This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

Brain Tumor Segmentation using Convolutional Neural Networks in MRI Images

Sérgio Pereira, Adriano Pinto, Victor Alves and Carlos A. Silva

***Abstract*—Among brain tumors, gliomas are the most common and aggressive, leading to a very short life expectancy in their highest grade. Thus, treatment planning is a key stage to improve the quality of life of oncological patients. Magnetic Resonance Imaging (MRI) is a widely used imaging technique to assess these tumors, but the large amount of data produced by MRI prevents manual segmentation in a reasonable time, limiting the use of precise quantitative measurements in the clinical practice. So, automatic and reliable segmentation methods are required; however, the large spatial and structural variability among brain tumors make automatic segmentation a challenging problem. In this paper, we propose an automatic segmentation method based on Convolutional Neural Networks (CNN), exploring small 3×3 kernels. The use of small kernels allows designing a deeper architecture, besides having a positive effect against overfitting, given the fewer number of weights in the network. We also investigated the use of intensity normalization as a pre-processing step, which though not common in CNN-based segmentation methods, proved together with data augmentation to be very effective for brain tumor segmentation in MRI images. Our proposal was validated in the Brain Tumor Segmentation Challenge 2013 database (BRATS 2013), obtaining simultaneously the first position for the complete, core, and enhancing regions in Dice Similarity Coefficient metric** (0.88, 0.83, 0.77) **for the Challenge data set. Also, it obtained the overall first position by the online evaluation platform. We also participated in the on-site BRATS 2015 Challenge using the same model, obtaining the second place, with Dice Similarity Coefficient metric of** 0.78, 0.65, **and** 0.75 **for the complete, core, and enhancing regions, respectively.**

***Index Terms*—Magnetic Resonance Imaging, Glioma, Brain Tumor, Brain Tumor Segmentation, Deep Learning, Convolutional Neural Networks**

I. Introduction

Gliomas are the brain tumors with the highest mortality rate and prevalence [1]. These neoplasms can be graded into Low Grade Gliomas (LGG) and High Grade Gliomas (HGG), with the former being less aggressive and infiltrative than the latter [1], [2]. Even under treatment, patients do not survive on average more than 14 months after diagnosis [3]. Current treatments include surgery, chemotherapy, radiotherapy, or a combination of them [4]. MRI is especially useful to assess gliomas in clinical practice, since it is possible to acquire MRI sequences providing complementary information [1].

S. Pereira, A. Pinto and C. A. Silva are with CMEMS-UMinho Research Unit, University of Minho, Campus Azurém, Guimarães, Portugal (e-mail: id5692@alunos.uminho.pt (S. Pereira), csilva@dei.uminho.pt (C. Silva))

S. Pereira and V. Alves are with Centro Algoritmi, Universidade do Minho, Braga, Portugal (e-mail: valves@di.uminho.pt (V. Alves))

The accurate segmentation of gliomas and its intra-tumoral structures is important not only for treatment planning, but also for follow-up evaluations. However, manual segmentation is time-consuming and subjected to inter- and intra-rater errors difficult to characterize. Thus, physicians usually use rough measures for evaluation [1]. For these reasons, accurate semi-automatic or automatic methods are required [1], [5]. However, it is a challenging task, since the shape, structure, and location of these abnormalities are highly variable. Additionally, the tumor mass effect change the arrangement of the surrounding normal tissues [5]. Also, MRI images may present some problems, such as intensity inhomogeneity [6], or different intensity ranges among the same sequences and acquisition scanners [7].

In brain tumor segmentation, we find several methods that explicitly develop a parametric or non-parametric probabilistic model for the underlying data. These models usually include a likelihood function corresponding to the observations and a prior model. Being abnormalities, tumors can be segmented as outliers of normal tissue, subjected to shape and connectivity constrains [8]. Other approaches rely on probabilistic atlases [9]–[11]. In the case of brain tumors, the atlas must be estimated at segmentation time, because of the variable shape and location of the neoplasms [9]–[11]. Tumor growth models can be used as estimates of its mass effect, being useful to improve the atlases [10], [11]. The neighborhood of the voxels provides useful information for achieving smoother segmentations through Markov Random Fields (MRF) [9]. Zhao at al. [5] also used a MRF to segment brain tumors after a first oversegmentation of the image into supervoxels, with a histogram-based estimation of the likelihood function. As observed by Menze et al. [5], generative models generalize well in unseen data, but it may be difficult to explicitly translate prior knowledge into an appropriate probabilistic model.

Another class of methods learns a distribution directly from the data. Although a training stage can be a disadvantage, these methods can learn brain tumor patterns that do not follow a specific model. This kind of approaches commonly consider voxels as independent and identically distributed [12], although context information may be introduced through the features. Because of this, some isolated voxels or small clusters may be mistakenly classified with the wrong class, sometimes in physiological and anatomically unlikely locations. To overcome this problem, some authors include information of the neighborhood by embedding the probabilistic predictions of the classifier into a Conditional Random Field [12]–[15]. Classifiers such as Support Vector Machines [12], [13] and,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

more recently, Random Forests (RF) [14]–[21] were successfully applied in brain tumor segmentation. The RF became very used due to its natural capability in handling multi-class problems and large feature vectors. A variety of features were proposed in the literature: encoding context [15], [16], [21], first-order and fractals-based texture [14], [15], [18], [21], [22], gradients [14], [15], brain symmetry [14], [15], [19], and physical properties [19]. Using supervised classifiers, some authors developed other ways of applying them. Tustison et al. [19] developed a two-stage segmentation framework based on RFs, using the output of the first classifier to improve a second stage of segmentation. Geremia et al. [20] proposed a Spatially Adaptive RF for hierarchical segmentation, going from coarser to finer scales. Meier et al. [23] used a semi-supervised RF to train a subject-specific classifier for post-operative brain tumor segmentation.

Other methods known as Deep Learning deal with representation learning by automatically learning an hierarchy of increasingly complex features directly from data [24]. So, the focus is on designing architectures instead of developing handcrafted features, which may require specialized knowledge [25]. CNNs have been used to win several object recognition [26], [27] and biological image segmentation [28] challenges. Since a CNN operates over patches using kernels, it has the advantages of taking context into account and being used with raw data. In the field of brain tumor segmentation, recent proposals also investigate the use of CNNs [29]–[35]. Zikic et al. [29] used a shallow CNN with two convolutional layers separated by max-pooling with stride 3, followed by one fully-connected (FC) layer and a softmax layer. Urban et al. [30] evaluated the use of 3D filters, although the majority of authors opted for 2D filters [31]–[35]. 3D filters can take advantage of the 3D nature of the images, but it increases the computational load. Some proposals evaluated two-pathway networks to allow one of the branches to receive bigger patches than the other, thus having a larger context view over the image [31], [32]. In addition to their two-pathway network, Havaei et al. [32] built a cascade of two networks and performed a two-stage training, by training with balanced classes and then refining it with proportions near the originals. Lyksborg et al. [33] use a binary CNN to identify the complete tumor. Then, a cellular automata smooths the segmentation, before a multi-class CNN discriminates the sub-regions of tumor. Rao et al. [34] extracted patches in each plane of each voxel and trained a CNN in each MRI sequence; the outputs of the last FC layer with softmax of each CNN are concatenated and used to train a RF classifier. Dvořák and Menze [35] divided the brain tumor regions segmentation tasks into binary sub-tasks and proposed structured predictions using a CNN as learning method. Patches of labels are clustered into a dictionary of label patches, and the CNN must predict the membership of the input to each of the clusters.

In this paper, inspired by the groundbreaking work of Simonyan and Zisserman [36] on deep CNNs, we investigate the potential of using deep architectures with small convolutional kernels for segmentation of gliomas in MRI images. Simonyan and Zisserman proposed the use of small 3×3 kernels to obtain deeper CNNs. With smaller kernels we can stack more convolutional layers, while having the same receptive field of bigger kernels. For instance, two 3×3 cascaded convolutional layers have the same effective receptive field of one 5×5 layer, but fewer weights [36]. At the same time, it has the advantages of applying more non-linearities and being less prone to overfitting because small kernels have fewer weights than bigger kernels [36]. We also investigate the use of the intensity normalization method proposed by Nyúl et al. [7] as a pre-processing step that aims to address data heterogeneity caused by multi-site multi-scanner acquisitions of MRI images. The large spatial and structural variability in brain tumors are also an important concern that we study using two kinds of data augmentation.

The remainder of this paper is organized as follows. In Section II, the proposed method is presented. The databases used for evaluation and the experimental setup are detailed in Section III. Results are presented and discussed in Section IV. Finally, the main conclusions are presented in Section V.

II. Method

Fig. 1 presents an overview of the proposed approach. There are three main stages: pre-processing, classification via CNN and post-processing.

A. Pre-processing

MRI images are altered by the bias field distortion. This makes the intensity of the same tissues to vary across the image. To correct it, we applied the N4ITK method [6]. However, this is not enough to ensure that the intensity distribution of a tissue type is in a similar intensity scale across different subjects for the same MRI sequence, which is an explicit or implicit assumption in most segmentation methods [37]. In fact, it can vary even if the image of the same patient is acquired in the same scanner in different time points, or in the presence of a pathology [7], [38]. So, to make the contrast and intensity ranges more similar across patients and acquisitions, we apply the intensity normalization method proposed by Nyúl et al. [7] on each sequence. In this intensity normalization method, a set of intensity landmarks $I_L = \{pc_1, i_{p_{10}}, i_{p_{20}}, \cdots, i_{p_{90}}, pc_2\}$ are learned for each sequence from the training set. pc_1 and pc_2 are chosen for each MRI sequence as described in [38]. i_{p_l} represents the intensity at the l^{th} percentile. After training, the intensity normalization is accomplished by linearly transforming the original intensities between two landmarks into the corresponding learned landmarks. In this way, the histogram of each sequence is more similar across subjects.

After normalizing the MRI images, we compute the mean intensity value and standard deviation across all training patches extracted for each sequence. Then, we normalize the patches on each sequence to have zero mean and unit variance[1].

[1]The mean and standard deviation computed in the training patches are used to normalize the testing patches.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

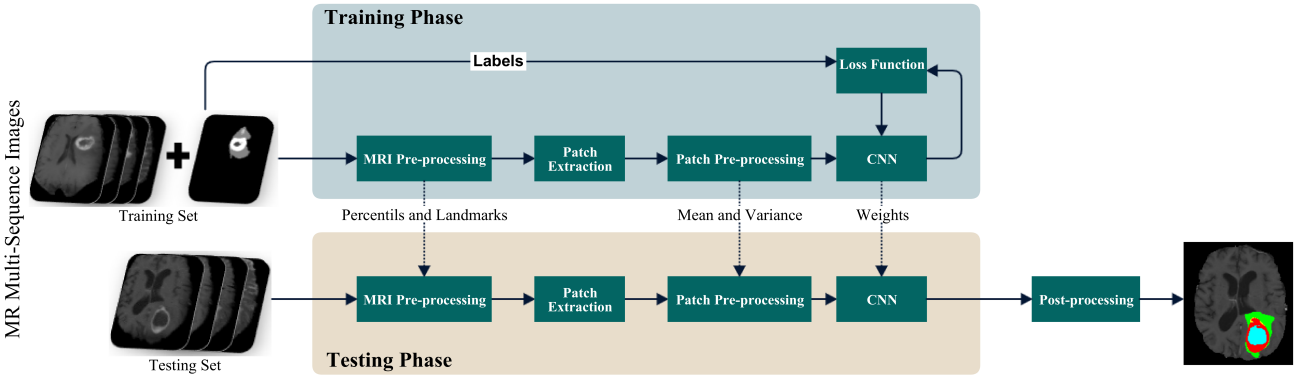


Fig. 1: Overview of the proposed method.

B. Convolutional Neural Network

CNN were used to achieve some breakthrough results and win well-known contests [26], [27]. The application of convolutional layers [39], [40] consists in convolving a signal or an image with kernels to obtain feature maps. So, a unit in a feature map is connected to the previous layer through the weights of the kernels. The weights of the kernels are adapted during the training phase by backpropagation, in order to enhance certain characteristics of the input. Since the kernels are shared among all units of the same feature maps, convolutional layers have fewer weights to train than dense FC layers, making CNN easier to train and less prone to overfitting. Moreover, since the same kernel is convolved over all the image, the same feature is detected independently of the location – translation invariance. By using kernels, information of the neighborhood is taken into account, which is an useful source of context information [25], [39], [40]. Usually, a non-linear activation function is applied on the output of each neural unit.

If we stack several convolutional layers, the extracted features become more abstract with the increasing depth. The first layers enhance features such as edges, which are aggregated in the following layers as motifs, parts, or objects [25].

The following concepts are important in the context of CNN:

a) Initialization: it is important to achieve convergence. We use the Xavier initialization [41]. With this, the activations and the gradients are maintained in controlled levels, otherwise back-propagated gradients could vanish or explode.

b) Activation Function: it is responsible for non-linearly transforming the data. Rectifier linear units (ReLU), defined as

$$f(x) = max(0,x), \tag{1}$$

were found to achieve better results than the more classical sigmoid, or hyperbolic tangent functions, and speed up training [26], [42]. However, imposing a constant 0 can impair the gradient flowing and consequent adjustment of the weights [43]. We cope with these limitations using a variant called leaky rectifier linear unit (LReLU) [43] that introduces a small slope on the negative part of the function. This function is defined as

$$f(x) = \max(0,x) + \alpha \min(0,x) \tag{2}$$

where α is the leakyness parameter. In the last FC layer, we use softmax.

c) Pooling: it combines spatially nearby features in the feature maps. This combination of possibly redundant features makes the representation more compact and invariant to small image changes, such as insignificant details; it also decreases the computational load of the next stages. To join features it is more common to use max-pooling or average-pooling [25].

d) Regularization: it is used to reduce overfitting. We use Dropout [44], [45] in the FC layers. In each training step, it removes nodes from the network with probability p. In this way, it forces all nodes of the FC layers to learn better representations of the data, preventing nodes from co-adapting to each other. At test time, all nodes are used. Dropout can be seen as an ensemble of different networks and a form of bagging, since each network is trained with a portion of the training data [44], [45].

e) Data Augmentation: it can be used to increase the size of training sets and reduce overfitting [26]. Since the class of the patch is obtained by the central voxel, we restricted the data augmentation to rotating operations. Some authors also consider image translations [26], but for segmentation this could result in attributing a wrong class to the patch. So, we increased our data set during training by generating new patches through the rotation of the original patch. In our proposal, we used angles multiple of 90°, although another alternative will be evaluated.

f) Loss Function: it is the function to be minimized during training. We used the Categorical Cross-entropy,

$$H = -\sum_{j \in voxels} \sum_{k \in classes} c_{j,k} \log(\hat{c}_{j,k}) \tag{3}$$

where $\hat{c}$ represents the probabilistic predictions (after the softmax) and c is the target.

In the next subsections, we discuss the architecture and training of our CNN.

1) Architecture: We aim at a reliable segmentation method; however, brain tumors present large variability in intra-tumoral structures, which makes the segmentation a challenging problem. To reduce such complexity, we designed a CNN and tuned

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

the intensity normalization transformation for each tumor grade – LGG and HGG.

The proposed architectures are presented in Tables I and II[2]. The architecture used for HGG is deeper than the one for LGG, because going deeper did not improve results in the latter. To go deeper, one must include more layers with weights, which may increase overfitting, given the smaller training set of LGG. This is supported by the need of setting Dropout with $p = 0.5$ in LGG, while it is $p = 0.1$ in HGG, since the database used for evaluation contained more HGG then LGG cases. Additionally, the appearance and patterns are different in HGG and LGG. Since we are doing segmentation, we need a precise sense of location. Pooling can be positive to achieve invariance and to eliminate irrelevant details, however, it can also have a negative effect by eliminating important details. We apply overlapping pooling with 3×3 receptive fields and 2×2 stride to keep more information of location. In the convolutional layers the feature maps are padded before convolution, so that the resulting feature maps could maintain the same dimensions. In the case of HGG there are 2,118,213 weights to train, while in LGG it lowers to 1,933,701 weights because it has two less convolutional layers. All sequences were used as input. LReLU is the activation function in all layers with weights, with the exception of the last that uses softmax. Dropout was used only in the FC layers.

TABLE I: Architecture of the HGG CNN. In inputs, the first dimension refers to the number of channels and the next two to the size of the patch, or feature maps. Conv. refers to convolutional layers and Max-pool. to max-pooling.

	Type	Filter size	HGG Stride	# filters	FC units	Input
Layer 1	Conv.	3×3	1×1	64	-	4×33×33
Layer 2	Conv.	3×3	1×1	64	-	64×33×33
Layer 3	Conv.	3×3	1×1	64	-	64×33×33
Layer 4	Max-pool.	3×3	2×2	-	-	64×33×33
Layer 5	Conv.	3×3	1×1	128	-	64×16×16
Layer 6	Conv.	3×3	1×1	128	-	128×16×16
Layer 7	Conv.	3×3	1×1	128	-	128×16×16
Layer 8	Max-pool.	3×3	2×2	-	-	128×16×16
Layer 9	FC	-	-	-	256	6272
Layer 10	FC	-	-	-	256	256
Layer 11	FC	-	-	-	5	256

TABLE II: Architecture of the LGG CNN. In inputs, the first dimension refers to the number of channels and the next two to the size of the patch, or feature maps. Conv. refers to convolutional layers and Max-pool. to max-pooling.

	Type	Filter size	LGG Stride	# filters	FC units	Input
Layer 1	Conv.	3×3	1×1	64	-	4×33×33
Layer 2	Conv.	3×3	1×1	64	-	64×33×33
Layer 3	Max-pool.	3×3	2×2	-	-	64×33×33
Layer 4	Conv.	3×3	1×1	128	-	64×16×16
Layer 5	Conv.	3×3	1×1	128	-	128×16×16
Layer 6	Max-pool.	3×3	2×2	-	-	128×16×16
Layer 7	FC	-	-	-	256	6272
Layer 8	FC	-	-	-	256	256
Layer 9	FC	-	-	-	5	256

[2]We also provide graphical representations in the online *Supplementary Materials*.

2) Training: To train the CNN the loss function must be minimized, but it is highly non-linear. We use Stochastic Gradient Descent as an optimization algorithm, which takes steps proportionally to the negative of the gradient in the direction of local minima. Nevertheless, in regions of low curvature it can be slow. So, we also use Nesterov's Accelerated Momentum to accelerate the algorithm in those regions. The momentum ν is kept constant, while the learning rate ε was linearly decreased, after each epoch. We consider an epoch as a complete pass over all the training samples.

C. Post-processing

Some small clusters may be erroneously classified as tumor. To deal with that, we impose volumetric constrains by removing clusters in the segmentation obtained by the CNN that are smaller than a predefined threshold τ_{VOL}.

III. Experimental Setup

A. Database

The proposed method was validated on the BRATS 2013 and 2015 databases[3] [5], [46]. For every patient in BRATS there are four MRI sequences available: T1-weighted (T1), T1 with gadolinium enhancing contrast (T1c), T2-weighted (T2) and FLAIR. The images of each subject were already aligned with the T1c and skull stripped. BRATS 2013 contains three data sets: Training, Leaderboard and Challenge, comprising 65 MR scans from different patients — histological diagnosis: astrocytomas or oligoastrocytomas, LGG, and anaplastic astrocytomas and glioblastoma multiforme tumors, HGG. The Training set contains 20 HGG and 10 LGG, with manual segmentations available. The Leaderboard set is composed by 21 HGG and 4 LGG, while the Challenge set includes 10 HGG. Metrics for these two sets are computed through the online evaluation platform [47], given that the manual segmentations are not publicly available. In BRATS 2015, the Training set comprises 220 and 54 acquisitions of HGG and LGG, respectively. The Challenge set contains 53 cases, including both grades. In this case, the evaluation metrics were computed by the organizers of the challenge. The manual segmentation identifies four types of intra-tumoral classes: necrosis, edema, non-enhancing, and enhancing tumor. However, the evaluation is performed for the enhancing tumor, the core (necrosis + non-enhancing tumor + enhancing tumor), and the complete tumor (all classes combined).

B. Setup

Some of the hyperparameters of the architectures were shown in Tables I and II. The remaining are depicted in Table III. All hyperparameters were found using the validation set, consisting of one subject in both HGG and LGG.

We approached brain tumor segmentation as a multi-class classification problem with 5 classes (normal tissue, necrosis, edema, non-enhancing, and enhancing tumor). However, in brain tumor, the classes are imbalanced. So, we used all samples from the underrepresented classes and randomly sampled

[3]The data set of BRATS 2014 is not currently available.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

TABLE III: Hyperparameters of the proposed method.

Stage	Hyperparameter	Value
Initialization	bias	0.1
	weights	Xavier
Leaky ReLU	α	0.333
Dropout	p – HGG	0.1
	p – LGG	0.5
Training	epochs – HGG	20
	epochs – LGG	25
	ν	0.9
	Initial ε	0.003
	Final ε	0.00003
	Batch	128
Post-processing	τ_{VOL} – HGG	10000
	τ_{VOL} – LGG	3000

from the other. Additionally, the number of samples of necrosis and enhancing tumor is small in the LGG training set. To cope with that, we also normalized the intensities of HGG using the landmarks calculated with LGG to extract samples of those classes from HGG to use as training samples in LGG. To train the CNNs for HGG and LGG, we extracted around 450,000 and 335,000 patches, respectively. Note that, with data augmentation, we end up having roughly four times these numbers as effective training samples. Approximately 40% of these patches represent normal tissue in HGG and 50% in LGG. The learning rate was linearly decreased after each epoch during the training stage.

The CNNs were developed using Theano [48], [49] and Lasagne [50]. The trained architectures are available online[4].

C. Evaluation

The evaluation of the segmentations considered three metrics: Dice Similarity Coefficient (DSC), Positive Predictive Value (PPV) and Sensitivity. The DSC [51] measures the overlap between the manual and the automatic segmentation. It is defined as,

$$DSC = \frac{2TP}{FP+2TP+FN}, \tag{4}$$

where TP, FP and FN are the numbers of true positive, false positive and false negative detections, respectively. PPV is a measure of the amount of FP and TP, defined as,

$$PPV = \frac{TP}{TP+FP}. \tag{5}$$

Finally, Sensitivity is useful to evaluate the number of TP and FN detections, being defined as

$$Sensitivity = \frac{TP}{TP+FN}. \tag{6}$$

[4]http://www.dei.uminho.pt/pessoas/csilva/brats_cnn/

IV. Experimental Results and Discussion

In this section, we analyze the effect of key components and the choice of the plane over which we extract patches on the performance of the proposed method. Also, we compare our method with the state of the art using the same database, including also methods based on deep learning for brain tumor segmentation. Lastly, we report our result during the participation on BRATS Challenge 2015.

A. Validation of Key Components

We evaluate the effect of each component on the proposed approach by studying the improvement in performance. This increment in performance is evaluated as the mean gain in the metrics (DSC, PPV and Sensitivity), which is obtained in the following way: we compute all metrics using the proposed method for the data sets; then, we remove or substitute the component under study, and compute the metrics for this alternative method. Finally, we subtract each metric for the two systems and calculate the average across the subtractions, obtaining the mean gain, μ_{gain}. The metric of each experiment is reported in Table IV, Fig. 2 and 3 present the boxplots in the Leaderboard and Challenge data set, respectively, and in Fig. 4 we exemplify the effect of the experiments in the segmentation of tumor in two patients (HGG and LGG). In the experiments, we maintained the hyperparameters presented in Table III as possible to preserve the same conditions[5]. Also, only the images in the Training data set are used in the learning phase of the intensity normalization method. All tests in this section use patches extracted from planes perpendicular to the Axial axis of the MRI image, except in subsection IV-B, where it is evaluated the choice of the best axis.

a) Pre-processing: The effect of the pre-processing on the segmentation was evaluated by comparing with an alternative method described in [19]. We chose this method, because it is also utilized in a CNN-based brain tumor segmentation method [32]. This alternative pre-processing starts by applying a 1% winsorizing over the intensities within the brain. Then, the N4ITK is used to correct the bias field in each MRI sequence and the intensities are linearly transformed to $[0,1]$. Finally, we normalized each sequence to have zero mean and unit variance. During the training stage of the CNN with this pre-processing for LGG, we found to be necessary to decrease the initial and final learning rate to 3×10^{-5} and 3×10^{-7}, respectively, otherwise the optimization would diverge. Observing Table IV, we verify that the pre-processing using the intensity normalization method by Nyúl et al. improved most of the metrics, obtaining a mean gain of 4.6% (Leaderboard: 4.2%, Challenge: 4.9%). This improvement was specially larger for LGG, indicating that the proposed pre-processing increased the detection of the complete as well as the core of the tumor, which is considered a difficult task [5]. Also, comparing the drop in performance, when removing our pre-processing, and the one verified when removing any other

[5]The learning rate was kept constant after 25 epochs; although the validation error may fluctuate, we verified that it stabilized before 30 epochs, so, we trained that amount of epochs and selected the one with the best validation metrics.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

TABLE IV: Study of key components of the proposed method. In each test, just the referred component was modified in the Proposed method. Results in bold represent metrics with p-value < 0.05 computed with the two-sided paired Wilcoxon Signed-Rank Test when comparing the results with each component of the Proposed method in each grade, or combination of grades; underlined results represent the one with the highest metric for each region in each grade, or combination of grades.

			DSC			PPV			Sensitivity		
Dataset	Method	Grade	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
Leaderboard	**Proposed**	HGG	0.88	0.76	0.73	0.91	0.90	0.72	0.86	0.74	0.81
		LGG	0.65	0.53	0.00	0.54	0.42	0.00	0.86	0.86	0.00
		Combined	0.84	0.72	0.62	0.85	0.82	0.60	0.86	0.76	0.68
	Using pre-processing as in [19]	HGG	0.87	0.74	**0.71**	**0.89**	**0.92**	0.73	0.86	0.69	0.75
		LGG	0.34	0.33	0.00	0.29	0.29	0.00	0.63	0.44	0.00
		Combined	0.78	0.67	**0.60**	**0.79**	0.82	0.61	0.82	0.65	0.63
	Using no training samples from HGG into LGG	HGG	0.88	0.76	0.73	0.91	0.90	0.72	0.86	0.74	0.81
		LGG	0.46	0.34	0.00	0.37	0.27	0.00	0.71	0.63	0.00
		Combined	0.81	0.69	0.62	0.82	0.80	0.60	0.84	0.72	0.68
	Using no rotations	HGG	0.87	0.77	**0.73**	**0.86**	**0.83**	**0.70**	**0.89**	**0.78**	0.83
		LGG	0.47	0.31	0.00	0.39	0.25	0.00	0.68	0.66	0.00
		Combined	**0.80**	0.69	**0.61**	**0.78**	**0.74**	**0.59**	**0.85**	0.76	0.70
	Random rotations (5.625°)	HGG	0.87	0.77	0.74	**0.92**	**0.89**	**0.76**	**0.84**	**0.76**	**0.79**
		LGG	0.62	0.49	0.00	0.49	0.38	0.00	0.91	0.87	0.00
		Combined	**0.83**	0.72	0.62	0.85	**0.81**	**0.64**	**0.85**	**0.78**	**0.66**
	Using ReLU	HGG	0.87	0.77	0.73	**0.87**	**0.88**	**0.69**	**0.89**	**0.77**	**0.86**
		LGG	0.53	0.47	0.00	0.40	0.37	0.00	0.86	0.89	0.00
		Combined	0.82	0.72	0.61	**0.79**	**0.80**	**0.58**	**0.88**	**0.79**	**0.72**
	Large/small kernels 1	HGG	**0.85**	0.74	0.72	**0.92**	**0.87**	**0.73**	**0.81**	**0.71**	**0.77**
		LGG	0.52	0.36	0.00	0.42	0.27	0.00	0.84	0.71	0.00
		Combined	**0.80**	0.68	0.60	**0.84**	0.77	**0.61**	**0.81**	**0.71**	**0.65**
	Large/small kernels 2	HGG	**0.85**	0.74	0.72	**0.92**	**0.91**	**0.78**	**0.81**	**0.71**	**0.77**
		LGG	0.52	0.34	0.00	0.42	0.26	0.00	0.85	0.71	0.00
		Combined	**0.79**	**0.67**	0.60	**0.84**	**0.81**	**0.66**	**0.81**	**0.71**	**0.64**
	Coronal patches	HGG	**0.86**	0.75	0.72	**0.88**	**0.83**	**0.74**	0.86	0.74	**0.76**
		LGG	0.59	0.44	0.00	0.46	0.34	0.00	0.92	0.86	0.00
		Combined	**0.82**	**0.70**	0.61	**0.81**	**0.75**	**0.62**	0.87	0.76	**0.64**
	Sagittal patches	HGG	**0.86**	0.75	**0.71**	**0.86**	**0.79**	**0.70**	0.87	**0.78**	**0.79**
		LGG	0.45	0.32	0.00	0.36	0.26	0.00	0.87	0.70	0.00
		Combined	**0.79**	**0.68**	**0.60**	**0.78**	**0.70**	**0.59**	0.87	0.76	**0.66**
Challenge	**Proposed**	HGG	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81
	Using pre-processing as in [19]	HGG	0.80	**0.78**	0.73	**0.75**	0.86	0.71	0.92	**0.74**	0.77
	Using no rotations	HGG	**0.85**	**0.79**	0.74	**0.81**	**0.78**	**0.70**	**0.91**	**0.86**	0.82
	Random rotations (5.625°)	HGG	0.88	0.82	0.76	**0.90**	**0.84**	**0.76**	**0.86**	0.84	**0.78**
	Using ReLU	HGG	**0.86**	0.81	**0.74**	**0.82**	**0.80**	**0.66**	**0.90**	**0.85**	**0.86**
	Large kernels/shallow arq. 1	HGG	0.87	**0.81**	**0.75**	**0.90**	**0.89**	0.76	**0.84**	**0.78**	**0.76**
	Large kernels/shallow arq. 2	HGG	0.87	**0.81**	**0.75**	**0.90**	0.88	**0.75**	**0.84**	**0.78**	**0.76**
	Coronal patches	HGG	**0.85**	**0.81**	**0.74**	**0.81**	**0.85**	0.75	0.90	0.79	**0.75**
	Sagittal patches	HGG	**0.84**	**0.76**	0.73	**0.78**	**0.73**	**0.67**	**0.92**	0.85	0.82

component, we verify that this pre-processing was the key component for improving the segmention in LGG. The result of this experiment in both grades is interesting, because we know that the features learned by the CNN are computed in local regions by a bank of band-pass filters at different scales, instead of point-wise properties as an intensity. Shah [37] presented a study regarding the segmentation of multiple sclerosis based on MRI images, showing that classifiers based on point-wise features, as intensity, improved after Nyúl normalization. This improvement was obtained by minimizing the data heterogeneity from multi-site multi-scanner MRI acquisitions. However, our experiment gives evidence that in MRI applications, CNN-based classifiers also improve after Nyúl normalization, at least in the context of brain tumor segmentation. Additionally, we further investigated the effect of increasing the number of training epochs until 90 epochs, but we obtained no improvement with the simpler pre-processing. Referring to Fig. 4, we can observe that the proposed pre-processing enabled a better training of the CNN, such that the segmentation presented a better delineation of the non-enhancing and the necrosis regions in both data sets.

b) Data Augmentation: Artificial data augmentation is a common procedure in the context of CNN, when the data set is relatively small. In the case of MRI images, we have a large number of samples for healthy and tumorous tissue, which may be the reason why most recent studies on brain tumor segmentation based on Deep Learning [30], [31], [33] did not explore data augmentation. Havaei et al. [32] considered its application, but found to be ineffective in their system.

We investigated two types of data augmentation. In the first case, we studied the effect of data augmentation by increasing the number of samples using rotations. In this study, we evaluated two variants. In the first, we used multiples of 90° (90°, 180° and 270°) for rotations (corresponding to the Proposed method), while in the second, we sampled three rotation angles from an array using an uniform distribution, whose angles were equally spaced. The angle step was defined as $\alpha \times 90^\circ$ with $\alpha \in \{1/8, 1/16, 1/32\}$. In this second variant, we consider $\alpha = 1/16$[6]. In Table IV, we present the results with each variant and without rotations in both the Leaderboard

[6]The case $\alpha = 1/16$ was the best result; the other results can be found in the online *Supplementary Material*.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

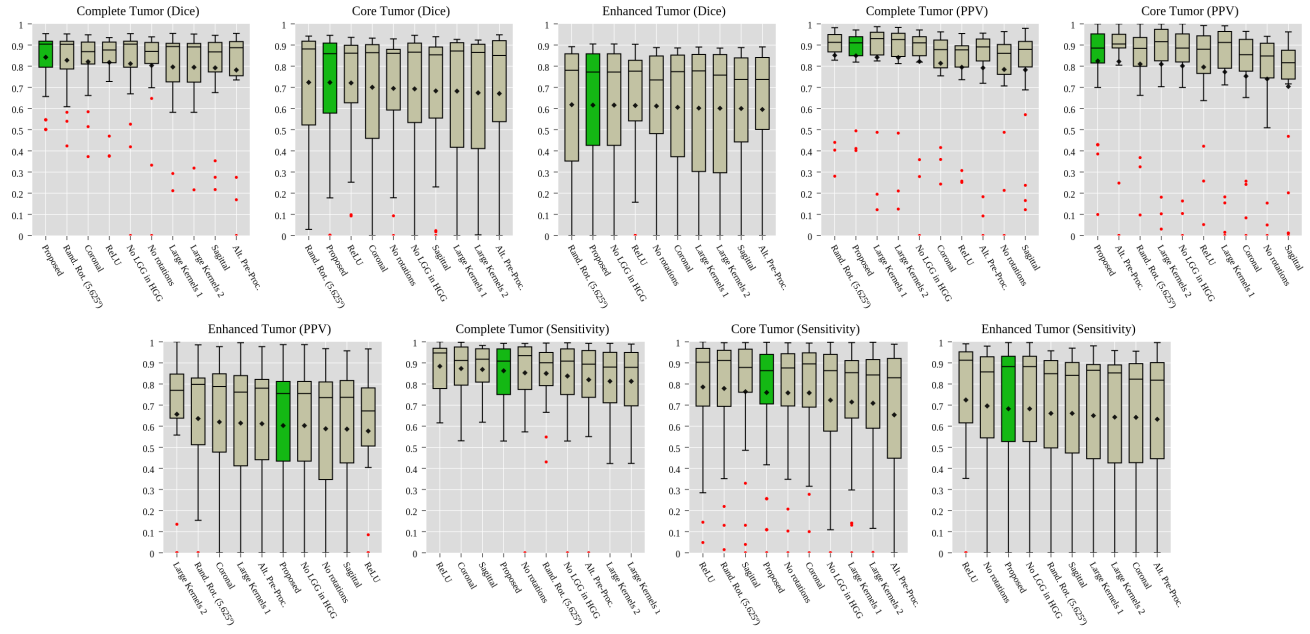


Fig. 2: Boxplot for each of the experiments in Table IV in the Leaderboard data set. The boxplot for the experiment of sampling training samples from HGG into LGG is not shown given the reduced number of subjects (4 LGG in 25 subjects for the Leaderboard data set). The diamond marks the mean.

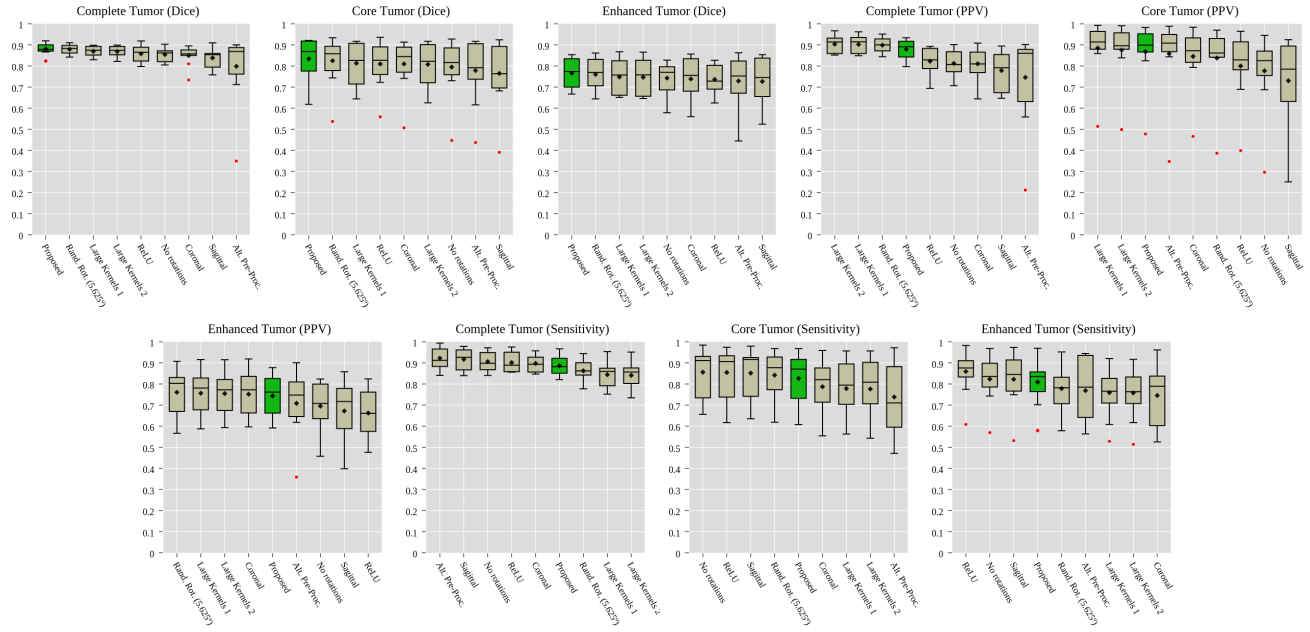
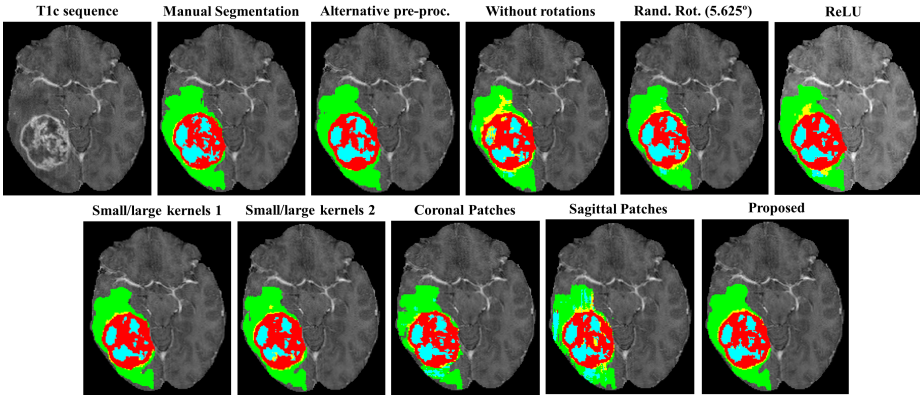


Fig. 3: Boxplot for each of the experiments in Table IV in the Challenge data set. The diamond marks the mean.

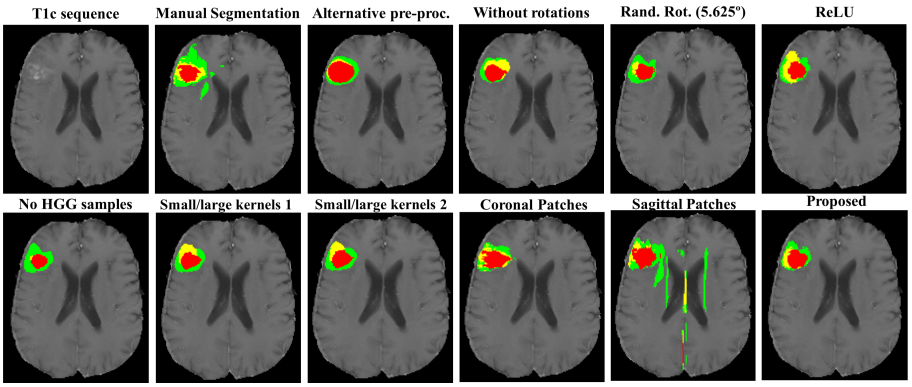
and Challenge data sets. As can be observed, the rotations improved the performance in all regions for the DSC and PPV; but, we also note a decrease in the sensitivity for both variants in the Challenge data set. However, the mean gain obtained by including rotations was 2.6% (Leaderboard: 2.6%, Challenge: 2.7%) for the first variant (Proposed) and 2.3% (Leaderboard: 2.7%, Challenge: 2.0%) for the second variant. Comparing the two variants, we obtain a mean gain of 0.3% of the first variant in relation to the second. Also, the first variant has the advantage of being faster to compute. Observing Fig. 4, we conclude that the extra information provided by the rotations of the first variant in training the CNN resulted in segmentations with a better delineation of the complete tumor as well as of the intra-tumoral structures. In both grades, we have an excess of non-enhancing class, when we trained without data augmentation, and for HGG this class is even found inside the region formed by enhancing and necrotic structures, which does not happen in the manual segmentation.

Brain tumors are constituted by intra-tumoral structures with very different volumes, resulting in an imbalanced number

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging



(a)



(b)

Fig. 4: Examples of segmentations obtained with cross-validation, showing the effect of each component of the proposed method. In the first row, we have a HGG, and in the bottom row a LGG. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor and red – enhancing tumor.

of samples of each class. This underrepresentation of some classes impairs the performance of the CNN. So, we investigated a second type of data augmentation to balance the number of samples of each class, which consisted in extracting samples from necrosis and enhancing tumor regions in HGG to use as training samples in LGG. In Table IV, we compare the proposed approach, in which the number of samples of each tumor class in LGG were more balanced, with another experiment that uses only samples from LGG. We verify that the extra samples from HGG improved all metrics for the complete and core regions in the Leaderboard data set with a mean gain of 1.9%. Examining Fig. 4(b), we note that by sampling from HGG to LGG, we improved the training of the CNN. Observe that the tumor segmentation presented a better delineation of the enhancing and non-enhancing regions, although the sampling was only for enhancing and necrosis regions. The improvement of the non-enhaced region could be explained by the context introduced by the patches of the enhancing samples, since these two regions are next to each other.

c) Activation Function: The gradients of ReLU are zero when the unit is not active, which may slow down the convergence during the optimization and lead to worst training. To avoid that problem, Maas [43] proposed LReLU as an alternative nonlinearity. So, we investigate the effectiveness of this activation function in brain tumor segmentation. In this experiment, only the activation function was changed in the proposed method. The results in both the Leaderboard and Challenge data sets are presented in Table IV. We verify that LReLU activation improved the performance of the proposed method in both data sets in the DSC and PPV, with the exception of the core in the DSC in the Leaderboard data set. ReLU activations presented better scores in the Sensitivity metric. However, the mean gain using LReLU instead of ReLU was 1.3% (Leaderboard: 0.44%, Challenge: 2.2%). Referring to Fig. 4, we find that using ReLU as an activation function resulted in an excessive segmentation of non-enhancing and necrosis regions outside the core for HGG.

d) Deeper architectures/small kernels: Using cascaded layers with small 3×3 kernels has the advantage of maintaining the same effective receptive field of bigger kernels, while reducing the number of weights, and allowing more non-linear transformations on the data. To evaluate the real impact of this technique on brain tumor segmentation, we changed the cascaded convolutional layers before each max-pooling of the proposed architecture by one layer with larger kernels with the equivalent effective receptive field. So, in HGG we changed the groups of layers 1, 2, 3 and 5, 6, 7

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

(Table I) by one convolutional layer with 7×7 kernels each, while in the LGG we changed the groups of layers 1 and 2, and 4 and 5 (Table II) by one layer with 5×5 kernels each. Using these architectures, we experimented two variants for both grades: 1) we maintained the 64 feature maps in the first convolutional layer and 128 in the second; 2) we increased the capacity of the CNN by using wider layers, namely, 128 feature maps in the first convolutional layer and 256 in the second. We present the results obtained in the Leaderboard and Challenge data sets in Table IV and the boxplots in Fig. 2 and 3. In relation to variant 1, the mean gain was 2.4% (Leaderboard: 3.1%, Challenge: 1.6%), while for variant 2 it was 2.1% (Leaderboard: 2.4%, Challenge: 1.8%). In the majority of metrics, the proposed method obtained higher scores than both variants with bigger kernels, with some of them with statistical significance, while the variants achieved better scores in PPV (HGG in both data sets). In the boxplots, both variants seem to have larger dispersion and more outliers. In the segmentations of Fig. 4, although the segmentations by the variants appear with good quality, the proposed method can capture more details, and variant 2 classified some non-enhanced tumor inside the enhancing ring, which does not happen in the manual segmentation in HGG; in LGG the architecture with bigger kernels also identified an excess of non-enhancing tumor.

B. Patch Extraction Plane

The use of 2D patches in a MRI image requires that we define a plane perpendicular to an axis to extract patches. So, following the procedure defined in the previous subsection, we investigated the use of patches extracted in a plane perpendicular to the Axial, Coronal, and Sagittal axis. The results in both the Leaderboard and Challenge data sets are presented in Table IV. As can be observed, extracting patches in the plane perpendicular to the Axial axis presented the best overall performance with a mean gain of 2.33% relative to the Coronal plane (Leaderboard: 1.89%, Challenge: 2.78%) and 4.00% relative to the Sagittal plane (Leaderboard: 3.56%, Challenge: 4.44%). The Axial plane presented better DSC and PPV scores for both data sets than the Sagittal plane, but worst sensitivity for the Challenge data set and for the complete region in the Leaderboard data set. Considering Fig. 4, this can be explained by an over-segmentation of the tumor, which is corroborated by the lower PPV score. A similar pattern is found for the Coronal plane, which was better in the enhanced region for the PPV score and in the complete region for the Sensitivity score. The better performance obtained using patches extracted in the Axial plane can be explained by some acquisitions having lower spatial resolution in the Coronal and Sagittal planes, which can be considered a limitation of the BRATS databases.

Finally, as an overall analysis, we note some general trends across all experiments. Considering the boxplots, Fig. 2 and 3, we verify a lower dispersion for the complete region, presenting also a higher mean value for the same region. This lower dispersion is less expressive in the Leaderboard than in the Challenge data set, which may be explained by the worst performance of the algorithms on LGG subjects in this data set. Another general trend is found in Table IV that shows that none of the algorithms found presence of enhanced region among the LGG subjects[7].

C. Global Validation

In Table V, we compile the results of the top 5 methods in the Leaderboard and Challenge data sets of BRATS 2013 (including the proposed method). We also include the proposals by Havaei et al. [32], Davy et al. [31], and Urban et al. [30] that are based on CNN. Appraising the results in Table V, we conclude that no method is yet able to achieve the first place in all metrics and regions for brain tumor segmentation; but, the proposed method obtained the first position in DSC in the three regions (Challenge data set), according to the online evaluation platform [47]. Also, based on the same evaluation, the proposed method obtained the overall first position in both data sets, outperforming the other methods.

Assessing the CNN-based methods, we observe that two of those methods [30], [31] had modest performances, comparing with others not based on CNN; however, the method proposed by Havaei et al. [32] exhibit higher metrics. They propose a novel and elaborated training and concatenation of two CNNs to capture more context into the training. Contrasting the two methods, while our method is better in Sensitivity, their method is in PPV in the complete region. According to Menze et al. [5], the most difficult tasks in brain tumor segmentation are the segmentation of the core region for LGG and the enhancing region for HGG. In these two tasks our method outperformed Havaei et al. [32]. We note a larger difference in the core region in the Challenge data, which is considered an easier region to segment according to Menze et al. [5]. Based on the analysis of the key components in the previous section, we conclude that although our architecture is simpler, those components permitted a better training of our CNN classifier, compensating the lack of information of a larger context, which according to the experiments reported by Havaei et al. [32] was found to be relevant.

The method proposed by Kwon et al. [11] is ranked in the second place in both data sets. They perform a joint segmentation and registration using a tumor growth model to transform an atlas of healthy patients into one with the tumor and its intra-tumoral structures. Given the complex shape of tumors, they refine the initial solution using the Expectation Maximization algorithm. Comparing their approach with ours in the Challenge data set (HGG tumors), our method obtained higher DSC in the enhancing region and in Sensitivity in the three regions, and their method was better in PPV in the complete and core regions. For the Leaderboard data set (LGG and HGG tumors), our method obtained higher metrics in the enhancing region in DSC and Sensitivity, and their method was better in the complete region in PPV and DSC, and in the core region in the three metrics. Another strong contender in the Leaderboard is the method proposed by Zhao [5]. His method was better in the core region in DSC and outperformed all methods in the complete and core regions in sensitivity;

[7]We currently are not able to confirm if these tumors contain enhanced tumor, since the expert segmentation is private.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

TABLE V: Results in the Leaderboard and Challenge data sets of BRATS 2013. The relative rank refers to the combination of the ranking in each metric for the referred class, while the position is the global ranking, as provided by the online evaluation platform [47].

	Methods	DSC			PPV			Sensitivity			Relative Rank			Position
		Complete	Core	Enh.	Complete	Core	Enh.	Complete	Core	Enh.	Complete	Core	Enh.	
Leaderboard	**Proposed**	0.84	0.72	0.62	0.85	0.82	0.60	0.86	0.76	0.68	3.67	3.33	1.67	1
	Kwon et al. [11]	0.86	0.79	0.59	0.88	0.84	0.60	0.86	0.81	0.63	3.33	1.67	5.00	2
	Zhao et al.[8] [5]	0.83	0.73	0.55	0.77	0.67	0.46	0.94	0.89	0.78	4.67	4.00	9.33	3
	agnm1[9]	0.83	0.71	0.54	0.85	0.73	0.59	0.84	0.82	0.58	6.00	4.33	10.33	4
	havam2[9]	0.82	0.69	0.56	0.83	0.77	0.62	0.83	0.69	0.58	7.67	7.00	8.00	5
	Urban et al.[9] [30]	0.70	0.57	0.54	0.65	0.55	0.52	0.87	0.67	0.60	14.00	18.67	12.33	17
	Havaei et al.[10] [32]	0.84	0.71	0.57	0.88	0.79	0.54	0.84	0.72	0.68	–	–	–	–
	Davy et al. [31]	0.72	0.63	0.56	0.69	0.64	0.50	0.82	0.68	0.68	–	–	–	–
Challenge	**Proposed**	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81	7.00	3.33	5.33	1
	Kwon et al. [11], [52]	0.88	0.83	0.72	0.92	0.90	0.74	0.84	0.78	0.72	9.33	5.00	13.00	2
	Tustison et al. [19]	0.87	0.78	0.74	0.85	0.74	0.69	0.89	0.88	0.83	10.33	11.67	9.00	3
	havam2[9]	0.88	0.78	0.73	0.89	0.79	0.68	0.87	0.79	0.80	8.33	10.67	13.33	4
	al-ss1[9]	0.87	0.78	0.70	0.89	0.83	0.75	0.86	0.78	0.70	9.67	8.67	14.67	5
	Urban et al.[8] [30]	0.86	0.75	0.73	0.82	0.75	0.79	0.92	0.79	0.70	11.67	16.00	11.67	12
	Havaei et al. [32]	0.88	0.79	0.73	0.89	0.79	0.68	0.87	0.79	0.80	–	–	–	–
	Davy et al. [31]	0.85	0.74	0.68	0.85	0.74	0.62	0.85	0.78	0.77	–	–	–	–

[8] Results retrieved from [47] using the cited method.

[9] Results retrieved from [47], but the method or author are unknown.

[10] Results provided by the author using the cited method.

however, since we note a significant drop in performance in PPV in the same regions, we may infer that probably the method by Zhao oversegmented the tumor. Also, we note another trend, both our and Kwon methods drop from the Challenge to the Leaderboard data set in most metrics (Kwon improved in the complete and core regions in sensitivity); however, appraising the separated metrics for HGG and LGG in the Leaderboard, Table V, we observe that the performance of our method was similar in the complete and enhanced regions but dropped more significantly in the core region in DSC and in sensitivity; therefore, given the lower metric of LGG, we hypothesize that the general drop in both methods from the Challenge data set to the Leaderboard was mainly due to the LGG subjects; however, the method proposed by Kwon dropped less in the core region. Considering the performance in both data sets, we argue that both methods were similar in segmenting the complete tumor, the method proposed by Kwon was in general better in the core region and our method in delineating the enhanced structure. Assessing the running times, Kwon reports an average running time of 85 min. on an Intel Core i7 3.4 GHz machine, while our full pipeline presents an average running time of 8 min. using a GPU NVIDIA GeForce GTX 980 equipped on an Intel Core i7 3.5 GHz machine. This difference in running times is explained by our method performing an optimization only during training, which permits a fast segmentation during normal use.

Considering the state of the art, we verify that current CNN-based approaches [29]–[35] have used larger filters and shallow architectures, with some using features computed by the CNN as input to a RF [34], or employing the network for structured prediction [35]. Also, these works did not explore the stacking of several layers to apply more non-linearities on the data, which we showed to be important. In the CNN, these authors have used more common non-linearities, as hyperbolic tangent or ReLU; however, our experiments indicate that LReLU is a strong alternative to ReLU and do not suffer of the limitations of the hyperbolic tangent [43]. Although some authors found no advantage in using data augmentation [32], we have shown that data augmentation and the adequate pre-processing have a significant impact on performance. Based on these facts, our conclusion is that the contributions in this article are orthogonal to current state of the art, existing potential for further improvement in brain tumor segmentation using MRI images by looking for synergies with the techniques studied by current works.

In Fig. 5, we present the segmentation of two patients with HGG and LGG, respectively, from the Leaderboard data set. Fig. 6 shows a patient with two tumors that were correctly detected and segmented from the Challenge data set.

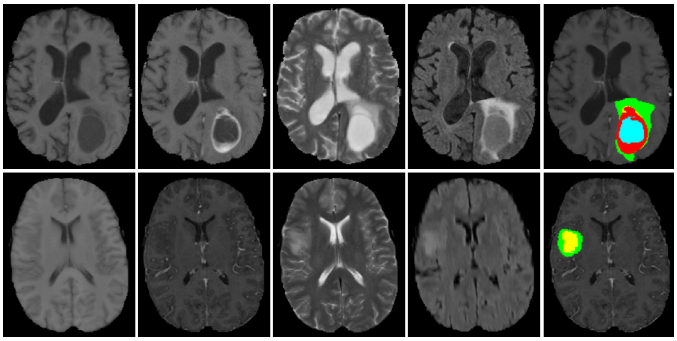


Fig. 5: Examples of segmentations in the Leaderboard data set, showing a HGG in the first row (subject id: 210) and a LGG in the bottom row (subject id: 105). From left to right: T1, T1c, T2, FLAIR, and the segmentation. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor, and red – enhancing tumor.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMI.2016.2538465, IEEE Transactions on Medical Imaging

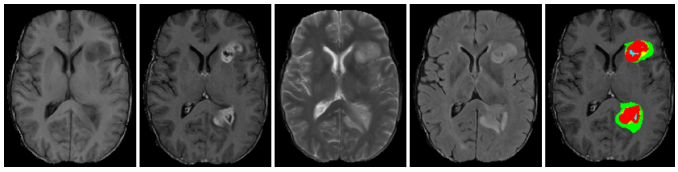


Fig. 6: Examples of segmentations in the Challenge data set (subject id: 310). From left to right: T1, T1c, T2, FLAIR, and the segmentation. Each color represents a tumor class: green – edema, blue – necrosis, yellow – non-enhancing tumor, and red – enhancing tumor.

D. Participation on BRATS 2015 Challenge

The proposed architecture was also used to segment the BRATS 2015 Challenge data set. The differences when comparing with the models trained in BRATS 2013 were the number of samples for training, given the bigger size of the Training set, and in Dropout (p) that was increased to 0.5 in the HGG architecture. In this data set, our method obtained the second position with a DSC score of 0.78, 0.65, and 0.75 in the complete, core, and enhanced regions, respectively, as computed by the BRATS organization; the boxplots are presented in Fig. 7.

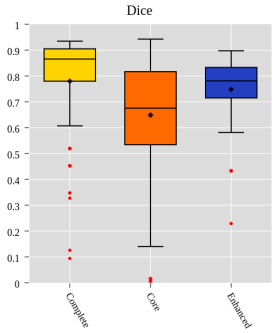


Fig. 7: Boxplots of DSC obtained in the Challenge set of BRATS 2015. The diamond marks the mean.

V. Conclusions

In summary, we propose a novel CNN-based method for segmentation of brain tumors in MRI images. We start by a pre-processing stage consisting of bias field correction, intensity and patch normalization. After that, during training, the number of training patches is artificially augmented by rotating the training patches, and using samples of HGG to augment the number of rare LGG classes. The CNN is built over convolutional layers with small 3×3 kernels to allow deeper architectures.

In designing our method, we address the heterogeneity caused by multi-site multi-scanner acquisitions of MRI images using intensity normalization as proposed by Nyúl et al. We show that this is important in achieving a good segmentation. Brain tumors are highly variable in their spatial localization and structural composition, so we have investigated the use of data augmentation to cope with such variability. We studied augmenting our training data set by rotating the patches as well as by sampling from classes of HGG that were underrepresented in LGG. We found that data augmentation was also quite effective, although not thoroughly explored in Deep Learning methods for brain tumor segmentation. Also, we investigated the potential of deep architectures through small kernels by comparing our deep CNN with shallow architectures with larger filters. We found that shallow architectures presented a lower performance, even when using a larger number of feature maps. Finally, we verified that the activation function LReLU was more important than ReLU in effectively training our CNN.

We evaluated the proposed method in BRATS 2013 and 2015 databases. Concerning 2013 database, we were ranked in the first position by the online evaluation platform. Also, it was obtained simultaneously the first position in DSC metric in the complete, core, and enhancing regions in the Challenge data set. Comparing with the best generative model [11], we were able to reduce the computation time approximately by ten-fold. Concerning the 2015 database, we obtained the second position among twelve contenders in the on-site challenge. We argue, therefore, that the components that were studied have potential to be incorporated in CNN-based methods and that as a whole our method is a strong candidate for brain tumor segmentation using MRI images.

Acknowledgments

The authors would like to thank the questions and suggestions of the Anonymous Reviewers that helped to improve this document. This work is supported by FCT with the reference project UID/EEA/04436/2013, by FEDER funds through the COMPETE 2020 Programa Operacional Competitividade e Internacionalização (POCI) with the reference project POCI-01-0145-FEDER-006941. Sérgio Pereira was supported by a scholarship from the Fundação para a Ciência e Tecnologia (FCT), Portugal (scholarship number PD/BD/105803/2014). Brain tumor image data used in this article were obtained from the MICCAI 2013 and 2015 Challenges on Multimodal Brain Tumor Segmentation. The challenge database contain fully anonymized images from the Cancer Imaging Atlas Archive and the BRATS 2012 challenge.

References

[1] S. Bauer *et al.*, "A survey of mri-based medical image analysis for brain tumor studies," *Physics in medicine and biology*, vol. 58, no. 13, pp. 97–129, 2013.

[2] D. N. Louis *et al.*, "The 2007 who classification of tumours of the central nervous system," *Acta neuropathologica*, vol. 114, no. 2, pp. 97–109, 2007.

[3] E. G. Van Meir *et al.*, "Exciting new advances in neuro-oncology: The avenue to a cure for malignant glioma," *CA: a cancer journal for clinicians*, vol. 60, no. 3, pp. 166–193, 2010.

[4] G. Tabatabai *et al.*, "Molecular diagnostics of gliomas: the clinical perspective," *Acta neuropathologica*, vol. 120, no. 5, pp. 585–592, 2010.

[5] B. Menze *et al.*, "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2015.

[6] N. J. Tustison *et al.*, "N4itk: improved n3 bias correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[7] L. G. Nyúl, J. K. Udupa, and X. Zhang, "New variants of a method of mri scale standardization," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 143–150, 2000.

[8] M. Prastawa *et al.*, "A brain tumor segmentation framework based on outlier detection," *Medical image analysis*, vol. 8, no. 3, pp. 275–283, 2004.

[9] B. H. Menze *et al.*, "A generative model for brain tumor segmentation in multi-modal images," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 151–159.

[10] A. Gooya *et al.*, "Glistr: glioma image segmentation and registration," *IEEE Transactions on Medical Imaging*, vol. 31, no. 10, pp. 1941–1954, 2012.

[11] D. Kwon *et al.*, "Combining generative models for multifocal glioma segmentation and registration," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 763–770.

[12] S. Bauer, L.-P. Nolte, and M. Reyes, "Fully automatic segmentation of brain tumor images using support vector machine classification in combination with hierarchical conditional random field regularization," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011*. Springer, 2011, pp. 354–361.

[13] C.-H. Lee *et al.*, "Segmenting brain tumors using pseudo–conditional random fields," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2008*. Springer, 2008, pp. 359–366.

[14] R. Meier *et al.*, "A hybrid model for multimodal brain tumor segmentation," in *Proceedings of NCI-MICCAI BRATS*, 2013, pp. 31–37.

[15] ——, "Appearance-and context-sensitive features for brain tumor segmentation," in *MICCAI Brain Tumor Segmentation Challenge (BraTS)*, 2014, pp. 20–26.

[16] D. Zikic *et al.*, "Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel mr," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012*. Springer, 2012, pp. 369–376.

[17] S. Bauer *et al.*, "Segmentation of brain tumor images based on integrated hierarchical classification and regularization," *Proceddings of MICCAI-BRATS*, pp. 10–13, 2012.

[18] S. Reza and K. Iftekharuddin, "Multi-fractal texture features for brain tumor and edema segmentation," in *SPIE Medical Imaging*. International Society for Optics and Photonics, 2014, pp. 903 503–903 503.

[19] N. Tustison *et al.*, "Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with antsr," *Neuroinformatics*, vol. 13, no. 2, pp. 209–225, 2015.

[20] E. Geremia, B. H. Menze, and N. Ayache, "Spatially adaptive random forests," in *2013 IEEE 10th International Symposium on Biomedical Imaging (ISBI)*. IEEE, 2013, pp. 1344–1347.

[21] A. Pinto *et al.*, "Brain tumour segmentation based on extremely randomized forest with high-level features," in *Engineering in Medicine and Biology Society (EMBC), 2015 37th Annual International Conference of the IEEE*. IEEE, 2015, pp. 3037–3040.

[22] A. Islam, S. Reza, and K. M. Iftekharuddin, "Multifractal texture estimation for detection and segmentation of brain tumors," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 11, pp. 3204–3215, 2013.

[23] R. Meier *et al.*, "Patient-specific semi-supervised learning for postoperative brain tumor segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014*. Springer, 2014, pp. 714–721.

[24] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1798–1828, 2013.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[26] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[27] S. Dieleman, K. W. Willett, and J. Dambre, "Rotation-invariant convolutional neural networks for galaxy morphology prediction," *Monthly Notices of the Royal Astronomical Society*, vol. 450, no. 2, pp. 1441–1459, 2015.

[28] D. Ciresan *et al.*, "Deep neural networks segment neuronal membranes in electron microscopy images," in *Advances in neural information processing systems*, 2012, pp. 2843–2851.

[29] D. Zikic *et al.*, "Segmentation of brain tumor tissues with convolutional neural networks," *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, pp. 36–39, 2014.

[30] G. Urban *et al.*, "Multi-modal brain tumor segmentation using deep convolutional neural networks," *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, pp. 1–5, 2014.

[31] A. Davy *et al.*, "Brain tumor segmentation with deep neural networks," *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, pp. 31–35, 2014.

[32] M. Havaei *et al.*, "Brain tumor segmentation with deep neural networks," *arXiv:1505.03540v1*, 2015. [Online]. Available: http://arxiv.org/abs/1505.03540

[33] M. Lyksborg *et al.*, "An ensemble of 2d convolutional neural networks for tumor segmentation," in *Image Analysis*. Springer, 2015, pp. 201–211.

[34] V. Rao, M. Sharifi, and A. Jaiswal, "Brain tumor segmentation with deep learning," *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, pp. 56–59, 2015.

[35] P. Dvořák and B. Menze, "Structured prediction with convolutional neural networks for multimodal brain tumor segmentation," *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, pp. 13–24, 2015.

[36] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556v6*, 2014. [Online]. Available: http://arxiv.org/abs/1409.1556

[37] M. Shah *et al.*, "Evaluating intensity normalization on mris of human brain with multiple sclerosis," *Medical image analysis*, vol. 15, no. 2, pp. 267–282, 2011.

[38] L. Nyúl and J. Udupa, "On standardizing the mr image intensity scale," *Magnetic Resonance in Medicine*, vol. 42, no. 6, pp. 1072–1081, 1999.

[39] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, 1989.

[40] ——, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[41] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International conference on artificial intelligence and statistics*, 2010, pp. 249–256.

[42] K. Jarrett *et al.*, "What is the best multi-stage architecture for object recognition?" in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 2146–2153.

[43] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, vol. 30, 2013.

[44] N. Srivastava *et al.*, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[45] G. E. Hinton *et al.*, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv preprint arXiv:1207.0580v1*, 2012. [Online]. Available: http://arxiv.org/abs/1207.0580

[46] M. Kistler *et al.*, "The virtual skeleton database: An open access repository for biomedical research and collaboration," *Journal of Medical Internet Research*, vol. 15, no. 11, Nov 2013.

[47] VirtualSkeleton, *BRATS 2013*, 2013. [Online]. Available: https://www.virtualskeleton.ch/BRATS/Start2013 [Accessed: September 30, 2015]

[48] F. Bastien *et al.*, "Theano: new features and speed improvements," Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop, 2012.

[49] J. Bergstra *et al.*, "Theano: a CPU and GPU math expression compiler," in *Proceedings of the Python for Scientific Computing Conference (SciPy)*, Jun. 2010.

[50] S. Dieleman *et al.*, "Lasagne: First release." Aug. 2015. [Online]. Available: http://dx.doi.org/10.5281/zenodo.27878

[51] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[52] D. Kwon *et al.*, "Multimodal brain tumor image segmentation using glistr," in *MICCAI Multimodal Brain Tumor Segmentation Challenge (BraTS)*, 2014, pp. 18–19.