00
Std.	0.05	0.05	0.07	0.08	0.08	0.10	0.01	0.02	0.03	0.08	0.07	0.10	0.00

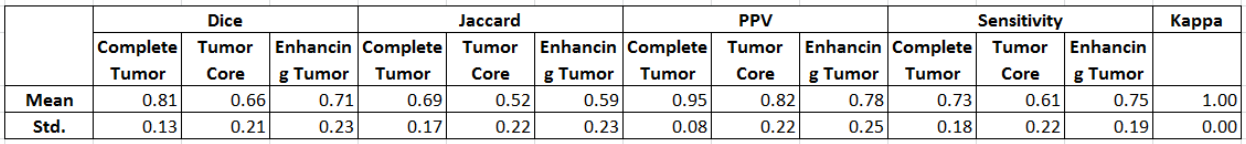
Table 1: Average results of 3-fold cross validation [1] on 30 patients of BRATS-2013.

The patient-wise cross validation results using our algorithm in Table 1 suggest that one may obtain reasonably good results for any representative patient dataset. In order to measure the robustness of the method, we use the trained RF classifier with BRATS-2013 data and test on BRATS-2014 dataset. Furthermore, we obtain significant improvement using the proposed morphological post processing. Quantitative scores of 216 training patients of BRATS-2014 with the basic algorithm [1] and the proposed method (Fig. 1) are shown in Table 2 and Table 3 respectively.



	Dice			Jaccard			PPV			Sensitivity			Kappa
	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	
Mean	0.76	0.62	0.67	0.64	0.49	0.55	0.82	0.69	0.66	0.74	0.64	0.79	1.00
Std.	0.16	0.22	0.25	0.19	0.21	0.24	0.17	0.27	0.28	0.18	0.21	0.19	0.00

Table 2: Average results of 216 patients of BRATS-2014 using the method [1]. RF classifier is trained with 20 HG patients of BRATS-2013.



	Dice			Jaccard			PPV			Sensitivity			Kappa
	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	Complete Tumor	Tumor Core	Enhancing Tumor	
Mean	0.81	0.66	0.71	0.69	0.52	0.59	0.95	0.82	0.78	0.73	0.61	0.75	1.00
Std.	0.13	0.21	0.23	0.17	0.22	0.23	0.08	0.22	0.25	0.18	0.22	0.19	0.00

Table 3: Average results of 216 patients of BRATS-2014 using the proposed method. RF classifier is trained with 20 HG patients of BRATS-2013.

Results in Table 2 and 3 show that the Dice score varies from 67% to 76% using the method [1], and from 71% to 81% using the proposed method respectively. From the patient-wise results we notice that the propose algorithm usually performs better on High grade (HG) tumors than Low grade (LG). Therefore, we observe that the MRI containing HG tumor surface may contain higher randomness in texture. Furthermore, the morphological filter is especially developed to reduce the misclassification of necrosis tissues in the core region. As the necrosis tissues are commonly found in HG tumors, the morphological filter improves the segmentation results of HG.

Conclusion

In this work, we have investigated the efficacy of our proposed method, which is our basic automatic segmentation method [1] followed by a post-processing morphological filter. Preliminary results of 246 glioma patients confirm the efficacy of the proposed method (Fig. 1). However, generalization of such morphological filter is challenging and need more investigation. Preliminary results from both BRATS-2013 and BRATS-2014 training dataset suggest that further BTS improvement may be achieved with the additional morphological filter. We further plan to obtain cross validated results using BRATS-2014 data for the final submission.

Acknowledgements

This work is partially supported through a grant from NCI/NIH (R15CA115464). Brain tumor image data used in this work were obtained from the NCI-MICCAI 2013 & 2014 Challenge on Multimodal Brain Tumor Segmentation (http://martinos.org/qtim/miccai2013/index.html) organized by K. Farahani, M. Reyes, B. Menze, E. Gerstner, J. Kirby and J. Kalpathy-Cramer . The challenge database contains fully anonymized images from the following institutions: ETH Zurich, University of Bern, University of Debrecen, and University of Utah and publicly available images from the Cancer Imaging Archive (TCIA)

References

[1] S. Reza and K. M. Iftekharuddin, "Multi-class abnormal brain tissue segmentation using texture features," *in Proceedings MICCAI-BRATS*, pp. 38-42, 2013.

[2] S. Reza and K. M. Iftekharuddin, “Multi-fractal texture features for brain tumor and edema segmentation,” *Proc. SPIE Med. Imag. Conf.*, vol. 9035, 2014.

[3] https://vsd.unibe.ch/WebSite/BRATS/Start2013/

[4] https://sites.google.com/site/miccaibrats2014/

[5] N. Tustison and J. Gee. N4ITK: Nick's N3 ITK implementation for MRI bias field correction. The Insight Journal, 2010.

[6] L. G. Nyul, J. K. Udupa, and X. Zhang, "New variants of a method of MRI scale standardization," IEEE Transaction on Medical Imaging, vol. 19, no. 2, pp. 143-150, 2000.

[7] S. Ahmed, K. Iftekharuddin, and A. Vossough, "Efficacy of texture, shape, and intensity feature fusion for posterior-fossa tumor segmentation in MRI," *IEEE Transactions on Information Technology in Biomedicine*, pp. 206-213, 2011.

[8] A. Islam, S. Reza, and K. M. Iftekharuddin, "Multi-fractal texture estimation for detection and segmentation of brain tumors," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 11, pp. 3204-15, 2013.

[9] T. Leung and J. Malik, "Representing and recognizing the visual appearance of materials using three-dimensional textons," *International Journal of Computer Vision*, vol. 43, no. 1, pp. 29 – 44, 2001.

[10] A. Cutler and L. Breiman, "Random forests-classification description," *Technical report, University of California, Berkeley*, 2004.

Multi-modal Brain Tumor Segmentation using Deep Convolutional Neural Networks

Gregor Urban[2], Martin Bendszus[1], Fred Hamprecht[2], and Jens Kleesiek[1,2,3]

[1] Division of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany
[2] Heidelberg University HCI/IWR, Heidelberg, Germany
[3] Division of Radiology, German Cancer Research Center, Heidelberg, Germany
kleesiek@uni-heidelberg.de

Abstract. We present the application of 3D-Convolutional Neural Networks for brain tumor segmentation in multi-modal magnetic resonance images. We are able to achieve Dice scores comparable to human interrater variability and are ranked among the top-scoring submission for the BraTS 2013 challenge data, where no ground truth is publicly available. As careful intensity calibration is crucial for discriminative models, we rely on a cerebrospinal fluid (CSF) normalization technique for preprocessing.

Keywords: Multi-modal MRI, Brain tumor segmentation, BraTS challenge, Convolutional Neural Network

1 Introduction

The majority ($\approx 70\,\%$) of primary cerebral malignancies originate from glial cells. Amongst those, the most frequent malignant primary brain tumor in humans, *glioblastoma multiforme* (GBM), is accompanied by rapid infiltrative growth and a very poor prognosis. This is reflected by an average survival time of about one year after diagnosis [8]. The overall survival rate of patients suffering from GBM is affected by a combination of extensive treatment strategies such as concomitant radio- and chemotherapy and/or surgical resection [8]. The gold standard to account for tumor growth in daily clinical routine is guided by the Response Assessment in Neuro-Oncology (RANO) criteria [9]. These guidelines only comprise surrogate measures (e.g. maximal 2D diameter of the contrast enhancing portion of the lesion) to estimate the development of the malignancy. For diagnosis, treatment planing and monitoring it is thus desirable and very important to have reliable and reproducible segmentation methods available that are able to quantify not only the whole tumor volume but also the volume of sub-regions of the mass, like non-enhancing portions and edema.

As human experts compare the texture and intensities of different MRI channels in order to rate the signal alterations, we trained a 3D-Convolutional Neural Network (CNN) to mimic this procedure. Using The CNN, we achieve accuracies comparable to human raters for the whole tumor and active core sub-regions. At the time of writing, our method is ranked second of all previously submitted results for the BraTS 2013 challenge data set.

2 Materials and Methods

2.1 Data

We use the BraTS 2013 training and challenge data set provided via the Virtual Skeleton Database (VSD) [4]. The synthetic data was excluded, because it is less variable in intensity and contains fewer artifacts than real data. Furthermore, higher Dice scores were reported for the synthetic data sets [7].

The data stems from MR scanners of different vendors and with different field strengths. It comprises co-registered native and contrast enhanced T1-weighted images, as well as T2-weighted and T2-FLAIR images. The images contain low grade (LG) and high grade (HG) tumors. For a detailed description please refer to Menze et al. [7].

We employ the same two-step pre-processing as described in [5], which comprises a normalization with the mean CSF value. However, in contrast to the other approach, we do not "augment" the data set with the differences between the channels, and thus only use the four canonical MRI channels as input for the CNN.

2.2 The Voxel-wise Classifier

We tackle the segmentation problem by applying a voxel-wise classifier on the data. Predictions are based on local information provided by small 3D patches, one for each input channel. These cubes of voxels are fed into the classifier, which then predicts the voxel(s) in the center of the cube. As we employ a Convolutional Neural Network (e.g. [6]) for this task, we can easily control the number of input voxels that are used for predicting the class of one voxel by changing the number of layers or the sizes of the convolutional filters of the network. Our Convolutional network uses 3D spatial convolutions instead of the usual 2D layout used in image classification. The data has three spatial dimensions (x,y,z) and one dimension for the channels. Thus, we effectively analyze 4D data (x,y,z,c) during the convolution operation.

The Network is a stack of multiple layers, each convolving their input with a set of filters. The filters are optimized on the training data using stochastic gradient descent; their initial values are drawn from a Gaussian distribution with zero mean. Following the convolution operation, we apply a nonlinear voxel-wise squashing function, the hyperbolic tangent function. The convolution operation reduces the 4D block of the preceding layer to filtered 3D blocks. All filtered 3D blocks are then combined to serve as 4D input for the next layer. The final convolution layer has as many filters as there are different classes to be predicted, in our case six (edema, enhancing tumor, non-enhancing tumor, necrosis, air, other/normal tissue). A final soft-max operation ensures that the values of the output layer sum to one, and thus can be interpreted as probabilities.

We achieve a speedup of several orders of magnitude by interpreting fully-connected layers (the final layer of the network) as convolutional layers with filters of size 1^3. Using this trick we can predict multiple neighboring voxels in one pass and benefit from a highly reduced computational overhead, as compared to making predictions for voxels independently. This idea has been described by Giusti et al. [2]. During training we predict 9^3 voxels per gradient optimization pass. This effectively enables one to train Convolutional Networks for segmentation on a single CPU-thread in reasonable time (in our case $\approx 30 - 40\,\mathrm{h}$). We also evaluated a GPU implementation that offers a further speedup, allowing to train the network in less than half a day. Generating predictions for an entire volume takes about one minute. The network is implemented using the Theano library [1].

We train different classifiers for LG and HG tumors, as they might have a different local structure, but can be distinguished globally. The data-flow of an exemplary 3D-CNN is shown in Fig. 1.

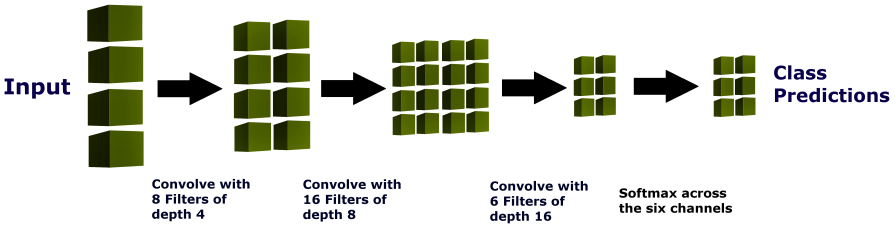

Input
Convolve with 8 Filters of depth 4
Convolve with 16 Filters of depth 8
Convolve with 6 Filters of depth 16
Softmax across the six channels
Class Predictions

Fig. 1. Visualisation of the memory and operations of an exemplary Convolutional Network. The input is a 3D image with four channels/modalities. Each filter has three spatial dimensions, with a typical size of 5^3 voxels in our experiments, as well as an additional fourth dimension in order to take all input channels into account (e.g. the first 8 filters are of size 5x5x5x4). The depth of a hidden layer is equal to the number of filters of the preceding layer. After the convolution the nonlinear activation function tanh is applied independently to all voxels in all channels (not shown). The employed convolution only emits a value at points where the filter fully overlaps with the data, thus the number of voxels per channel decreases after each convolution when filters larger than 1^3 are used.

We trained one network with four layers, the first three layers all have filters with a size of 5^3 (plus one dimension that corresponds to the depth of the channel of the input to the layer, e.g. the first layer's filters have a shape of (5,5,5,4) to account for the four input channels). We used 15 filters in the first layer, 25 for the next two and six filters in the last layer (one for each of the six different classes), respectively. We trained a second network that is identical to the first one, except that an extra layer containing 40 filters of size 5^3 was added in front of the last layer.

2.3 Post-processing

We identified connected components (CC) of the thresholded class predictions and discarded all those that contain less than 3000 voxels. This procedure removes disconnected and likely false-positive segmentations. CC removal was realized with the VIGRA[4] library.

2.4 Evaluation of the Results

We evaluate the challenge data, for which no ground truth is publicly available, through the challenge website[5].

3 Results

We present results (Tab. 1) for the challenge data for an average of the voxel-wise predicted probabilities of two Convolutional Networks, that slightly vary in their architecture (cf. 2.2).

Table 1. Dice scores for BratTS 2013 challenge data

Method	whole	core	active
Human Rater [7]	88	93	74
Best 2013	87	78	74
Current Best	88	83	72
Averaged Network	87	77	73

4 Discussion

We demonstrated the successful application of deep learning for segmenting tumorous regions of MR scans, yielding results among the top-scoring submissions for the BratTS 2013 challenge (ranked second at the time of writing). A clear advantage of this approach is that it does not rely on hand-crafted features.

The other approach that we submitted for this years challenge employs a random forest for predicting the segmentations [5]. For both methods we noticed the occurrence of 'holes' (healthy neuronal tissue) within tumorous tissue. Again, we chose not to fill those in, as they are biologically plausible. Besides using pre-processing that normalizes each channel with its mean CSF-value we also experimented using the provided raw data directly as an input for the CNN. This resulted in a weaker performance, emphasizing the importance of a suitable intensity calibration.

[4] https://github.com/ukoethe/vigra

[5] http://www.virtualskeleton.ch

Further improvements might be expected from using larger Networks or using dropout [3], a method that helps to prevent over-fitting. However, the errors of the predictions are already close to the range of inter-rater variability and it is therefore not likely to yield large improvements when training the network with a single ground truth labeling only.

As future work we plan to introduce two additional neurons in the output layer, coding for low and high grade, respectively. This will allow to generate pseudo-probability maps that indicate areas of different malignancies and thus might help to characterize tumor sub-regions.

Acknowledgments. This work was supported by a postdoctoral fellowship from the Medical Faculty of the University of Heidelberg.

References

1. Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y.: Theano: a CPU and GPU math expression compiler. In: Proceedings of the Python for Scientific Computing Conference (SciPy) (Jun 2010), oral Presentation
2. Giusti, A., Cireşan, D.C., Masci, J., Gambardella, L.M., Schmidhuber, J.: Fast image scanning with deep max-pooling convolutional neural networks. arXiv preprint arXiv:1302.1700 (2013)
3. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR abs/1207.0580 (2012)
4. Kistler, M., Bonaretti, S., Pfahrer, M., Niklaus, R., Büchler, P.: The virtual skeleton database: an open access repository for biomedical research and collaboration. J Med Internet Res 15(11), e245 (2013)
5. Kleesiek, J., Biller, A.B., Urban, G., Koethe, U., Bendszus, M., Hamprecht, F.: ilastik for Multi-modal Brain Tumor Segmentation (2014), submitted to BraTS 2014 Workshop
6. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. The handbook of brain theory and neural networks 3361 (1995)
7. Menze, B., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS), submitted to IEEE Transactions on Medical Imaging
8. Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., Curschmann, J., Janzer, R.C., Ludwin, S.K., Gorlia, T., Allgeier, A., Lacombe, D., Cairncross, J.G., Eisenhauer, E., Mirimanoff, R.O.: Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine 352(10), 987–996 (2005), http://www.nejm.org/doi/full/10.1056/NEJMoa043330, pMID: 15758009
9. Wen, P.Y., Macdonald, D.R., Reardon, D.A., Cloughesy, T.F., Sorensen, A.G., Galanis, E., Degroot, J., Wick, W., Gilbert, M.R., Lassman, A.B., Tsien, C., Mikkelsen, T., Wong, E.T., Chamberlain, M.C., Stupp, R., Lamborn, K.R., Vogelbaum, M.A., van den Bent, M.J., Chang, S.M.: Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. J Clin Oncol 28(11), 1963–72 (Apr 2010)

Segmentation of Brain Tumor Tissues with Convolutional Neural Networks

Darko Zikic[1], Yani Ioannou[1,2], Matthew Brown[2], and Antonio Criminisi[1]

[1] Microsoft Research, Cambridge, UK
[2] University of Bath, Bath, UK

Abstract In this work, we investigate the possibility to directly apply convolutional neural networks (CNN) to segmentation of brain tumor tissues. As input to the network, we use multi-channel intensity information from a small patch around each point to be labeled. Only standard intensity pre-processing is applied to the input data to account for scanner differences. No post-processing is applied to the output of the CNN. We report promising preliminary results on the high-grade training data from the BraTS 2013 challenge. Work for the final submission will include architecture modifications, parameter tuning and training on the BraTS 2014 training corpus.

1 Introduction

In this work, we apply convolutional neural networks (CNNs) to the problem of brain tumor segmentation. The work is motivated by the recent success of CNNs for object recognitionion 2D images [1], and the availability of efficient off-the-shelf implementations such as Caffe [2].

CNNs are currently primarily used for object recognition, i.e. if an image contains an object, the complete image is assigned the corresponding label. Two exceptions are [3,4], where CNNs are used inside more complex frameworks in order to perform the segmentation. In the domain of medical image analysis, CNNs have been very successfully applied for mitosis detection in 2D histology images [5]. The intermediate step of [5] can be seen as a binary segmentation of mitotic cells, and the use of CNNs in that work as a per-pixel classifier is similar to the one we use here.

In this work, we explore the possibility of applying CNNs to segmentation of brain tumors *directly.* The CNNs operate on standardly pre-processed intensity information, and we apply no further post-processing to their output.

2 Method

For the segmentation task, we use a standard CNN implementation based on multi-channel 2D convolutions, and adapt it such that it operates on multi-channel 3D data usually available for the brain tumor segmentation task.

We apply the CNN in a sliding-window fashion in the 3D space, for each point inside the brain masks. At each point x, the CNN takes as input a multi-channel 3D patch around this point $P(x)$. Given $P(x)$, the CNN is trained to make a class prediction for the central patch point x.

2.1 Input Data Representation

For each case in the BraTS database, the multi-channel 3D data consists of 4 different 3D MR contrast images: contrast enhanced T1 (T1c), T1, T2 and FLAIR. While T1c usually has an isotropic resolution, the other channels originally have a slice distance which is larger than the in-slice element spacing. In the BraTS challenge, all data is resampled to fit the T1c resolution. For each point x to be labeled, we extract a multi-channel patch $P(x)$ around it, which has spatial dimensions d_1, d_2, d_3. Here, $d1$ and $d2$ are taken to be in-slice dimensions corresponding to high resolution, and d_3 is the lower-resolution axial direction.

Having 4 channels in our task, each 4-channel 3D patch $P(x)$ of size $(d_1 \times d_2 \times d_3 \times 4)$ can also be interpreted as a $(4 \cdot d_3)$-channel 2D patch of size $(d_1 \times d_2 \times 4d_3)$, where the 2D space d_1-d_2 corresponds to original MR-slices, in which the original data generally has the highest resolution. We use this interpretation to apply a standard 2D-CNN convolutional architecture to our 3D problem. Thusly, in the first convolutional layer, we use convolutional filters of size $5 \times 5 \times 4d_3$, and perform a 2D convolution with this filter along the dimensions d_1 and d_2 within each patch $P(x)$ of size $19 \times 19 \times 4d_3$.

This approach is taken for two reasons. First, we can use existing efficient off-the-shelf CNN implementations for 2D convolutions without large modifications. Second, performing 2D instead of 3D convolution is computationally more efficient. The justification for this step is that due to lower resolution in d_3 dimensions, we expect that omitting the convolution in this direction will have a minor impact on accuracy.

2.1.1 Pre-processing As additional pre-processing for the BraTS data, we perform inhomogeneity correction in each channel by [6], set the median of each channel to a fixed value of 0, and downsample the images by a factor of two with nearest-neighbor interpolation. Testing is also performed on down-sampled images, and the results are correspondingly upsampled before quantitative evaluation.

2.2 CNN Architecture and Optimization

We use a standard CNN framework following [1], with the following per layer characteristics of the architecture:

- layer 0: input patch of size $19 \times 19 \times 4$,
  (i.e. we currently only use a single slice from each of the 4 channels)
- layer 1: 64 filters of size $5 \times 5 \times 4$,
  (resulting in $15 \times 15 \times 64$ nodes)

- layer 2: max-pooling with kernel size 3 and stride of 3, (resulting in $5 \times 5 \times 64$ nodes)
- layer 3: 64 filters of size $3{\times}3{\times}64$, (resulting in $3 \times 3 \times 64$ nodes)
- layer 4: fully connected with 512 nodes
- layer 5: soft-max (fully-connected) with 5 output nodes (for the 5 classes)

All inner nodes in the network use a rectified linear unit (ReLU) as a non-linearity term.

We use log-loss as the energy function for training, and optimization is performed with a stochastic gradient descent with momentum.

3 Preliminary Evaluation

Since we did not have access to the BraTS evaluation platform at the time of this submission, we perform the preliminary evaluation on the training data set from the BraTS 2013 challenge. We focus on the 20 high-grade cases from training set. To provide some context, we relate to results of our previous method from [7], which is based on randomized forests (RF).

We perform the evaluation of the CNN approach with a 2-fold validation where, based on the ascending ordering of the test cases IDs, the first fold contains the odd cases, and the second fold contains the even ones. Thus each fold contains 10 cases. Results for each fold are computed by a CNN which was trained on the other fold. For training, we use all samples available for the tumor classes, and we randomly subsample the number of background/brain samples to correspond to the total of the tumor samples for each case.

The results for the RF approach are computed in a leave-1-out manner, where for each case, the RF method was trained on the remaining 19 high-grade cases. For RF training, the background is randomly subsampled by a factor of 0.1 which is very similar to the ones used for the CNN training. Thus, the RF approach has access to almost double the amount of training data compared to the CNN approach, which seems like an advantage.

The results are summarized in Table 1 and Figure 1, and show a promising performance of the CNN-based approach.

4 Discussion and Future Work

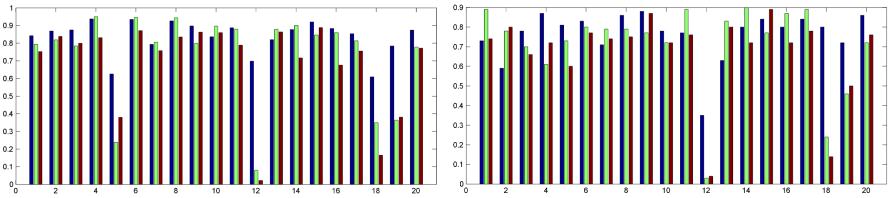
The preliminary results indicate that the unoptimized CNN architecture is already capable of achieving acceptable results. Our work for the final submission will include training on the large BraTS 2014 training corpus, improvements of the network architecture, and parameter tuning.

References

1. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems 25. (2012)

Method	Training HG (BraTS 2013)		
	complete	core	enhancing
RF	76.3±12.4	70.9±22.5	67.4±21.7
CNN	83.7±9.4	73.6±25.6	69.0±24.9

Table 1: Quantitative summary of results on the high-grade training data from the BraTS 2013 challenge (10 cases). The results for CNN are obtained by a 2-fold data split for training and testing. The results for RF are obtained with a leave-one-out experiment. This means that for CNN each prediction is based on a classifier trained on 10 cases, while for RF, each classifier is trained on 19 cases, i.e. nearly the double the amount of data.



(a) RF (leave-1-out)

(b) CNN (2-fold)

Figure 1: Visualization of the results on the training data from BraTS 2013, and relation to results of a randomized forest from [7]. Results are shown for the complete tumor (blue), core tumor (green) and enhancing tumor (red).

2. Jia, Y.: Caffe: An open source convolutional architecture for fast feature embedding. `http://caffe.berkeleyvision.org/` (2013)
3. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence, IEEE Transactions on **35**(8) (2013) 1915–1929
4. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014)
5. Ciresan, D.C., Giusti, A., Gambardella, L.M., Schmidhuber, J.: Mitosis detection in breast cancer histology images with deep neural networks. In: MICCAI. Volume 2. (2013) 411–418
6. Tustison, N., Avants, B., Cook, P., Zheng, Y., Egan, A., Yushkevich, P., Gee, J.: N4ITK: Improved N3 Bias Correction. Medical Imaging, IEEE Transactions on (2010)
7. Zikic, D., Glocker, B., Konukoglu, E., Shotton, J., Criminisi, A., Ye, D., Demiralp, C., Thomas, O.M., Das, T., Jena, R., Price, S.J.: Context-sensitive classification forests for segmentation of brain tumor tissues. In: MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation (BraTS). (2012)