NeuroImage 129 (2016) 460–469

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Deep MRI brain extraction: A 3D convolutional neural network for skull stripping

Jens Kleesiek [a,b,c,d,*,1], Gregor Urban [a,1], Alexander Hubert [a], Daniel Schwarz [a], Klaus Maier-Hein [b], Martin Bendszus [a], Armin Biller [a,d]

[a] MDMI Lab, Division of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany
[b] Junior Group Medical Image Computing, German Cancer Research Center, Heidelberg, Germany
[c] Heidelberg University HCI/IWR, Heidelberg, Germany
[d] Division of Radiology, German Cancer Research Center, Heidelberg, Germany

ARTICLE INFO

Article history:
Received 11 August 2015
Accepted 11 January 2016
Available online 22 January 2016

Keywords:
MRI
Brain extraction
Brain mask
Skull stripping
Deep learning
Convolutional networks

ABSTRACT

Brain extraction from magnetic resonance imaging (MRI) is crucial for many neuroimaging workflows. Current methods demonstrate good results on non-enhanced T1-weighted images, but struggle when confronted with other modalities and pathologically altered tissue. In this paper we present a 3D convolutional deep learning architecture to address these shortcomings. In contrast to existing methods, we are not limited to non-enhanced T1w images. When trained appropriately, our approach handles an arbitrary number of modalities including contrast-enhanced scans. Its applicability to MRI data, comprising four channels: non-enhanced and contrast-enhanced T1w, T2w and FLAIR contrasts, is demonstrated on a challenging clinical data set containing brain tumors (N = 53), where our approach significantly outperforms six commonly used tools with a mean Dice score of 95.19. Further, the proposed method at least matches state-of-the-art performance as demonstrated on three publicly available data sets: IBSR, LPBA40 and OASIS, totaling N = 135 volumes. For the IBSR (96.32) and LPBA40 (96.96) data set the convolutional neuronal network (CNN) obtains the highest average Dice scores, albeit not being significantly different from the second best performing method. For the OASIS data the second best Dice (95.02) results are achieved, with no statistical difference in comparison to the best performing tool. For all data sets the highest average specificity measures are evaluated, whereas the sensitivity displays about average results. Adjusting the cut-off threshold for generating the binary masks from the CNN's probability output can be used to increase the sensitivity of the method. Of course, this comes at the cost of a decreased specificity and has to be decided application specific. Using an optimized GPU implementation predictions can be achieved in less than one minute. The proposed method may prove useful for large-scale studies and clinical trials.

Introduction

Brain Extraction, a.k.a. skull stripping, from magnetic resonance imaging (MRI) data is an essential step in many neuroimaging applications, amongst others surgical planning, cortical structure analysis (Fischl et al., 1999; Thompson et al., 2001), surface reconstruction (Tosun et al., 2006) and thickness estimation (MacDonald et al., 2000), as well as image registration (Klein et al., 2010a; Woods et al., 1993) and tissue segmentation (de Boer et al., 2010; Menze et al., 2014; Shattuck et al., 2001; Wang et al., 2010; Zhang et al., 2001; Zhao et al., 2010). It is desirable to automate this procedure to reduce human rater variance and eliminate time-consuming manual processing steps that potentially impede not only the analysis, but also the reproducibility of large-scale (clinical) studies.

It has been shown that several factors affect the outcome of methods devised for removing non-brain tissue. These include imaging artifacts, different MRI scanners and protocols that in turn lead to contrast and intensity variations. Further, anatomical variability, as well as age and the extent of brain atrophy, e.g. due to neurodegeneration, have been reported as influencing the results of current brain extraction methods (Fennema-Notestine et al., 2006). The problem becomes even more severe when considering MR brain scans of pathological conditions such as brain tumors. The disease and its treatment-related changes, e.g. resection cavities and radiation effects usually considerably alter the brain structure. The voxel intensity distribution is impaired either by the disease itself, for instance due to a large edema, or due to administration of contrast agent during the examination (Speier et al.,

Abbreviations: GT, ground truth; T1w, T1-weighted; nT1w, non-enhanced T1w; ceT1w, contrast-enhanced T1-weighted; T2w, T2-weigthed; FLAIR, fluid attenuated inversion recovery; RF, random forest; CSF, cerebrospinal fluid; MFP, max-fragment-pooling.

* Corresponding author at: MDMI Lab, Division of Neuroradiology, Heidelberg University Hospital, Heidelberg, Germany.
E-mail address: kleesiek@uni-heidelberg.de (J. Kleesiek).

[1] Contributed equally.

http://dx.doi.org/10.1016/j.neuroimage.2016.01.024
1053-8119/

2011).This usually leads to a decreased segmentation accuracy of existing methods.

Histologically the border of the brain is defined at the transition of myelination from oligodendroglia to Schwann cells. This microscopy delineation cannot be captured in MR images yet. Thus, under physiological conditions alone, detecting the boundaries of the brain in MR images is considered a difficult problem (Iglesias et al., 2011; Wang et al., 2014). This is due to the complex nature of the data that is often of low contrast and resolution, and contains partial volume effects leading to ill-defined boundaries. Moreover, the problem itself is not well defined because the "ground truth" (GT) varies based on the guidelines specified for the manual delineation by experts (Eskildsen et al., 2012). These include gross differences, such as the decision whether or not to include the brain stem (Iglesias et al., 2011), but also more subtle choices regarding the in- or exclusion of the cerebral sinuses and what demarcation of sulci depth is appropriate for the study at hand.

Previous work

Myriad methods have been proposed in the last decades (Eskildsen et al., 2012; Galdames et al., 2012; Iglesias et al., 2011; Rex et al., 2004; Smith, 2002; Speier et al., 2011; Wang et al., 2014), emphasizing the important nature of the brain extraction problem that has yet to be solved satisfactorily in a generic and robust way (for a comprehensive listing of existing methods please see Eskildsen et al. (2012)). Most methods work well for deep structures of the brain. The delineation of the cortical surface, which exhibits a large degree of morphological variability across humans (Rex et al., 2004), is more challenging. Of course, an even larger variability is introduced by pathological changes, like multiple sclerosis or brain tumors.

In this study we compare the proposed method to six popular algorithms that are freely available and commonly used for comparison (Galdames et al., 2012; Iglesias et al., 2011; Wang et al., 2014). The Brain Extraction Tool[2] (BET) is a fast and versatile method that uses a deformable model (Smith, 2002). In this method a spherical mesh is initialized at the center of gravity and then expanded towards the surface of the brain. Forces based on local intensity values guide this process. Intensity and shape variations related to tumor presence might hinder the evolution of the mesh (Speier et al., 2011). Nevertheless, BET performs quite well on these kinds of images.

Another popular method is the Hybrid Watershed Algorithm[3] (HWA), combining edge detection with a deformable surface model that takes atlas-derived shape restrictions into account (Segonne et al., 2004). This method has problems with data sets containing brain tumors. The reason is two-fold. Firstly, the watershed edge detection requires an approximately constant "basin", i.e. intact white matter (WM) regions, to function properly. This is usually not true for tumor data. Secondly, the shape restrictions are derived from an atlas of healthy subjects and do not necessarily hold for more diversified cancerous data.

3dSkullStrip[4] is another software package devoted to skull stripping. It is a modified version of BET. The modifications are designed to cope with some known drawbacks of BET, e.g. leakage into the skull and incorrect inclusion of eyes and ventricles. Furthermore, not only points inside the surface, but also points lying outside it are used to guide the evolution of the mesh. Signal alterations due to tumor presence tend to interfere with the modified pipeline as well.

Brain Surface Extractor[5] (BSE) utilizes a pipeline of anisotropic diffusion filtering (Perona and Malik, 1990), Marr-Hildreth (Mexican hat) edge detection and morphological operations to accomplish the task. Anisotropic diffusion filtering is an edge-preserving filtering technique that smoothens small gradients while preserving strong ones. This operation is supposed to facilitate edge detection; and yet this parameter-sensitive method struggles if confronted with data of tumor patients (Speier et al., 2011). For normal images it has been reported that this technique works better, the higher the quality of the data. Thus, if quality requirements are met, it is able to achieve results with a high specificity (Fennema-Notestine et al., 2006).

A more recently published tool is the Robust Learning-Based Brain Extraction[6] (ROBEX) presented by Iglesias et al. (2011). After standardizing signal intensities and applying a bias field correction, a discriminative model is combined with a generative model. This is accomplished by training a random forest (Breiman, 2001) to detect voxels that are located on the brain boundary. The RF prediction is used as a cost function for fitting a shape model (point distribution model), followed by a refinement step that grants small free deformations outside the model. An extension has been proposed for scans containing brain tumors (Speier et al., 2011). To this end, an adaptive thresholding algorithm is employed at the brain boundary to detect resection cavities, complemented by a Random Walker algorithm that prevents leakage into the ventricles.

Another contemporary method proposed by Eskildsen et al. (2012) is entitled Brain Extraction Based on nonlocal Segmentation Technique (BEaST). This multi-resolution, patched-based framework uses the sum of squared differences metric to determine a suitable patch from a library of priors. For this procedure it is important to perform spatial and intensity normalization of the data (input and prior library). Further, the library needs to be representative of the given data for the segmentation to work optimally and therefore should be manually curated. By design the user is able to add custom priors to the library. When populating the library appropriately, the method shows state-of-the art performance for several data sets, amongst others for the Alzheimer's Disease Neuroimaging Initiative (Mueller et al., 2005). However, the authors state that "[…] pathologies, such as tumors and lesions, may impose a problem for the segmentation".

Most of the existing methods work well on certain datasets, but fail on others. Except ROBEX (and to some extent BEaST), all the described methods have numerical parameters. In our experience, to obtain reasonable results for real-world images, as for instance can be found in a clinical setting, a case-specific parameter tuning is frequently required. It is also not unlikely that contrast agent (not to speak of tumor images) either reduces the segmentation accuracy of the enumerated tools or requires labor-intensive parameter tweaking. Another restriction is that the methods have been designed to primarily work with T1w data. A notable exception is BET, which can additionally or solely work with T2w and related (e.g. T2*-, diffusion-weighted) images.

Our approach

Our motivation for devising a novel brain extraction algorithm was multifaceted. We aimed at establishing a method that requires minimal to no parameter tuning and still handles images from clinical routine, i.e. from a wide age range, possibly including (motion) artifacts, contrast agent and pathologically altered brain tissue. Further, we demanded that it should work with any single modality or the combination of several modalities (e.g. T1, T2, T2-FLAIR, etc.) because it seems intuitive to take all the available information into account if the task is to discriminate between brain and non-brain tissue.

Deep Neural Networks have gained more and more popularity in the recent past because they achieve state-of-the-art performance for many applications. For a recent review please see LeCun et al. (2015). Next to applications like speech and text mining they are especially effective for image segmentation tasks (Krizhevsky et al., 2012; Long et al., 2015), and thus seem to be predestined for the given classification problem.

[2] Part of FMRIB's Software Library (FSL): http://fsl.fmrib.ox.ac.uk/.
[3] Part of the FreeSurfer package: http://freesurfer.net/.
[4] Part of the AFNI package: http://afni.nimh.nih.gov/.
[5] Part of BrainSuite: http://brainsuite.org/.
[6] http://www.nitrc.org/projects/robex.

For an impression of what kind of other neuroimaging problems can be solved by deep learning, please see Plis et al. (2014). One advantage over other classifiers is that there is no need for humans to design features as they are learned from the data during training. This automatic feature tuning is one reason for their remarkable accuracy.

For our purposes we implemented a 3D Deep Convolutional Neural Network (CNN). Besides demonstrating its favorable performance on three publicly available data sets (IBSR (Rohlfing, 2012), LPBA40 (Shattuck et al., 2008) and OASIS (Marcus et al., 2007), comprising a total of N = 135 brain scans) commonly used for benchmarking of skull stripping algorithms, we conduct an experiment with our own data set of recurrent brain tumors (N = 53). Instead of only considering T1w data, we used four modalities (non-enhanced and contrast-enhanced T1, T2 and T2-FLAIR) for training and prediction. An expert generated the ground truth brain masks for this experiment.

Material and methods

Data sets

In total we used N = 188 scans for evaluation. The first data set IBSR_V2.0 came from the Internet Brain Segmentation Repository[7] (IBSR) and consisted of N = 18 nT1w scans (0.94 × 1.5 × 0.94 mm) from healthy subjects (Rohlfing, 2012). The images were provided with manual expert segmentations of gray matter and white matter, whose union served as ground truth for our experiments.

The second data set was taken from the LONI Probabilistic Brain Atlas project[8] and was entitled LPBA40 (Shattuck et al., 2008). It contained N = 40 nT1w scans (0.86 × 1.5 × 0.86 mm) of healthy subjects. Again, the union of the manually delineated tissue classifications served as a GT.

The third publicly available data set comprises the first two discs of the Open Access Series of Imaging Studies (OASIS) project[9] (Marcus et al., 2007). The reason for taking the first two discs only, corresponding to N = 77 nT1w data sets (1 × 1 × 1 mm), was to make our results comparable to Iglesias et al. (2011). In contrast to the previously described data sets, this data comprises not only healthy subjects, but includes 20 subjects that were categorized as demented and possibly suffering from Alzheimer's disease. Here, human experts did not create the brain masks. Instead, a customized automatic method was used and experts only checked the results (Marcus et al., 2007). The quality of the masks is not as good as for the other data sets but it was sufficient in order to demonstrate the general applicability and robustness of our approach.

The fourth data set was acquired at our institution and consisted of N = 53 multimodal images from patients suffering from recurrent Glioblatoma Multiforme (GBM). All subjects provided written informed consent based on institutional guidelines and the local review board. The sequences comprised nT1w, ceT1w, T2w and FLAIR images. The data acquisition was performed on a 3 Tesla MR-System (Magnetom Verio, Siemens Healthcare, Erlangen, Germany) with the following specifications: nT1w and ceT1w Magnetization Prepared Rapid Acquisition Gradient Echo (MPRAGE) sequence with TE = 3.4 ms, TR = 1740 ms and a voxel resolution of 1 × 1 × 1 mm; T2-weighted Turbo Spin Echo (TSE) imaging (T2) with TE = 85 ms, TR = 5520 ms and a voxel size of 0.63 × 0.63 mm, slice spacing 5 mm; FLAIR TSE imaging with TE = 135 ms, TR = 8500 ms and a voxel size of 0.94 × 0.94 mm, and slice spacing of 5 mm. A neuroradiologist generated the brain masks using the interactive learning and segmentation toolkit (ilastik), a semi-automatic tool that utilizes a random forest for segmentation (Sommer et al., 2011). Before binarizing, the RF pseudo probabilities were smoothed using a guided filter (He et al., 2013) with the Gaussian gradient magnitude image (sigma = 1.0 voxels) of the corresponding ceT1w scan as reference. For the guided filter the parameters were set to radius = 1.2 voxels and epsilon = 0.001. A second expert manually reviewed and approved the obtained results. For the construction we used the following definition of a brain mask. We included all cerebral and cerebellar white and gray matter as well as the brainstem (pons, medulla). We excluded CSF in ventricles (lateral, 3rd and 4th) and the quadrigeminal cistern, as well as in the major sulci and along the surface of the brain. Further, dura mater, exterior blood vessels, the sinuses and the optic chiasm were excluded. All other non-brain tissue was also discarded. For an additional experiment a senior neuroradiologist manually segmented the ventricular system (inferiorly confined by the 4th ventricle at the level of foramina of Luschka, and superiorly limited by the roof of the lateral ventricles) for half of the data (N = 26). This was realized using the wand tool of 3D Slicer (Fedorov et al., 2012). In an automated postprocessing step the segmented ventricular system was aligned with the ventricular system of the brain masks.

Convolutional neural network

Convolutional neural networks are the architecture of choice for analyzing structural data like images and 3D-volumes. In each layer the input data is convolved by a number of local filters (with a size of 5 × 5 pixels for image data) followed by a nonlinear transformation of the results. Several layers are then stacked on top of each other, where each receives the output of the previous layer as its input. This “deep” stacking gave birth to the name deep learning. Optionally, the outputs of some layers are down-sampled. With relatively low computational effort this increases the field of view of each neuron, i.e. the effectively analyzed area, and introduces translation invariance. During training the filters of the CNN are optimized to minimize a given loss function. A frequently used loss function is the Kullback–Leibler divergence. It measures the discrepancy between the CNN's prediction (i.e. the output of the very last layer of the CNN) and the labels/ground-truth of the training data.

Implementation

The deep learning architecture that was used for our experiments contains seven convolutional hidden layers and one convolutional soft-max output-layer (see Table 1 for details). The receptive field (input for each predicted pixel) of this model is 53^3 pixels, providing sufficient context for achieving a decent performance level. We evaluated different architectures while focusing on the speed of prediction and training, and only found empirically negligible effects on accuracy when adding or removing one hidden layer (less than 0.5% difference).

The transformations of the input consist of a succession of spatial 3D convolutions and a voxel−/point-wise non-linear transformation (also referred to as activation or squashing function) of the results of the convolutions. In our case we use the absolute value function, which performs, according to our experiments, on par with the rectified linear function (ReLu), often used in the literature (Krizhevsky et al., 2012; Nair and Hinton, 2010).

The only deviation from this pattern is in the first and last layer: after convolving the input to the CNN with the 16 different filters of the first

Table 1
CNN architecture details.

	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6	Layer 7	Layer 8
Filter size	4	5	5	5	5	5	5	1
Number of filters	16	24	28	34	42	50	50	2
Layer input size [voxels^3]	65	31	27	23	19	15	11	7
Layer output size [voxels^3]	31	27	23	19	15	11	7	7

[7] http://www.cma.mgh.harvard.edu/ibsr/.
[8] https://ida.loni.usc.edu/.
[9] http://www.oasis-brains.org.

layer and applying the activation function, we down-sample the result with 2^3 voxel wide windows and keep only the highest value per 2^3 cube. This procedure is known as max-pooling. It reduces the size of the intermediate representations and allows for a far wider field of view of the overall architecture. The final layer does not apply the activation function, but instead combines the two linear filter responses per voxel with the soft-max function to generate pseudo-probabilities as votes for the two given classes: brain and non-brain tissue.

During training we construct mini-batches consisting of four 65^3 voxel large cubes of the data. As the receptive field of the CNN is only 53^3 voxels, it will predict $((65-53)/2+1)^3 = 7^3$ voxels for a given 65^3 cube (thus we provide 7^3 labels at the correct positions, for a total of 1372 labels per mini-batch). The divisor (2) is a result of the max-pooling in the first layer. If we had provided cubes of size 67^3, the CNN would make predictions for 8^3 points per cube; if we had provided 53^3 cubes, it would predict only 1 voxel per cube. The reason for providing more than 53^3 voxels is that the computation (per predicted voxel) is much more efficient, the larger the input volume is.

The filters of the CNN are initialized as described by Glorot and Bengio (2010). We train the network using stochastic gradient descent, which takes on average about 10 h. Initially, the learning rate is chosen to be $1e^{-5}$. This value was determined empirically and was the highest possible value that did not cause learning to diverge. The learning rate is halved automatically when the loss on the training set does not decrease for 5000 update steps. Training ends after the tenth reduction.

Our CNN is "fully-convolutional", i.e. it can predict a variable number of voxels in one pass (depending on the input size). For predictions, we feed the CNN with volumes of size 180^3 voxels, resulting in 128^3 predicted voxels in one forward pass utilizing a technique called max-fragment pooling (Masci et al., 2013). For training we chose not to use max-fragment-pooling as it increases the computational load of a single parameter-update/optimization step by one order of magnitude and thus significantly slows down training. The difference to training with MFP is that each update step only uses every second voxel of the training labels/ground truth along the spatial axes, instead of a dense volume. This effectively smaller "batch size" introduces more noise into the stochastic gradient descent optimization of the filters, but this is in our experience not detrimental.

It is no problem to apply any fully-convolutional neural network to data with varying sizes: the number of voxels in the training or test data does not significantly influence the predictions. On the other hand, testing the CNN on data with a significantly different resolution (cm/voxel) than the data used for training will have a noticeable impact on the quality of predictions. But it is easy to accommodate to this case by employing data augmentation. During training the network could be fed the same data scaled to varying resolutions.

The network is implemented in python utilizing the theano library (Bastien et al., 2012; Bergstra et al., 2010), whereby the 3D convolution as well as the max-fragment prediction are custom, highly optimized routines that are not part of the library.

Other methods—parameter and version

The CNN was compared to BET (FSL 5.0.8), BEaST (1.15), BSE (build #:2162), ROBEX (v1.2), HWA (stable5) and 3dSkullStrip (AFNI_2011_12_21_1014). For a brief description of the corresponding methods, please see the Introduction or refer to their original publications. As suggested for most of the data sets by the authors of the respective methods (Galdames et al., 2012; Iglesias et al., 2011), we used default parameters throughout the experiments. For the evaluation of BEaST on the publicly available data sets we only used the example prior library included with the package. For the tumor experiment we additionally populated the library with custom priors derived from our data set. This version was denoted as BEaST*.

Data pre- and postprocessing

No further processing was applied to the publicly available data sets. Our institutional tumor data set was processed as follows. First, the N3 bias field correction (Sled et al., 1998) was applied to nT1w, ceT1w and T2w images, but not to the FLAIR images. Next, all four modalities were resampled to a resolution of $1 \times 1 \times 1$ mm^3. For bias field correction and resampling the FreeSurfer (Fischl, 2012) tools *nu_correct* (version 1.10) and *mri_convert* (stable5) were employed. Finally, all sequences were intra-individually registered to the respective nT1w volume using a 6-DOF linear registration as implemented in FLIRT (Jenkinson et al., 2002). A senior neuroradiology resident visually confirmed the registration accuracy.

Before feeding the data to the CNN, each volume was rescaled to be within the range [0.0, 1.0]. This rescaling reduced the number of outliers produced by the model. In order to further boost the robustness of our model, we modified the input volumes on the fly, as seen by the CNN during training, by shifting all gray-values and scaling them by a small amount. For shifting we added a random value in the range of $[-0.05, 0.05]$ to the training block; for scaling, we multiplied all voxels in a block by a random value in the range of [0.85, 1.3]. This ensured that the model learned to focus on relative variations in the data instead of learning the exact gray-value of brain vs. non-brain regions (as those might vary significantly). Usually the CNN outputs a single connected brain mask. However, as a safety measure we included a postprocessing step that identified the largest connected component and discarded all smaller (disconnected) ones. We also included an optional flag that "fills" the obtained brain masks in order to remove enclosed holes. This flag was used for the publicly available data sets but not for the tumor experiment, as here the GT deliberately contains holes that correspond to the ventricles.

Experiments

Publicly available data sets

We evaluated the performance of the CNN and the other methods on three publicly available data sets. For this purpose we performed 2-fold cross validation by randomly creating two overall folds mixing scans from all data sets and evaluated against the GT. In a second approach we trained on two of the available data sets and evaluated on the remaining one. This was performed for all combinations.

Tumor experiment

The tumor data set was evaluated independently using 2-fold cross validation. For the CNN we used the multi-modal data, whereas the nT1w scans were used to compute the brain masks with all other methods. Next to the out-of-the-box version of BEaST, we also evaluated a version denoted as BEaST*. Here we used the CNN training data to populate the custom prior library.

Measures for comparison

Using the notion of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) permitted several measures of comparison to be determined. FNs were defined as voxels that are removed by a method, yet are present in the reference mask. FPs, on the other hand, were defined as voxels which have been predicted incorrectly as brain tissue. TP voxels corresponded to correctly identified brain tissue, whereas TN voxels comprised correctly identified non-brain tissue.

Volumetric measures already utilized in several studies (Fennema-Notestine et al., 2006; Galdames et al., 2012; Iglesias et al., 2011; Rex et al., 2004; Wang et al., 2014) were used to compare a predicted brain segmentation P with the associated reference mask R. The Dice coefficient (Dice, 1945) is probably the most widespread measure used for

the comparison of two segmentations. The ratio is defined as twice the size of the intersection of the two masks normalized by the sum of their combined sizes:

$$D = \frac{2|P \cap R|}{|P| + |R|} = \frac{2TP}{2TP + FP + FN}.$$

The coefficient takes values in the range [0.0, 1.0], where 1.0 is obtained for identical segmentations. In addition Sensitivity ($TP/(TP+FN)$) and Specificity ($TN/(TN+\mathrm{FP})$) scores were computed.

Visualization of absolute error, false positives and false negatives

To visualize the spatial distribution of errors and emphasize areas with systematic problems we computed average maps of FPs, FNs and absolute errors. First, both the reference mask and the predicted mask were resampled (FreeSurfer's *mri_convert*) to a resolution of $1 \times 1 \times 1$ mm and binarized (*fslmaths*). Next, a linear 6-DOF registration using FLIRT (Jenkinson et al., 2002) was utilized to register the reference mask on the 1 mm MNI152 T1w average brain mask (Grabner et al., 2006). This linear registration was followed by a non-linear transformation using elastix (Klein et al., 2010b). The transformation matrices of both steps were stored.

After computing the FP, FN and absolute error maps between reference and predicted mask, these maps were mapped into MNI space using the stored transformation matrices. For each method and data set, the particular error maps were combined to yield an average map. For display purposes the natural logarithm of the 3D volumes collapsed (averaged) along each axis was plotted.

Statistical analysis

All statistics were computed using R version 3.1.3 (R Core Team, 2014).

Results

Publicly available data sets

We evaluated the performance of our deep neural architecture for brain extraction in comparison to several popular algorithms. Combined results (mean and standard deviation) for the IBSR, LPBA40 and OASIS data sets are presented in Table 2. Bold values indicate best result in a given category; underlined values are the second best result of the respective category. Supplementary Fig. 1 shows corresponding boxplots (median values) including outliers.

Dunnett's test was used to compare the CNN with the other methods. It revealed significant differences ($p < 0.003$) for the Dice score between our method and all other groups except Robex. For sensitivity there were significant differences in comparison to Robex ($p < 0.05$) and BEaST ($p < 0.001$), as well as to HWA ($p < 0.01$), the top scoring algorithm of this category. Significant differences of specificity were found comparing CNN to BSE ($p < 0.001$), HWA ($p < 0.001$) and BET ($p < 0.02$). The network required slightly less than 3 GB of GPU-Memory (NVIDIA Titan with Kepler™ architecture) and allowed us to predict a single volume in less than 40 s on average (IBSR 26.3 s ± 0.01 s; LPBA40 36.51 s ± 0.08 s; OASIS 40.99 s ± 0.17 s). Training took roughly 15 h. The runtimes of the compared methods are listed in Supplementary Table 1.

Table 2
Combined results for the IBSR, LPBA40 and OASIS data sets. Bold values indicate the best result; underlined values represent the second best result of the respective category.

	Dice	Sensitivity	Specificity
CNN	**95.77 (±0.01)**	94.25 (±0.03)	**99.36 (±0.003)**
BEaST	89.96 (±0.12)	87.89 (±0.12)	98.54 (±0.01)
BET	93.05 (±0.03)	95.47 (±0.05)	97.71 (±0.02)
Robex	95.30 (±0.02)	95.99 (±0.03)	98.81 (±0.01)
3dSkullStrip	92.34 (±0.04)	91.88 (±0.07)	98.45 (±0.01)
HWA	91.37 (±0.03)	**99.06 (±0.01)**	95.88 (±0.01)
BSE	78.77 (±0.10)	95.57 (±0.06)	85.12 (±0.12)

Because the data sets differ in quality and demographic properties, we also looked at the groups individually. The boxplots are depicted in Fig. 1; a detailed statistical evaluation can be found in Supplement S1. Our method obtained the highest average specificity measure for all data sets and the highest average Dice score for the IBSR and LPBA40 data sets, albeit not showing statistical significance to the second best performing method. For the OASIS data set, the second best Dice results were achieved, showing no statistical difference in comparison to the best performing tool Robex. HWA demonstrated top performance with respect to sensitivity for all data sets that could not be met by any method.

Training on two data sets and evaluating on the remaining had an impact on the median Dice score for two of the three data sets. For IBSR, i.e. training on LPBA40 and OASIS, there are no major differences observable. For OASIS and LPBA40 a decrease of the Dice score was present. However, we are able to demonstrate that a brief re-training of the neural network with one data set of the respective target domain sufficiently compensated for this loss of accuracy (Supplementary Fig. 1).

Fig. 2 shows the absolute error of the LPBA40 data for the different methods in MNI space. The illustration emphasizes the performance of the proposed method and also captures the strengths and weaknesses of the existing methods well. For details please see the Discussion section. The absolute error maps for the other two data sets, as well as false negative and false positive maps for the various data sets, can be found in the supplementary material.

Brain tumor experiment

In this experiment we evaluated the performance of the CNN brain extraction on a challenging multimodal MRI data set containing brain tumors ($N = 53$). Again Dunnett's test was used to compare the CNN with the other methods. It revealed significant differences ($p < 0.01$) for the Dice score and specificity measure ($p < 0.001$) between our method and all other methods. Robex demonstrated the highest sensitivity. In an auxiliary experiment we demonstrated that the ventricular system, which is not detected by the compared methods, only has a negligible effect on these results (Supplementary Figure 10). Results are summarized in Table 3 and Fig. 3. Note, varying the cut-off threshold used for generating the binary masks from the CNN's probability output can be used to increase the sensitivity of the method (Supplementary Figure 12). Of course, this comes at the cost of a reduced specificity. In comparison to sensitivity and specificity, varying the threshold has a weaker effect on the Dice score (Supplementary Figure 13). Prediction times were 58.54 s ± 2.75 s.

Qualitative examples of the results are depicted in Fig. 4. The brain masks of the CNN (red), the second best performing method 3dSkullStrip (w.r.t. the Dice score, yellow) and the ground truth (cyan) are outlined in a single image. It nicely can be seen how the CNN learned to detect the resection cavities.

Discussion

We present a deep 3D convolutional neural network for brain extraction in MRI images. In comparison to existing methods, this learning-based approach matches at least state-of-the-art performance. Trained appropriately it handles arbitrary multi-channel modalities. In the tumor experiment these modalities comprised non-enhanced and contrast enhanced T1w, T2w and FLAIR images, but in principle can include any conceivable modality, even non-MRI channels. For instance, even skull stripping of CT images could be performed. Further, the approach is generic enough to work with images from all kind of different species, e.g. non-primates like mice (Wang et al., 2014).

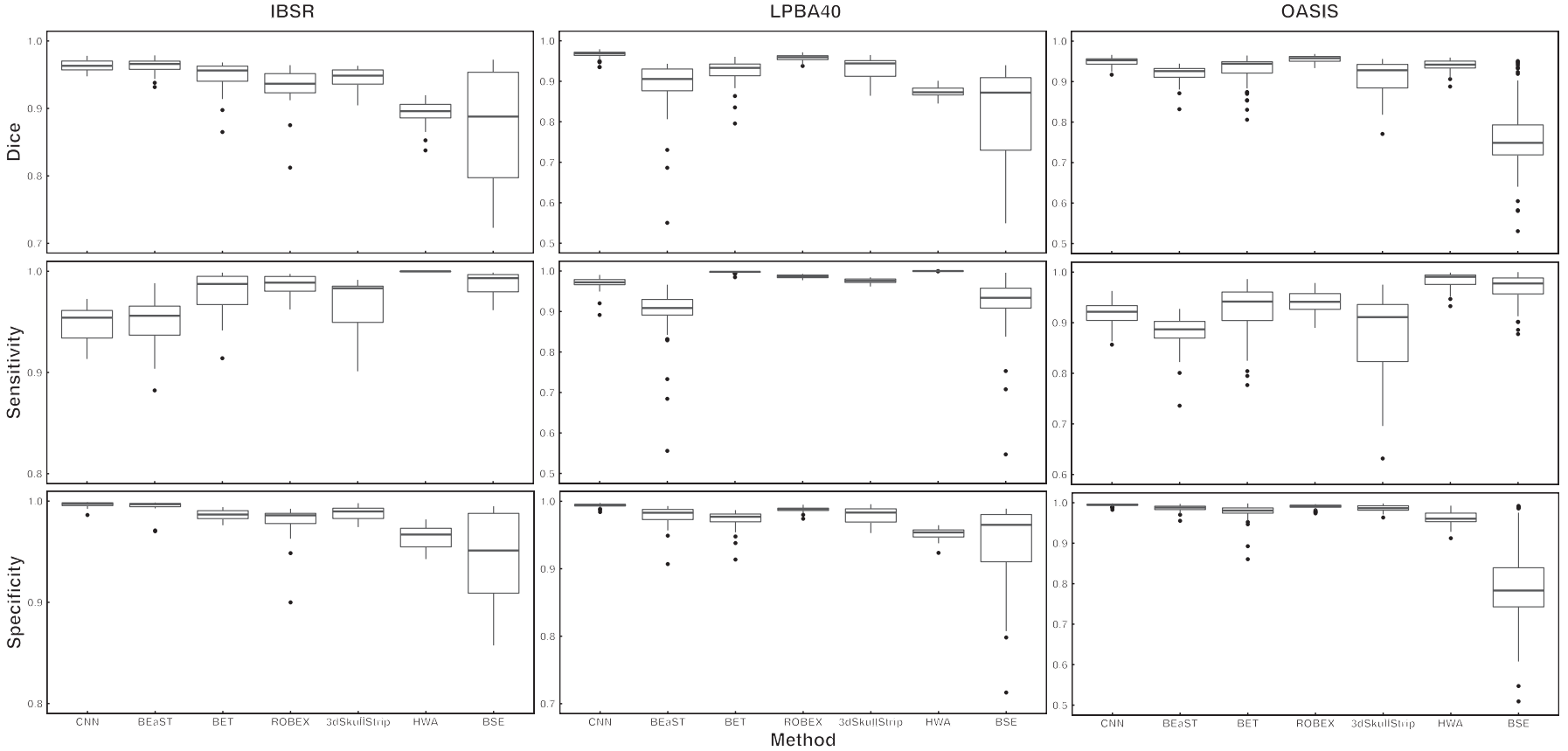

Fig. 1. Evaluation scores for three data sets. Median is displayed in boxplots; black dots represent outliers outside 1.5 times the interquartile range of the upper and lower quartile, respectively.

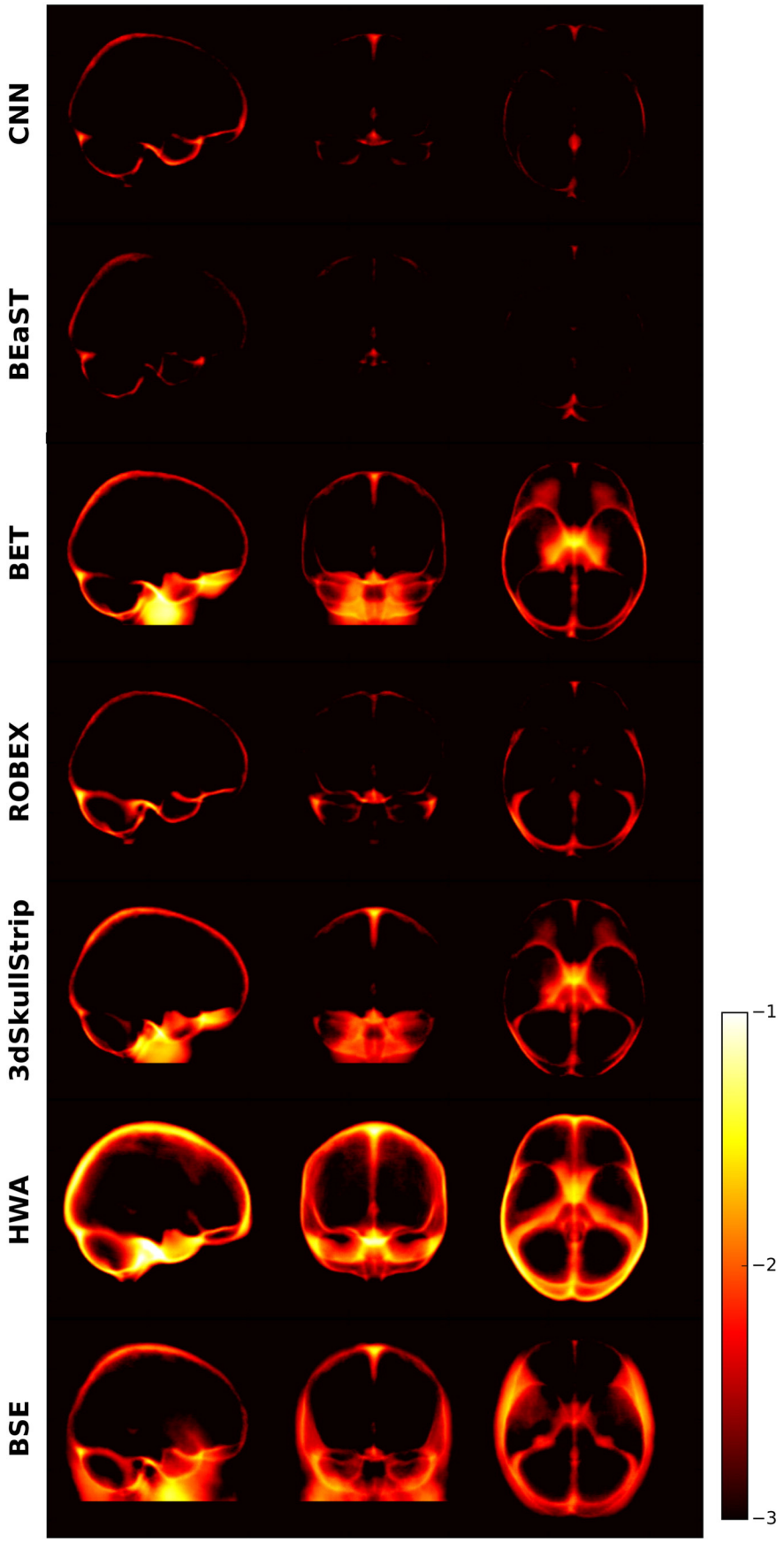

Fig. 2. Absolute error maps for the LPBA40 data. Results emphasize the excellent performance of the CNN with comparatively very low error rates in regions of the anterior and posterior cranial fossa, the paranasal sinuses, the clivus and orbits. Further, the strengths and weaknesses of the existing methods are in good agreement with findings reported in the literature. For display purposes we visualized the natural logarithm of the error.

Table 3
Mean and standard deviation for the tumor data set. CNN is compared with Dunnett's test to the other methods. Bold values indicate the best result; underlined values represent the second best result of the respective category. BEaST* denotes that data set specific priors were included in its library.

	Dice	Sensitivity	Specificity
CNN	**95.19 (±0.01)**	96.25 (±0.02)	**99.24 (±0.003)**
BEaST*	84.64 (±0.16)	89.23 (±0.19)	97.62 (±0.01)
BEaST	83.55 (±0.16)	91.31 (±0.2)	96.88 (±0.01)
BET	77.54 (±0.08)	95.26 (±0.1)	93.53 (±0.02)
Robex	86.33 (±0.03)	**99.78 (±0.002)**	95.87 (±0.01)
3dSkullStrip	88.7 (±0.03)	99.25 (±0.005)	96.81 (±0.008)
HWA	78.28 (±0.09)	97.76 (±0.13)	93.56 (±0.01)
BSE	86.46 (±0.05)	96.42 (±0.02)	96.49 (±0.02)

In comparison to other learning-based approaches, one clear advantage of neural networks is that no features have to be hand-crafted. Instead, suitable features for the given task arise during training automatically (Plis et al., 2014). Further, in contrast to voxel-wise classifiers, neural networks are based on image patches and, thus, take neighborhood information (possibly across modalities, tumor experiment) into account. The large field of view of the CNN favors continuities and the extracted brains are usually in one piece. Yet, if the ground truth used for training contains holes, e.g. the ventricles are excluded, this can also be learned by the architecture (cf. Fig. 4).

In the first experiment we demonstrated that the proposed approach achieves the highest average specificity scores for all three publicly available data sets, as well as the highest Dice score for the IBSR and

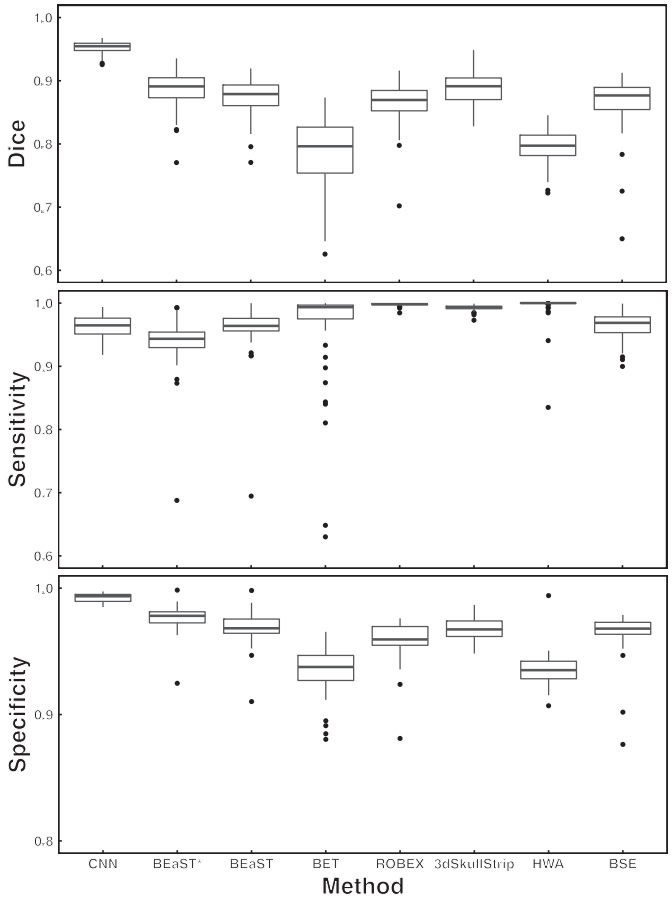

Fig. 3. Performance of the methods on the tumor data set. Results were evaluated against human GT segmentations. For the Dice and Specificity measure, the CNN significantly outperforms the existing methods. BEaST* denotes that data set specific priors were included in the library.

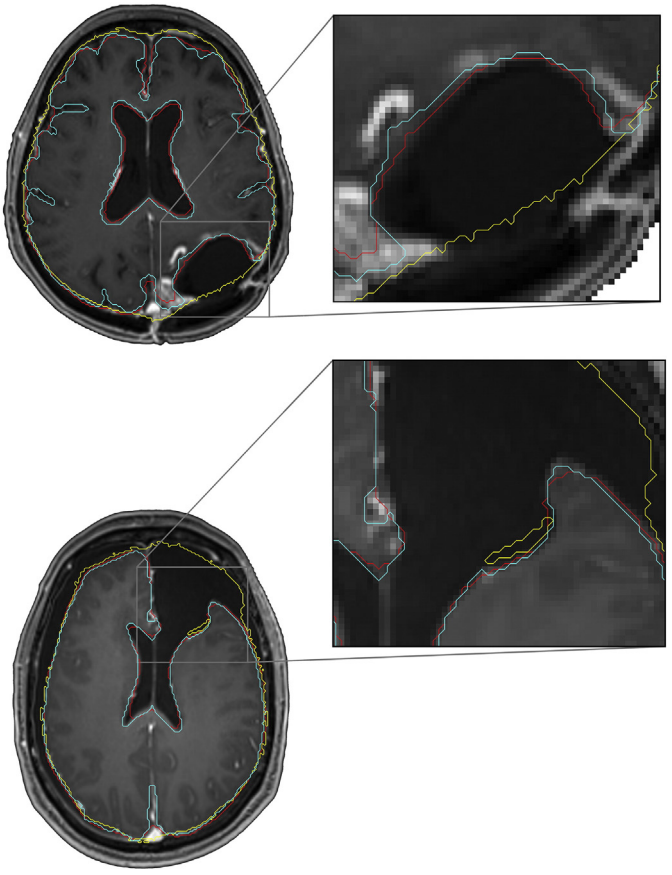

Fig. 4. Example segmentations for the tumor data set. Brain masks generated by the neural architecture compare favorably to the masks generated by all other methods w.r.t the Dice score and specificity measure (cf. Table 3). Masks generated by the CNN are outlined in red, the ones generated by the method with the second best Dice score (3dSkullStrip) in yellow and the expert constructed GT in cyan.

LPBA40 data. For the latter two human-delineated ground truth exists. For OASIS, another public data set with automatically generated and subsequently manually validated GT, we achieved second highest Dice scores that are not statistically different from the best scoring method of the respective category. However, the evaluation of the IBSR and LPBA40 data revealed an increased false negative rate for the CNN (Supplementary Figs. 4 and 6), leading to a reduced sensitivity measure (Fig. 1, Supplementary Tables 3 and 4). BEaST (Eskildsen et al., 2012), one of the methods used for comparison, displayed a similar behavior. The reasons for this most probably can be attributed to the fact that we did not populate its prior library with representative examples of the given data. The prior library that per default is included with the package is rather small. A reduced sensitivity measure may for example have an impact on cortical thickness computations and has to be addressed in future research. As the CNN does not generate a binary segmentation, but probability values instead, one possibility to cope with this shortcoming is to adjust the cut-off threshold (currently at 0.5). Of course, this comes at the cost of a reduced specificity (cf. Supplementary Figure 12).

Except Robex (and to some extent also BEaST), our method was the only parameter-free tool amongst the methods used for comparison. It has to be noted that the other methods were applied with their default parameter settings. Thus, it is very likely that a case-specific tuning would improve their results. Needless to say, this is not a very practical approach.

For the competing tools, the qualitative results are in good agreement with previously published findings (except BEaST, see above). For instance, it is known that BET often produces false positive regions ventral to the brain stem (Iglesias et al., 2011). This problem can be

seen nicely in Fig. 2 (and Supplementary Figs. 2 and 5). Further, it can be seen in these figures that 3dSkullStrip exhibits similar problematic regions to BET. This is not surprising, as it is essentially a modified version of the same tool. Another well-known result that we were able to reproduce is that HWA is the method with the highest sensitivity (Fennema-Notestine et al., 2006; Shattuck et al., 2009), at the cost of a reduced specificity with erroneous inclusion of dura and other non-brain tissue. Further, it is known that BSE used with default parameters leads to results with a very low specificity, especially on the OASIS data set (Galdames et al., 2012). These problems are revealed exemplarily in Supplementary Figs. 3, 5 and 8. It is also known that existing methods especially struggle to delineate particular anatomical structures such as the sagittal sinus and posterior fossa, displaying the highest false positive rates in these areas (Galdames et al., 2012). In comparison to existing methods, Fig. 2 shows very low error rates for our method in the regions of the anterior and posterior cranial fossa, the paranasal sinuses, the clivus and orbits. This is highly relevant in a clinical setting as these areas are important, for instance, when planning radiation therapy. Although it displays superior performance in the aforementioned regions, the CNN still exhibits isolated false positives within the cranial sinuses (Supplementary Figs. 3 and 5). One reason might reside in the GT, as even for human raters these regions are challenging. For the OASIS data, caudal parts of the cerebellum display elevated false positive rates (not only) for the CNN predictions (Supplementary Figure 8). Manual inspection revealed that the CNN results in these regions (and possibly also the results of other methods) are more likely to be correct than the supplied GT used for reference and training. This is not unexpected as the GT for this particular data is automatically generated and thus seems to suffer from a loss of quality in these well-known problematic areas.

It was previously proposed that different skull stripping methods be combined, to cope with their weaknesses and combine their strengths, using a meta-algorithm (Rex et al., 2004). For our setting it is a conceivable option to take the brain masks generated by several algorithms, combine them with e.g. STAPLE, an expectation–maximization algorithm designed to probabilistically estimate the underlying true segmentation (Warfield et al., 2004), and then use this segmentation for training the deep neural network. This would be a feasible approach if the goal were to automatically train the CNN with larger amounts of training data, usually beneficial for deep learning architectures, and avoid tedious manual segmentations that can take up to 6–8 h for a 1 mm^3 isotropic volume (Eskildsen et al., 2012). Further work will evaluate this approach.

It has been reported in the literature that scanners from different vendors have an impact on the outcome of the brain extraction (Fennema-Notestine et al., 2006; Rex et al., 2004). Therefore, a practical guide for the application of the proposed method would be to collect training data retrospectively from several studies that utilized similar scanners and protocols, and then train a CNN tailored to the conditions of the respective (home) institution or the multicenter clinical trial. When dealing with patient data, one immediate advantage is that e.g. contrast-enhanced T1w images or even exotic in-house custom sequences could be employed. Another possibility is to take a previously trained network and adapt it to the target domain. Already re-training the neural network with a single data set from the target domain has a beneficial effect on the Dice score (Supplementary Figure 11).

This scenario was mimicked by our second experiment. We took $N = 53$ multimodal MRI images, four channels comprising nT1w, ceT1w, T2w and FLAIR images, from patients suffering from brain tumors. For this challenging data set experts semi-automatically generated the GT as specified in the method section Our method significantly outperformed the other methods regarding Dice score and specificity (Fig. 3 and Table 3). Regarding the sensitivity measure, it performed about average. However, varying the threshold when generating the binary masks can be used to tune between sensitivity and specificity (Supplementary Figure 12) and to optimize the Dice score (Supplementary Figure 11 and 13). As the compared methods per default include the ventricles as part of the brain, we assessed the impact of the ventricular system on the measures for a subset of the data. For this purpose, we combined the output of each method with manual segmentations of the ventricles. For all three measures only a negligible effect is detectable (Supplementary Figure 10). We also evaluated BEaST including data set specific priors in its library (denoted BEaST*). This only led to a mild increase in performance and affirmed the concerns of the authors, that this method might have troubles with lesions like brain tumors (Eskildsen et al., 2012). Representative examples of the output of the neural network, the second best performing method 3dSkullStrip as well as the GT are demonstrated in Fig. 4. Resection cavities were nicely delineated by the method. On the other hand, the CSF within the deep sulci was not always captured as detailed as in the GT. We ascribe this to the regularization properties of the architecture. Of course, registration errors of the multimodal sequences may act as an error source as well. However, we only performed an intermodal intra-subject registration that is usually less problematic than an inter-subject registration or a registration to a template.

In future work, we plan to further improve the method. One idea is to employ an edge-preserving filter during pre-processing, e.g. the guided filter (He et al., 2013), to facilitate the learning. Next, it is also a conceivable option to extend the method to not only distinguish between brain and non-brain tissue, but to directly segment various other healthy and non-healthy tissue types like gray matter and lesions. In fact, we were able to win the 2014 brain tumor segmentations challenge (BRaTS) with a similar architecture (Urban et al., 2014). However, for this endeavor more annotated training data needs to be available and the size of the network might have to be increased.

Conclusions

In this paper we presented a 3D convolutional deep learning architecture for brain extraction of MR images. The described method yields at least state-of-the-art performance (Dice score and specificity) and addresses several problems of existing methods. Our method is not limited to non-enhanced T1w images but can also deal with contrast-enhanced scans. Secondly, when trained appropriately, the approach handles an arbitrary number of modalities. This was demonstrated on a challenging clinical data set of patients suffering from brain tumors. We believe that the proposed approach will prove to be useful for large-scale studies, as well as clinical trials.

Acknowledgments

We thank Andreas Bartsch for inspiring discussions and detailed comments on the manuscript draft as well as the anonymous reviewers. This work was supported by a postdoctoral fellowship from the Medical Faculty of the University of Heidelberg.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2016.01.024.

References

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I.J., Bergeron, A., Bouchard, N., Bengio, Y., 2012. Theano: New Features and Speed Improvements.

Bergstra, J., Breuleux, O., Bastien, F., Lamblin, P., Pascanu, R., Desjardins, G., Turian, J., Warde-Farley, D., Bengio, Y., 2010. Theano: A CPU and GPU Math Expression Compiler.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

de Boer, R., Vrooman, H.A., Ikram, M.A., Vernooij, M.W., Breteler, M.M., van der Lugt, A., Niessen, W.J., 2010. Accuracy and reproducibility study of automatic MRI brain tissue segmentation methods. NeuroImage 51, 1047–1056.

Dice, L.R., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.
Eskildsen, S.F., Coupe, P., Fonov, V., Manjon, J.V., Leung, K.K., Guizard, N., Wassef, S.N., Ostergaard, L.R., Collins, D.L., Alzheimer's disease Neuroimaging, I., 2012. BEaST: brain extraction based on nonlocal segmentation technique. NeuroImage 59, 2362–2373.
Fedorov, A., Beichel, R., Kalpathy-Cramer, J., Finet, J., Fillion-Robin, J.C., Pujol, S., Bauer, C., Jennings, D., Fennessy, F., Sonka, M., Buatti, J., Aylward, S., Miller, J.V., Pieper, S., Kikinis, R., 2012. 3D slicer as an image computing platform for the quantitative imaging network. Magn. Reson. Imaging 30, 1323–1341.
Fennema-Notestine, C., Ozyurt, I.B., Clark, C.P., Morris, S., Bischoff-Grethe, A., Bondi, M.W., Jernigan, T.L., Fischl, B., Segonne, F., Shattuck, D.W., Leahy, R.M., Rex, D.E., Toga, A.W., Zou, K.H., Brown, G.G., 2006. Quantitative evaluation of automated skull-stripping methods applied to contemporary and legacy images: effects of diagnosis, bias correction, and slice location. Hum. Brain Mapp. 27, 99–113.
Fischl, B., 2012. FreeSurfer. NeuroImage 62, 774–781.
Fischl, B., Sereno, M.I., Tootell, R.B., Dale, A.M., 1999. High-resolution intersubject averaging and a coordinate system for the cortical surface. Hum. Brain Mapp. 8, 272–284.
Galdames, F.J., Jaillet, F., Perez, C.A., 2012. An accurate skull stripping method based on simplex meshes and histogram analysis for magnetic resonance images. J. Neurosci. Methods 206, 103–119.
Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS). Society for Artificial Intelligence and Statistics.
Grabner, G., Janke, A.L., Budge, M.M., Smith, D., Pruessner, J., Collins, D.L., 2006. Symmetric atlasing and model based segmentation: an application to the hippocampus in older adults. Med. Image Comput. Comput. Assist. Interv. 9, 58–66.
He, K., Sun, J., Tang, X., 2013. Guided image filtering. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1397–1409.
Iglesias, J.E., Liu, C.Y., Thompson, P.M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. IEEE Trans. Med. Imaging 30, 1617–1634.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage 17, 825–841.
Klein, A., Ghosh, S.S., Avants, B., Yeo, B.T., Fischl, B., Ardekani, B., Gee, J.C., Mann, J.J., Parsey, R.V., 2010a. Evaluation of volume-based and surface-based brain image registration methods. NeuroImage 51, 214–220.
Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P., 2010b. Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging 29, 196–205.
Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems, Lake Tahoe, Nevada, USA.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521, 436–444.
Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. CVPR.
MacDonald, D., Kabani, N., Avis, D., Evans, A.C., 2000. Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. NeuroImage 12, 340–356.
Marcus, D.S., Wang, T.H., Parker, J., Csernansky, J.G., Morris, J.C., Buckner, R.L., 2007. Open access series of imaging studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. J. Cogn. Neurosci. 19, 1498–1507.
Masci, J., Giusti, A., Ciresan, D., Fricout, G., Schmidhuber, J., 2013. A Fast Learning Algorithm for Image Segmentation with Max-Pooling Convolutional Networks. IEEE, pp. 2713–2717.
Menze, B., Reyes, M., Van Leemput, K., 2014. The multimodal brain tumorimage segmentation benchmark (BRATS). IEEE Trans. Med. Imaging.
Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., Jack, C., Jagust, W., Trojanowski, J.Q., Toga, A.W., Beckett, L., 2005. The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin. N. Am. 15 (869–877), xi–xii.
Nair, V., Hinton, G.E., 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. pp. 807–814.
Perona, P., Malik, J., 1990. Scale-space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. 12, 629–639.
Plis, S.M., Hjelm, D.R., Salakhutdinov, R., Allen, E.A., Bockholt, H.J., Long, J.D., Johnson, H.J., Paulsen, J.S., Turner, J.A., Calhoun, V.D., 2014. Deep learning for neuroimaging: a validation study. Front. Neurosci. 8, 229.
R Core Team, 2014. R: A Language and Environment for Statistical Computing.
Rex, D.E., Shattuck, D.W., Woods, R.P., Narr, K.L., Luders, E., Rehm, K., Stoltzner, S.E., Rottenberg, D.A., Toga, A.W., 2004. A meta-algorithm for brain extraction in MRI. NeuroImage 23, 625–637.
Rohlfing, T., 2012. Image similarity and tissue overlaps as surrogates for image registration accuracy: widely used but unreliable. IEEE Trans. Med. Imaging 31, 153–163.
Segonne, F., Dale, A.M., Busa, E., Glessner, M., Salat, D., Hahn, H.K., Fischl, B., 2004. A hybrid approach to the skull stripping problem in MRI. NeuroImage 22, 1060–1075.
Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G., Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W., 2008. Construction of a 3D probabilistic atlas of human cortical structures. NeuroImage 39, 1064–1080.
Shattuck, D.W., Prasad, G., Mirza, M., Narr, K.L., Toga, A.W., 2009. Online resource for validation of brain segmentation methods. NeuroImage 45, 431–439.
Shattuck, D.W., Sandor-Leahy, S.R., Schaper, K.A., Rottenberg, D.A., Leahy, R.M., 2001. Magnetic resonance image tissue classification using a partial volume model. NeuroImage 13, 856–876.
Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17, 87–97.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.
Sommer, C., Straehle, C., Kothe, U., Hamprecht, F.A., 2011. Ilastik: interactive learning and segmentation toolkit. Biomedical Imaging: From Nano to Macro. 2011 IEEE International Symposium on, pp. 230–233.
Speier, W., Iglesias, J.E., El-Kara, L., Tu, Z., Arnold, C., 2011. Robust skull stripping of clinical glioblastoma multiforme data. Med. Image Comput. Comput. Assist. Interv. 14, 659–666.
Thompson, P.M., Mega, M.S., Woods, R.P., Zoumalan, C.I., Lindshield, C.J., Blanton, R.E., Moussai, J., Holmes, C.J., Cummings, J.L., Toga, A.W., 2001. Cortical change in Alzheimer's disease detected with a disease-specific population-based brain atlas. Cereb. Cortex 11, 1–16.
Tosun, D., Rettmann, M.E., Naiman, D.Q., Resnick, S.M., Kraut, M.A., Prince, J.L., 2006. Cortical reconstruction using implicit surface evolution: accuracy and precision analysis. NeuroImage 29, 838–852.
Urban, G., Bendszus, M., Hamprecht, F.A., Kleesiek, J., 2014. Multi-modal brain tumor segmentation using deep convolutional neural networks. Proceedings MICCAI BraTS (Brain Tumor Segmentation Challenge), Boston, Massachusetts, pp. 31–35.
Wang, L., Chen, Y., Pan, X., Hong, X., Xia, D., 2010. Level set segmentation of brain magnetic resonance images based on local Gaussian distribution fitting energy. J. Neurosci. Methods 188, 316–325.
Wang, Y., Nie, J., Yap, P.T., Li, G., Shi, F., Geng, X., Guo, L., Shen, D., Alzheimer's Disease Neuroimaging, I., 2014. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLoS One 9, e77810.
Warfield, S.K., Zou, K.H., Wells, W.M., 2004. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans. Med. Imaging 23, 903–921.
Woods, R.P., Mazziotta, J.C., Cherry, S.R., 1993. MRI-PET registration with automated algorithm. J. Comput. Assist. Tomogr. 17, 536–546.
Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation–maximization algorithm. IEEE Trans. Med. Imaging 20, 45–57.
Zhao, L., Ruotsalainen, U., Hirvonen, J., Hietala, J., Tohka, J., 2010. Automatic cerebral and cerebellar hemisphere segmentation in 3D MRI: adaptive disconnection algorithm. Med. Image Anal. 14, 360–372.