Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach

Sergi Valverde[a,*], Mariano Cabezas[a], Eloy Roura[a], Sandra González-Villà[a], Deborah Pareto[b], Joan C. Vilanova[c], Lluís Ramió-Torrentà[d], Àlex Rovira[b], Arnau Oliver[a], Xavier Lladó[a]

[a] Research institute of Computer Vision and Robotics, University of Girona, Spain
[b] Magnetic Resonance Unit, Dept of Radiology, Vall d'Hebron University Hospital, Spain
[c] Girona Magnetic Resonance Center, Spain
[d] Multiple Sclerosis and Neuroimmunology Unit, Dr. Josep Trueta University Hospital, Spain

ARTICLE INFO

ABSTRACT

In this paper, we present a novel automated method for White Matter (WM) lesion segmentation of Multiple Sclerosis (MS) patient images. Our approach is based on a cascade of two 3D patch-wise convolutional neural networks (CNN). The first network is trained to be more sensitive revealing possible candidate lesion voxels while the second network is trained to reduce the number of misclassified voxels coming from the first network. This cascaded CNN architecture tends to learn well from a small ($n \leq 35$) set of labeled data of the same MRI contrast, which can be very interesting in practice, given the difficulty to obtain manual label annotations and the large amount of available unlabeled Magnetic Resonance Imaging (MRI) data. We evaluate the accuracy of the proposed method on the public MS lesion segmentation challenge MICCAI2008 dataset, comparing it with respect to other state-of-the-art MS lesion segmentation tools. Furthermore, the proposed method is also evaluated on two private MS clinical datasets, where the performance of our method is also compared with different recent public available state-of-the-art MS lesion segmentation methods. At the time of writing this paper, our method is the best ranked approach on the MICCAI2008 challenge, outperforming the rest of 60 participant methods when using all the available input modalities (T1-w, T2-w and FLAIR), while still in the top-rank (3rd position) when using only T1-w and FLAIR modalities. On clinical MS data, our approach exhibits a significant increase in the accuracy segmenting of WM lesions when compared with the rest of evaluated methods, highly correlating ($r \geq 0.97$) also with the expected lesion volume.

Introduction

Multiple Sclerosis (MS) is the most common chronic immune-mediated disabling neurological disease affecting the central nervous system (Steinman, 1996). MS is characterized by areas of inflammation, demyelination, axonal loss, and the presence of lesions, predominantly in the white matter (WM) tissue (Compston and Coles, 2008). Nowadays, magnetic resonance imaging (MRI) is extensively used in the diagnosis and monitoring of MS (Polman et al., 2011), due to the sensitivity of structural MRI disseminating WM lesions in time and space (Rovira et al., 2015; Filippi et al., 2016). Although expert manual annotations of lesions is feasible in practice, this task is both time-consuming and prone to inter-observer variability, which has been led progressively to the development of a wide number of automated lesion segmentation techniques (Lladó et al., 2012; García-Lorenzo et al., 2013).

Among the vast literature in the field, recent techniques proposed for MS lesion segmentation include supervised learning methods such as decision random forests (Geremia et al., 2011; Jesson and Arbel, 2015), ensemble methods (Cabezas et al., 2014), non-local means (Guizard et al., 2015), k-nearest neighbors (Steenwijk et al., 2013; Fartaria et al., 2016) and combined inference from patient and healthy populations (Tomas-Fernandez and Warfield, 2015). Several unsupervised methods have been also proposed, based either in probabilistic models (Harmouche et al., 2015; Strumia et al., 2016) and thresholding methods with post-processing refinement (Schmidt et al., 2012; Roura et al., 2015a).

During the last years, a renewed interest in deep neural networks has been observed. Compared to classical machine learning approaches, deep neural networks require lower manual feature engineering, which in conjunction with the increase in the available computational power -mostly in graphical processor units (GPU)-,

and the amount of available training data, make these type of techniques very interesting (LeCun et al., 2015). In particular, convolutional neural networks (CNN) have demonstrated breaking performance in different domains such as computer vision semantic segmentation (Simonyan and Zisserman, 2014) and natural language understanding (Sutskever et al., 2014).

CNNs have also gained popularity in brain imaging, specially in tissue segmentation (Zhang et al., 2015; Moeskops et al., 2016a) and brain tumor segmentation (Kamnitsas et al., 2016; Pereira et al., 2016; Havaei et al., 2017). However, only a few number of CNN methods have been introduced so far to segment WM lesions of MS patients. Brosch et al. (2016) have proposed a cross-sectional MS lesion segmentation technique based on deep three-dimensional (3D) convolutional encoder networks with shortcut connections and two interconnected pathways. Furthermore, Havaei et al. (2016) have also introduced another lesion segmentation framework with independent image modality convolution pipelines that reduces the effect of missing modalities of new unseen examples. In both cases, authors reported a very competitive performance of their respective methods in public and private data such as the MS lesion segmentation challenge MICCAI2008 database[1], which is nowadays considered as a performance benchmark between methods.

In this paper, we present a new pipeline for automated WM lesion segmentation of MS patient images, which is based on a cascade of two convolutional neural networks. Although similar cascaded approaches have been used with other machine learning techniques in brain MRI (Moeskops et al., 2015; Wang et al., 2015), and also in the context of CNNs for coronary calcium segmentation (Wolterink et al., 2016), to the best of our knowledge this is the first work proposing a cascaded 3D CNN approach for MS lesion segmentation. Within our approach, WM lesion voxels are inferred using 3D neighboring patch features from different input modalities. The proposed architecture builds on an initial prototype that we presented at the recent Multiple Sclerosis Segmentation Challenge (MSSEG2016) (Commowick et al., 2016)[2]. That particular pipeline showed very promising results, outperforming the rest of participant methods in the overall score of the challenge. However, the method presented here has been redesigned based on further experiments to determine optimal patch size, regularization and post-processing of lesion candidates. As in previous studies (Roura et al., 2015a; Guizard et al., 2015; Strumia et al., 2016; Brosch et al., 2016), we validate our approach with both public and private MS datasets. First, we evaluate the accuracy of our proposed method with the MICCAI2008 public dataset to compare the performance with respect to state-of-the-art MS lesion segmentation tools. Secondly, we perform an additional evaluation on two private MS clinical datasets, where the performance of our method is also compared with different recent public available state-of-the-art MS lesion segmentation methods.

Methods

Input features

Training features are composed by multi-channel 3D patches sampled from a set of training images, where a new channel is created for each input image sequence available in the training database, but not restricted to. 3D feature patches take advantage of the spatial contextual information in MRI. 3D feature patches take advantage of the spatial contextual information in MRI, which may be beneficial in WM lesion segmentation due to the 3D shape of WM lesions (Rovira et al., 2015). The following steps are used to select the input features:

(i) Each training image is first subtracted by its mean and divided from its variance, in order to normalize image intensities across different images and speed-up training (LeCun et al., 1998).

(ii) For each training image, we compute 3D axial patches of size p centered on the voxel of interest, where p denotes the size in each of the dimensions.

(iii) The set of all computed patches P is stacked as $P = [n \times c \times p \times p \times p]$, where n and c denote the number of training voxels and available input modalities, respectively.

In all experiments, input patch size p is set to $p=11$. This value has been empirically selected after performing different optimization experiments with several patches sizes equal to $p = \{9, 11, 13, 15\}$.

Cascade based training

When large amounts of training data are available, several studies have shown that the addition of more CNN layers improves the accuracy in neural networks classification tasks (Simonyan and Zisserman, 2014; He et al., 2015), usually followed by an additional increase in the number of parameters in comparison to shallower networks (Krizhevsky et al., 2012). However, in MS lesion segmentation, the number of available images with labeled data may be limited by the high number of MRI slices to annotate manually (Lladó et al., 2012). More importantly, from the entire number of available voxels, only a very small number of those are lesions (∼1.5% of total brain volume for a patient with 20 ml lesion volume), which drastically reduces the number of positive training samples.

These limitations clearly affect the designed architecture, as CNNs tends to suffer from overfitting when they are not trained with enough data. In this aspect, our particular approach has been designed to deal with these two issues. By adequately sampling the training data and splitting the training procedure into two different CNN networks, we design a pipeline with fewer parameters while not compromising the precision and the accuracy of segmenting WM lesions. The next points present in detail each of the necessary steps used by our cascade based training procedure. For a more graphical explanation, see Fig. 1(a).

(i) First, the set of input patches $P = [n \times c \times p \times p \times p]$ is computed from the available training images and input modalities, as described in Section 2.1. An additional list of labels L_n is also computed using the manual expert lesion annotations, where $L_n=1$ if the current voxel n is a lesion and $L_n=0$ afterwards.

(ii) From the entire set P, patches where the intensity of the voxel of interest in the FLAIR channel is ($i_n^{FLAIR} < 0.5$) are removed. WM lesions are placed in either WM or Gray Matter (GM) boundaries, so by simply thresholding voxels on the negative class to signal intensities $i_n^{FLAIR} < 0.5$, we increase the chance of more challenging negatives examples after sampling.

(iii) In order to deal with data imbalance, we randomly under-sample the negative class in P. The training feature set F_1 is composed by all positive patches (WM lesion voxels) and the same number of negative patches (healthy voxels) after randomly sampling them.

(iv) The network (CNN_1) is trained using the balanced training feature set F_1. Exact details of the CNN architecture and training are described in Sections 2.3 and 2.4, respectively.

(v) All patches in P are afterwards evaluated using the already trained CNN_1, obtaining the probability Y^1_n of each voxel n to belong to the positive class.

(vi) Based on Y_n^1, a new balanced training feature set F_2 is created by using again all positive voxels in P and the same number of randomly selected negative examples that have been misclassified in Y^1_n, so $Y^1_n > 0.5$ with $L_n=0$.

(vii) Finally, the second CNN_2 is trained from scratch using the balanced training feature set F_2. The output of the CNN_2 is the probability Y_n^2 for each voxel of being part of a WM lesion.

(a) Cascade based pipeline
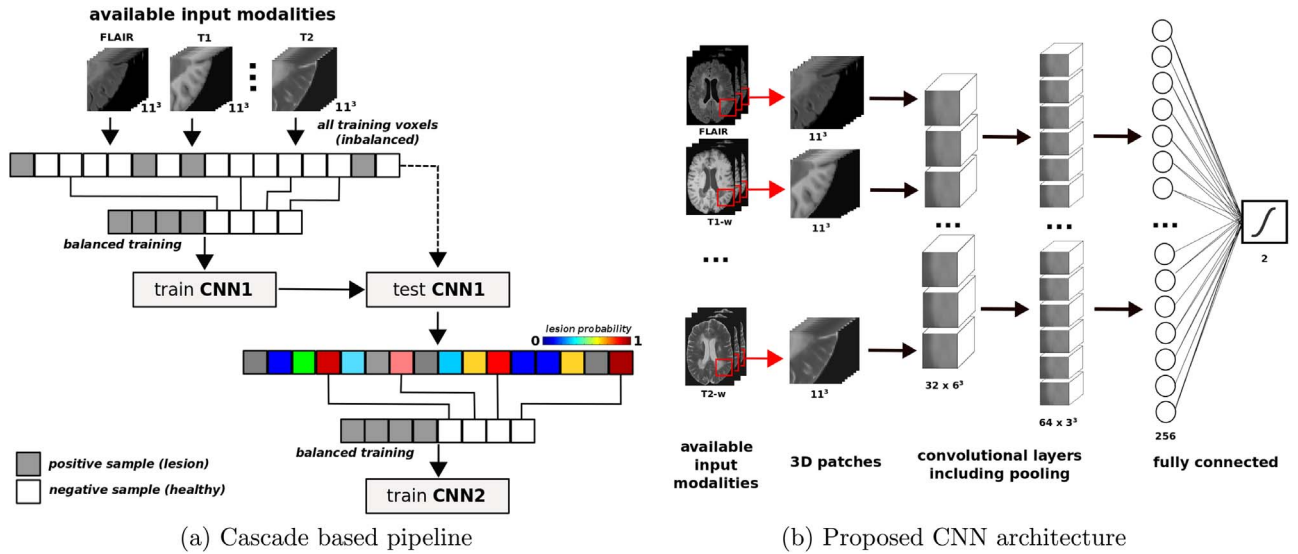

(b) Proposed CNN architecture

Fig. 1. Proposed pipeline for WM lesion segmentation. (a) Cascade based pipeline, where the output of the first CNN is used to select the input features of the second CNN. (b) The proposed 7-layer CNN model is trained using multi-sequence 3D image patches sampled from a set of training images, where each channel is created from each of the image sequences available in the training database, but not restricted to. See Table 1 for details about each layer parameters.

CNN architecture

Although increasingly deep CNN architectures have been proposed in brain MRI lately (Chen et al., 2016; Cicek et al., 2016; Moeskops et al., 2016b), still those tend to be shallower than other computer vision CNN architectures proposed for object recognition or image segmentation of natural images with up to 150 layers (He et al., 2015), mostly due to factors such as a lower number of training samples, image resolution, and a poorer contrast between classes when compared to natural images. However, compared with the latter, the lower variation in MRI tissues permits the use of smaller networks, which tends to reduce overfitting, specially in small training sets. Here, a 7-layer architecture is proposed for both CNN_1 and CNN_2 (see Fig. 1b). Each network is composed by two stacks of convolution and max-pooling layers with 32 and 64 filters, respectively. Convolutional layers are followed by a fully-connected (FC) layer of size 256 and a soft-max FC layer of size 2 that returns the probability of each voxel to belong to the positive and negative class. Exact parameters of each of the layers are shown in Table 1.

CNN training

To optimize CNN weights, training data is split into training and validation sets. The training set is used to adjust the weights of the neural network, while the validation set is used to measure how well the trained CNN is performing after the epoch, defined as a measure of the number of times all of the training samples are used once to update the architecture's weights. Each CNN is trained individually without

parameter sharing. The rectified linear activation function (ReLU) (Nair and Hinton, 2010) is applied to all layers. All convolutional layers are initialized using the method proposed by Glorot and Bengio (2010). Network parameters are learned using the adaptive learning rate method (ADADELTA) proposed by Zeiler (2012) with batch size of 128 and categorical cross-entropy as loss cost. In order to reduce over-fitting, batch-normalization regularization (Ioffe and Szegedy, 2015) is used after both convolutional layers, and Dropout (Srivastava et al., 2014) with (p=0.5) before the first fully-connected layer. Additionally, the CNN model is implemented with early stopping, which permits also to prevent over-fitting by stopping training after a number of epochs without a decrease in the validation error. Hence, final network parameters are taken from the epoch with the lowest error before stopping.

Data augmentation

Data augmentation has been shown to be an effective method to increase the accuracy of CNN networks in brain MRI (Pereira et al., 2016; Havaei et al., 2017; Kamnitsas et al., 2016). Following a similar approach, we perform data augmentation on-the-fly at batch time by multiplying the number of training samples by four following the next transformations: for each mini-batch, all patches are first rotated with 180 degrees in the axial plane. From the original and rotated versions of the patches, new versions are computed by flipping those horizontally. Other rotations than 180 degrees are avoided, in order to roughly maintain the symmetry of the brain and avoid artificial rotations of brain structures. In our empirical evaluations, rotated patches were found to increase the segmentation accuracy of the proposed method in ~1.5% when compared to non-rotated patches.

CNN testing

Once the proposed pipeline has been trained, new unseen images are processed using the same CNN cascade architecture (see Fig. 2). For each new subject to test, input feature patches for all brain voxels are obtained using the approach proposed in Section 2.1. All image patches are then evaluated using the first trained CNN. The first network discards all voxels with low probability to be lesion. The rest of the voxels are re-evaluated using the second CNN obtaining the final probabilistic lesion mask.

Binary output masks are computed by linear thresholding t_{bin} of

Table 1
Proposed 7-layer CNN architecture for input image patch size of $11 \times 11 \times 11$ with c input modalities as channels. Layer description: 3D convolutional layer (CONV), 3D max-pooling layer (MP) and fully-convolutional layer (FC). Same architecture is proposed for both CNNs.

Layer	Type	Input size	Maps	Size	Stride	Pad
0	*input*	$c \times 11 \times 11 \times 11$				
1	CONV	$c \times 11 \times 11 \times 11$	32	3^3	1^3	1^3
2	MP	$32 \times 5 \times 5 \times 5$	–	2^3	2^3	0
3	CONV	$64 \times 5 \times 5 \times 5$	64	3^3	1^3	1^3
4	MP	$64 \times 2 \times 2 \times 2$	–	2^3	2^3	0
5	FC	256	256	1	–	–
6	Softmax	2	2	1	–	–

testing image (all voxels)

FLAIR   T2

11^3 11^3

trained CNN1

**select only voxels
lesion prob > 0.5**

FLAIR   T2

11^3 11^3

trained CNN2

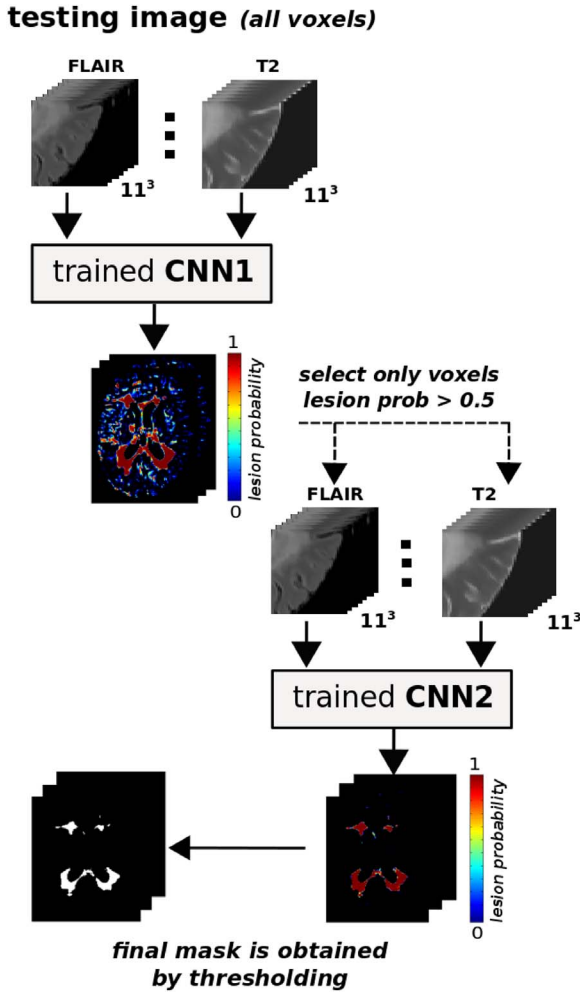**final mask is obtained
by thresholding**

Fig. 2. Proposed CNN testing procedure. New unseen subjects are evaluated using the same cascade architecture of trained networks. Voxels with ≥0.5 probability of being lesion in CNN_1 are re-evaluated using the second CNN_2. Final segmentation masks are obtained by thresholding the probabilistic mask obtained from the second CNN_2.

probabilistic lesion masks. Afterwards, an additional false-positive reduction is performed by discarding binary regions with lesion size below l_{min} parameter. Both parameters t_{bin} and l_{min} are automatically optimized by averaging the best values for each image used for training. Note that using this process, t_{bin} and l_{min} are only computed once after training the network and can be afterwards applied to an arbitrary number of unseen testing images.

Implementation

The proposed method has been implemented in the Python language[3], using Lasagne[4] and Theano[5] (Bergstra et al., 2011) libraries. All experiments have been run on a GNU/Linux machine box running Ubuntu 14.04, with 32 GB RAM memory. CNN training has been carried out on a single Tesla K-40c GPU (NVIDIA corp, United States) with 12 GB RAM memory. The proposed method is currently available for downloading at our research website[6].

___
[3] https://www.python.org/.
[4] http://lasagne.readthedocs.io.
[5] http://deeplearning.net/software/theano/.
[6] https://github.com/NIC-VICOROB/cnn-ms-lesion-segmentation.

Experimental results

MICCAI 2008 MS lesion segmentation

Data

The MICCAI 2008 MS lesion segmentation challenge is composed by 45 scans from research subjects acquired at Children's Hospital Boston (CHB, 3T Siemens) and University of North Carolina (UNC, 3T Siemens Alegra) (Styner et al., 2008). For each subject, T1-w, T2-w and FLAIR image modalities are provided with isotropic resolution of $0.5 \times 0.5 \times 0.5$ mm^3 in all images. Data is distributed in two sets:

- 20 training cases (10 CHB and 10 UNC) are provided with manual expert annotations of WM lesions from a CHB and UNC expert rater.
- 25 testing cases (15 CHB and 10 UNC) provided without expert lesion segmentation.

Evaluation

The evaluation is done blind for the teams by submitting the segmentation masks of the 25 testing cases to the challenge website[7]. Submitted segmentation masks are compared with manual annotations of both UNC and CHB raters. The evaluation metrics are based on the following scores:

- The % error in lesion volume in terms of the absolute difference in lesion volume (*VD*) between manual annotations masks and output segmentation masks:

$$VD = \frac{|TP_s - TP_{gt}|}{TP_{gt}} \times 100 \tag{1}$$

where TP_s and TP_{gt} denote the number of segmented voxels in the output and manual annotations masks, respectively.

- Sensitivity of the method in terms of the True Positive Rate (*TPR*) between manual lesion annotations and output segmentation masks, expressed in %:

$$TPR = \frac{TP_d}{TP_d + FN_d} \times 100 \tag{2}$$

where TP_d and FN_d denote the number of correctly and missed lesion region candidates, respectively.

- False discovery rate of the method in terms of the False Positive Rate (*FPR*) between manual lesion annotations and output segmentation masks, also expressed in %:

$$FPR = \frac{FP_d}{FP_d + TP_d} \times 100 \tag{3}$$

where FP_d denotes number of lesion region candidates incorrectly classified as lesion.

From these evaluation metrics, a single score is computed to rank each of the participant strategies, being 90 points comparable to human expert performance (Styner et al., 2008).

Experiment details

Provided FLAIR and T2-w image modalities were already rigidly co-registered to the T1-w space. All images were first skull-stripped using BET (Smith et al., 2002) and intensity normalized using N3 (Sled et al., 1998) with smoothing distance parameter to 30-50 mm (Boyes et al., 2008; Zheng et al., 2009). All training and testing images were then resampled from $512 \times 512 \times 512$ ($0.5 \times 0.5 \times 0.5$ mm) to $256 \times 256 \times 256$ ($1 \times 1 \times 1$ mm) to reduce the computational training time. Although original data could be just down-sampled, we observed

___
[7] http://www.ia.unc.edu/MSseg/about.html.

a poor alignment between images and annotation masks. Instead, a two-step approach was followed in order to maintain the consistency between image modalities and manual annotation masks: each of the modalities was first down-sampled to $(256 \times 256 \times 256)$ by local averaging. Then, each image modality in the original space was registered against the same down-sampled image, using the Statistical Parametric Mapping (SPM) toolkit (Estimate and reslice), with normalized mutual information as objective function. The resulting transformation matrix of the FLAIR image was then used to resample also manual annotations to $(256 \times 256 \times 256)$.

For this particular experiment, we trained two different pipelines:

(i) one trained using all input image modalities available (T1-w, FLAIR and T2-w).
(ii) one trained using only T1-w and FLAIR images.

Both pipelines were trained using the 20 available images provided with manual expert annotations, resulting on a balanced training database of 800.000 patches. 25% of the entire training dataset was set to validation and the rest to train the network. As reported in Styner et al. (2008), UNC manual annotations were adapted to match closely those from CHB, so in our experiments only CHB annotations were used to train the CNNs. The number of maximum training epochs was set 400 with early-stopping of 50 for each network of the cascade. Test parameters t_{bin} and l_{min} were optimized during training to $t_{bin} = 0.8$, $l_{min} = 5$ and $t_{bin} = 0.7$, $l_{min} = 5$ in the first and second pipeline, respectively.

Results

Table 2 shows the mean *VD*, *TPR*, *FPR* and final overall scores of the two proposed pipelines. Our results are compared with other 10 out of 60 other online participant strategies. At the time of writing this paper, our proposed pipeline using all available modalities (T1-w, FLAIR and T2-w) has been ranked in the first position of the challenge, outperforming the rest of participant strategies. Moreover, when only using T1-w and FLAIR images, our pipeline was still very competitive and it was ranked in the top three, only outperformed by the approach proposed by Jesson and Arbel (2015).

In terms of *VD*, the proposed method returned the lowest absolute difference in lesion volume of all participants for either UNC manual annotations (using T1-w and FLAIR) or CHB annotations (using T1-w, FLAIR and T2-w). Additionally, our CNN approach depicted a very high sensitivity detecting WM lesions (*TPR*), being only outperformed by Jerman et al. (2016). As seen in Table 2, other pipelines with high sensitivity such as the same work proposed by Jerman et al. (2016) or the CNN presented in Brosch et al. (2016) tended also to increase the number of false positives. Compared to these methods, our proposed pipeline showed a high *TPR* while maintaining lower false positive bounds.

Clinical MS dataset

Data

The MS cohort is composed by 60 patients with clinically isolated syndrome from the same hospital center (Hospital Vall d'Hebron, Barcelona, Spain). In all patients, the initial clinical presentation was clearly suggestive of MS:

Patients were scanned on a 3T Siemens with a 12-channel phased-array head coil (Trio Tim, Siemens, Germany), with acquired input modalities: 1) transverse PD and T2-w fast spin-echo (TR=2500 ms, TE=16-91 ms, voxel size=0.78×0.78×3 mm^3); 2) transverse fast T2-FLAIR (TR=9000 ms, TE=93 ms, TI=2500 ms, flip angle=120°, voxel size=0.49×0.49×3 mm^3); and 3) sagittal 3D T1 magnetization prepared rapid gradient-echo (MPRAGE) (TR=2300 ms, TE=2 ms, TI=900 ms, flip angle=9 °; voxel size=1×1×1.2 mm^3). In 25 out of the 60 subjects, the PD image modality was not available, so data was separated in two datasets MS1 and MS2 as follows:

- MS1 dataset: 35 subjects containing T1-w, FLAIR and PD modalities. Within this dataset, WM lesion masks were semi-automatically delineated from PD using JIM software (Xinapse Systems, http://www.xinapse.com/home.php) by an expert radiologist of the same hospital center. Mean lesion volume was 2.8 ± 2.5 ml (range 0.1-10.7 ml).
- MS2 dataset: 25 patients containing only T1-w and FLAIR input modalities. WM lesion masks were also semi-automatically delineated from FLAIR using JIM software by the same expert radiologist. Mean lesion volume was 4.1 ± 4.7 ml (range 0.2-18.3 ml).

Evaluation

Evaluation scores proposed in Section 3.1 are complemented with the following metrics:

- The overall % segmentation accuracy in terms of the Dice Similarity Coefficient (*DSC*) between manual lesion annotations and output segmentation masks:

$$DSC = \frac{2 \times TP_s}{FN_s + FP_s + 2 \times TP_s} \times 100 \tag{4}$$

where TP_s and FP_s denote the number of voxels correctly and incorrectly classified as lesion, respectively, and FN denotes the number of voxels incorrectly classified as non-lesion.
- Precision rate of the method expressed in terms of the Positive Predictive Value rate (*PPV*) between manual lesion annotations and output segmentation masks, also expressed in %:

$$PPV = \frac{TP_s}{TP_s + FP_s} \times 100 \tag{5}$$

Table 2

Segmentation results on the MICCAI 2008 MS lesion segmentation test set. Results are shown for 12 out of 60 participant methods. Mean *VD*, *TPR* and *FPR* and final scores are shown split by CHB and UNC raters, as in the original submission website. An overall score of 90 is considered comparable to human performance. The best value for each score is depicted in bold. For a complete ranking of all the participant methods, please refer to the challenge website http://www.ia.unc.edu/MSseg/about.html.

Rank	Method	UNC rater			CHB rater			Score
		VD	TPR	FPR	VD	TPR	FPR	
–	*Human rater*	–	–	–	–	–	–	90.0
1	Proposed method (T1-w, FLAIR, T2-w)	62.5	55.5	46.8	**40.8**	**68.7**	46.0	**87.12**
2,6	Jesson and Arbel (2015)	46.9	43.9	**32.3**	113.4	53.5	**24.2**	86.93
3	Proposed method (T1-w, FLAIR)	34.4	50.6	44.1	54.2	60.2	41.8	86.70
4,5	Guizard et al. (2015)	46.3	47.0	43.5	51.3	52.7	42.0	86.11
7	Tomas-Fernandez and Warfield (2015)	**37.8**	42.0	44.1	53.4	51.8	45.1	84.46
8,11	Jerman et al. (2016)	58.1	**59.0**	64.7	96.8	71.3	62.8	84.16
10	Brosch et al. (2016)	63.5	47.1	52.7	52.0	56.0	49.8	84.07
12	Strumia et al. (2016)	56.9	37.7	34.6	113.7	42.9	30.5	83.92
14	Roura et al. (2015a)	65.2	44.9	43.2	158.9	55.4	40.5	82.34

where TP_s and FP_s denote the number of correctly and incorrectly lesion region candidates, respectively.

Experiment details

For each dataset, T1-w and FLAIR images were first skull-stripped using BET (Smith et al., 2002) and intensity normalized using N3 (Sled et al., 1998) with smoothing distance parameter to 30-50 mm. Then, T1-w and FLAIR images were co-registered to the T1-w space using also the SPM toolbox, with normalized mutual information as objective function and tri-linear interpolation with no warping.

In order to investigate the effect of the training procedure on the accuracy of the method, two different pipelines were considered:

(i) one trained with leave-one-out cross-validation: for each patient, an independent pipeline was trained from the rest of the patient images of the MS1 and MS2 datasets. Testing parameters T_{bin}, l_{min} were computed at global level from all input images used for training. Final values for test parameters were set to $t_{bin} = 0.8$, $l_{min} = 20$.

(ii) one trained with independent cross training-testing databases: all images in MS1 were used to train a single pipeline, which was used to evaluate all the images in MS2. Afterwards, the process was inverted and MS2 images were used to train a single pipeline, which was used to evaluate MS1 images. Optimized test parameters t_{bin} and l_{min} were set equal to the previous pipeline.

In both experiments, each of the cascaded networks was trained using FLAIR and T1-w image modalities for a maximum number of 400 epochs with early-stopping of 50. For each subject, a training set of approximately 200.000 patches was generated, where 25% was set to validation and the rest to train the network. The same modalities were used for testing.

Comparison with other available methods

The accuracy of the proposed method is compared with two available state-of-the-art MS lesion segmentation approaches such as LST (Schmidt et al., 2012) and SLS (Roura et al., 2015a). Parameters for LST and SLS were optimized by grid-search on the MS1 database. As in Roura et al. (2015a), given that both MS1 and MS2 images were acquired using the same scanner, same parameters were used for MS2. In LST, initial threshold was set to $\kappa = 0.15$ and lesion belief map was set to $lgm = gm$. In the case of SLS, smoothing distribution parameter was set to $\alpha = 3$, the percentage of GM/WM tissue to $\lambda_{ts} = 0.6$, and percentage of neighboring WM to $\lambda_{nb} = 0.6$ in the first iteration. In the second iteration, parameters were set to $\alpha = 1.7$, $\lambda_{ts} = 0.75$ and $\lambda_{nb} = 0.7$, respectively.

Results

Table 3 shows the mean *DSC*, *VD*, *TPR*, *FPR* and *PPV* scores for all the evaluated pipelines. As seen in the Table 3, our proposed approach clearly outperformed the rest of available MS pipelines by a large margin in terms of detection (*TPR*, *FPR*), precision (*PPV*) and segmentation (*VD*, *DSC*). Furthermore, our method yielded a similar performance in both datasets and independently of the training procedure employed, highlighting the capability of the network architecture to detect new lesions on unseen images.

Fig. 3 shows the response operating characteristic (ROC) curves and parameter optimization plots for the proposed CNN approach on the MS1 database. Compared to the same CNN architecture without cascading, the proposed approach yielded a higher sensitivity, lower false negative rates and significantly higher DSC overlaps with respect to manual annotations, due to the addition of the second network, which drastically reduced the number of misclassified voxels.

Fig. 4 depicts a qualitative evaluation of each of the evaluated methods for a single subject of the MS1 dataset. Compared to the SLS and LST pipelines, both proposed CNN pipelines present a significant increase in the number of detected lesions (depicted in green). SLS and LST methods are designed to reduce the *FPR*, so as a counterpart they exhibit a higher number of missed lesions (depicted as blue squares). In contrast, the high sensitivity of the proposed method detecting WM lesions was not compromised by a remarkably increase in the false positives (depicted in red), still lower than the rest of pipelines (see Table 3).

We also compared the correlation between expected and estimated lesion volume for each of the evaluated pipelines and datasets. Table 4 shows the linear correlation between expected and obtained lesion volumes for each of the evaluated pipelines on the MS1 and MS2 databases. Both the Spearman's and Pearson's coefficients are shown, as the distribution of lesion volumes is highly concentrated in low lesion volumes and samples with high lesion load may influence the Pearson's coefficient, specially given the limited number of samples. Fig. 5 shows the linear regression models fitted based on the volume estimations of each method (solid lines). For each regression model, 95% confidence interval areas and expected lesion volume (dash lines) were also shown for comparison. In general, distances between expected and computed lesion volumes were lower in CNN architectures when compared to the rest of pipelines, as shown by the Pearson linear correlation coefficients obtained ($r \geq 0.97$ in all cases). In addition, confidence intervals for our proposed methods were distinctively lower in both datasets, specially in images with higher lesion load, where more variability was introduced.

Discussion

In this paper, we have presented a novel automated WM lesion segmentation method in brain MRI with application to MS patient images. The proposed patch-wise pipeline relies on a cascade of two identical convolutional neural networks, where the first network is trained to be more sensitive revealing possible candidate lesion voxels while the second network is trained to reduce the number of misclassified voxels coming from the first network. Although CNN cascade-based features have been used in coronary calcium segmentation (Wolterink et al., 2016), still this approach had been not applied in the context of MS lesion segmentation. In our opinion, the proposed cascaded architecture of two CNN is an interesting contribution of the present study. Our experiments have shown that the proposed method outperforms the rest of participant methods in the MICCAI2008 challenge, which is considered nowadays a benchmark for new proposed strategies. Moreover, additional experiments with two clinical MS datasets also confirms our claims about the accuracy of our proposed method. Compared to other state-of-the-art available methods, our approach exhibits a significant increase in the detection and segmentation of WM lesions, highly correlating ($r \geq 0.97$) with the expected lesion volume on clinical MS datasets. Furthermore, the performance of the pipeline is consistent when trained with different image datasets, acquisition protocols and image modalities, which highlights its robustness in different training conditions.

Automated WM lesion segmentation pipelines are usually designed as a trade-off between the sensitivity detecting lesions and the capability to avoid false positives, which may affect expected lesion segmentation. In all the experiments handled in this paper, the proposed approach yielded a high sensitivity detecting WM lesions while maintaining the number of false positives reasonably low. Related to that, differences in *VD* with respect to manual lesion annotations were in average the lowest when using our cascade architecture, specially in images with high lesion size, as seen in the correlation plots between expected and estimated lesion volume of MS patients. In our opinion, the obtained results show the capability of the cascade architecture to reduce false positives without compromising lesion segmentation, which is relevant, as it can have a direct benefit in several tasks such as automated delineation of focal lesions for MS diagnosing (Polman et al., 2011), measuring the impact of a given

Table 3

Mean segmentation results for each of the evaluated methods on the two MS clinical datasets. Results are shown for SLS, LST and our proposed approach trained with either leave-one-out experiments (LOU) or different training-testing dataset. Mean *DSC*, *VD*, *TPF*, *FPF* and *PPV* are shown for each method and database. The best value for each metric is shown in bold.

(a) MS1 dataset (n=35)

Method	DSC	VD	TPR	FPR	PPV
SLS Roura et al. (2015a)	33.4	81.0	49.5	38.2	61.7
LST Schmidt et al. (2012)	32.2	49.7	44.4	41.2	58.8
Proposed method (LOU)	**53.5**	**30.8**	77.0	**30.5**	**70.3**
Proposed method (train MS2)	50.8	38.8	**79.1**	35.6	65.3

(b) MS2 dataset (n=25)

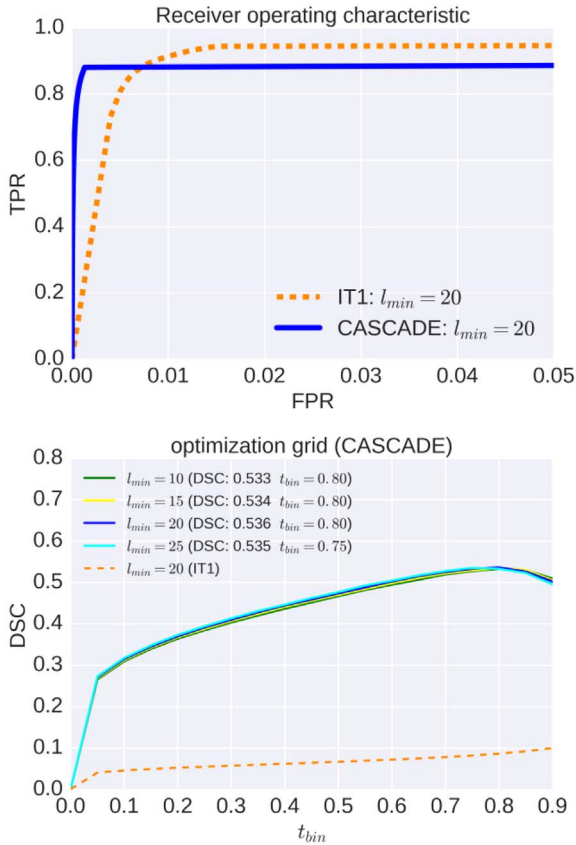| Method	DSC	VD	TPR	FPR	PPV
SLS Roura et al. (2015a)	29.7	65.1	35.7	46.7	53.2
LST Schmidt et al. (2012)	27.3	59.4	58.9	40.2	59.7
Proposed method (LOU)	**56.0**	27.5	**68.2**	33.6	66.1
Proposed method (train MS1)	51.9	**26.1**	63.7	**27.2**	**73.0**

Fig. 3. Receiver Operation Characteristic (ROC) curves and parameter optimization plots on the MS1 database. First row: ROC curves with fixed best minimum lesion size ($l_{min} = 20$) for both the proposed cascaded approach (solid blue) and the same architecture using only the network without cascading (dotted orange). Second row: parameter optimization for the cascaded CNN (t_{bin} and l_{min}) against DSC coefficient. Additionally, the best configuration for the CNN approach without cascading is also shown for comparison (dotted orange).

treatment on WM lesions (Calabrese et al., 2012) or reducing the effects of WM lesions in brain volumetry by refilling WM lesions before tissue segmentation (Valverde et al., 2014, 2015).

In terms of the network architecture, our pipeline has been designed to handle with the lack of large amounts of training data and most importantly, with the imbalanced nature of MRI images of MS patients, where lesion voxels are highly underrepresented. The proposed cascade of two identical CNNs can be also thought as a two-step optimization network, where a more general specific network learns to find candidate lesions while a second more specific network learns to reduce the number of false positives. At testing time, this means that voxels with a low probability to belong to WM lesion are easily discarded by the first network, while challenging voxels flow from the first to the second network, which has been explicitly trained to discriminate between lesion voxels and these challenging non-lesion voxels. Each network is trained independently in different portions of the training data, so the reduced size of each of these networks makes them less prone to overfitting. This makes this particular architecture suitable for small datasets without massive amounts of training samples, as there is a relative low number of parameters to optimize at each of the two steps.

In our proposed design, the sampling procedure followed to deal with data imbalance is equally important. In the first network, the negative class has been under-sampled to the same number of existing lesion voxels, increasing the sensitivity of the network detecting WM lesions at testing time. However, by under-sampling the negative voxel distribution, the probability to misclassify healthy voxels as lesion also increases due to a poorer representation of the negative class. In order to reduce the number of misclassified voxels, the network may be redesigned to incorporate more layers with the aim to learn more abstract features. However, a deeper network increases the number of parameters, and more training data is needed in order to avoid over-fitting. In contrast, within the designed proposal, false positives are reduced by re-training an identical CNN architecture with the most challenging samples derived from the first network.

In contrast to other CNN approaches used in brain MRI (Moeskops et al., 2016a; Brosch et al., 2016), the training stage is here split into two independent 7-layer CNNs that are exposed to different portions of equal data, so the number of parameters to be optimized at each step is remarkably small (<190000 with input patches of size 11^3) when compared to other methods. This can be specially interesting when training is performed with a cohort of MS patients with low lesion load, as seen in our experiment containing only around 200.000 training samples after balancing equally lesion and non-lesion voxels. In this aspect, our results suggest that the proposed pipeline is a valid alternative to train CNN networks without the need of large amounts of training data. In our opinion, this is one of the major contributions of the present study, given the difficulty to obtain labeled MRI data, in comparison to the huge number of available unlabeled data. This suggests that a similar approach may be useful in other similar detection problems like automated segmentation of WM hyperintensities (Griffanti et al., 2016), longitudinal MS lesion segmentation
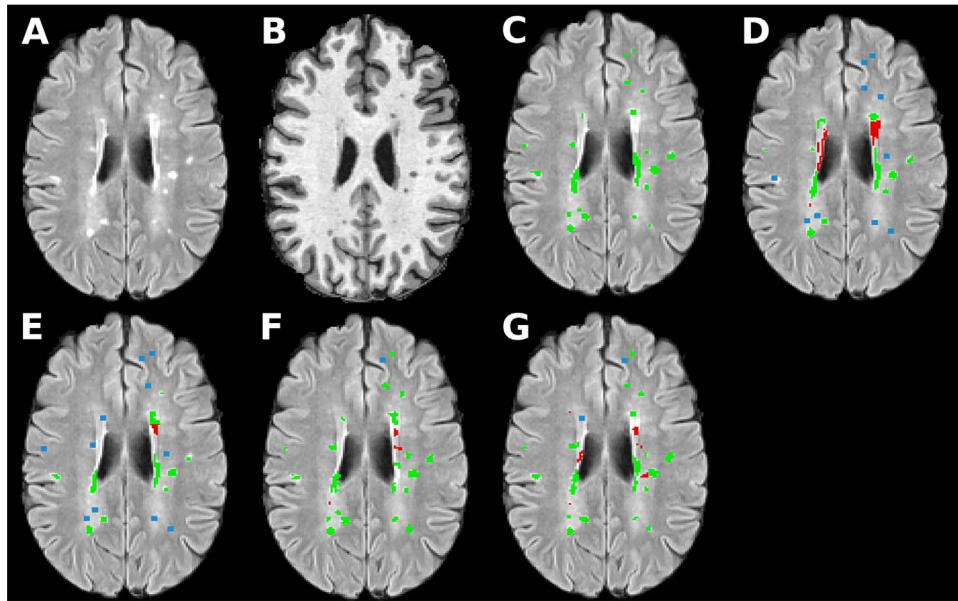
Fig. 4. Output segmentation masks for each of the evaluated methods on the first subject of the MS1 clinical dataset. (A) FLAIR image. (B) T1-w image. (C) Manual WM lesion annotation. Output segmentation masks for SLS (D), LST (E) and our approach when trained using either leave-one-out (F) or the MS2 dataset (E). On all images, true positives are denoted in green, false positives in red and false negatives with a blue square.

Table 4
Spearman's and Pearson's correlation between expected and obtained lesion volume for each of the evaluated pipelines on the MS1 (a) and MS2 (b) databases. Numbers represent the respective coefficient *r* and (p-value) for each method.

(a) MS1 dataset (n=35)

pipeline	Spearman's correlation	Pearson's correlation
SLS	0.65 (<0.001)	0.65 (<0.001)
LST	0.87 (<0.001)	0.82 (<0.001)
CNN (trained LOU)	0.96 (<0.001)	0.97 (<0.001)
CNN (trained on MS2)	0.95 (<0.001)	0.97 (<0.001)

(b) MS1 dataset (n=25)

pipeline	Spearman's correlation	Pearson's correlation
SLS	0.92 (<0.001)	0.95 (<0.001)
LST	0.93 (<0.001)	0.89 (<0.001)
CNN (trained LOU)	0.98 (<0.001)	0.97 (<0.001)
CNN (trained on MS1)	0.97 (<0.001)	0.99 (<0.001)

(Cabezas et al., 2016), Lupus lesion detection (Roura et al., 2015b), traumatic brain injury (Lee and Newberg, 2005) and brain tumor segmentation (Menze et al., 2015).

The proposed method presents also some limitations. When compared to SLS and LST, although our CNN approach clearly yields a better accuracy on the MS dataset, this has to be trained and tested for each of the datasets evaluated, which is time consuming and may require more expertise. Related to that, the abstract representations learned by the classifier are most probably provided by the FLAIR and T2-w image sequences, as these sequences are more sensitive revealing the majority of MS lesions when compared with T1-w, for instance. However, the geometry and image contrast of FLAIR / T2-w tend to vary considerably within acquisition protocols. Although changes in intensity scale can be corrected by the internal regularization of the CNN, differences between image domains still may exist, being the feature representations learned by the classifier highly dependent of the dataset used, which suggests us to use it only on the same image domain.

Conclusions

Automated WM lesion segmentation is still an open field, as shown by the constant number of proposed methods during the last years. In this paper, we have presented a novel automated WM lesion segmentation method with application to MS patient images that relies on a cascade of two convolutional neural networks. Experimental results presented in this paper have shown the consistent performance of the proposed method on both public and private MS data, outperforming the rest of participant methods in the MICCAI2008 challenge, which is considered nowadays as a benchmark for new proposed strategies. Compared to other available methods, the performance of our proposed approach shows a significant increase in the sensitivity while maintaining a reasonable low number of false positives. As a result, the method exhibits a lower deviation in the expected lesion volume in manual lesion annotations with different input image modalities and image datasets. In addition, the proposed cascaded CNN architecture tends to learn well from small sets of data, which can be very interesting in practice, given the difficulty to obtain manual label annotations and the amount number of available unlabeled MRI data.
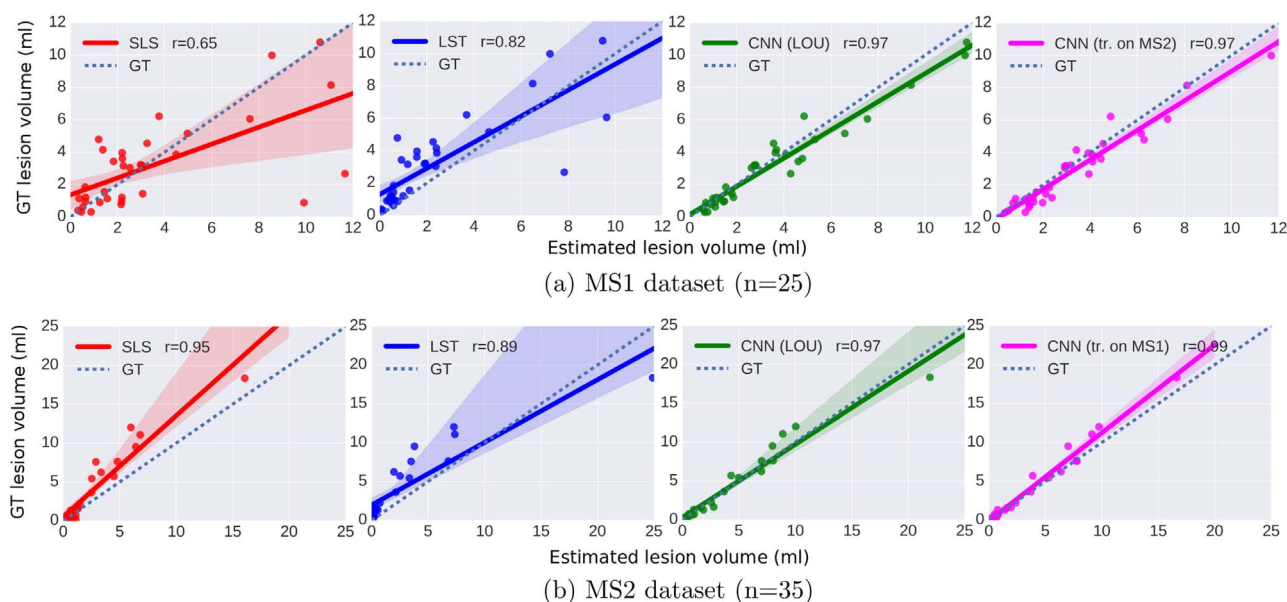
(a) MS1 dataset (n=25)

(b) MS2 dataset (n=35)

Fig. 5. Correlation between estimated lesion volume and manual WM lesion annotations for each of the evaluated methods on MS clinical databases MS1 (a) and MS2 (b). The linear regression model relating estimated and manual segmentation are plotted (solid lines) along with confidence intervals at 95%. Estimated models and confidence intervals are compared with respect to expect lesion size (dashed line). For each method and dataset, the Pearson's linear correlations between manual and estimated masks are also shown (p ≤ 0.001). See Table 4 for the Spearman's correlation coefficients.

The obtained results are encouraging, yielding our CNN architecture closer to human expert inter-rater variability. However, still more research is needed to accomplish this task. Meanwhile, we strongly believe that the proposed method can be a valid alternative for automated WM lesion segmentation.

Acknowledgements

References

Bergstra, J., Bastien, F., Breuleux, O., Lamblin, P., Pascanu, R., Delalleau, O., Desjardins, G., Warde-Farley, D., Goodfellow, I., Bergeron, A., Bengio, Y., 2011. Theano: Deep Learning on GPUs with Python. J. Mach. Learn. Res. 1, 1–48.

Boyes, R.G., Gunter, J.L., Frost, C., Janke, A.L., Yeatman, T., Hill, D.L.G., Bernstein, M.A., Thompson, P.M., Weiner, M.W., Schuff, N., Alexander, G.E., Killiany, R.J., DeCarli, C., Jack, C.R., Fox, N.C., 2008. Intensity non-uniformity correction using N3 on 3-T scanners with multichannel phased array coils. NeuroImage 39, 1752–1762.

Brosch, T., Tang, L.Y.W., Yoo, Y., Li, D.K.B., Traboulsee, A., Tam, R., 2016. Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, 35(5): pp. 1229–1239.

Cabezas, M., Oliver, A., Valverde, S., Beltran, B., Freixenet, J., Vilanova, J.C., Ramió-Torrentà, L., Rovira, À., Lladó, X., 2014. BOOST: A supervised approach for multiple sclerosis lesion segmentation. J. Neurosci. Methods 237, 108–117.

Cabezas, M., Corral, J.F., Oliver, A., Díez, Y., Tintoré, M., Auger, C., Montalban, X., Lladó, M., Pareto, D., Rovira, À., 2016. Improved automatic detection of new T2 lesions in multiple sclerosis using deformation fields. AJNR. Am. J. Neuroradiol., 1–8.

Calabrese, M., Bernardi, V., Atzori, M., Mattisi, I., Favaretto, A., Rinaldi, F., Perini, P., Gallo, P., 2012. Effect of disease-modifying drugs on cortical lesions and atrophy in relapsing-remitting multiple sclerosis. Mult. Scler. 18 (4), 418–424.

Chen, H., Dou, Q., Yu, L., Heng, P.-A., 2016. VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation. arXiv:1608.05895v1, (August):pp. 1–9.

Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O., 2016. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation.

Commowick, O., Cervenansky, F., Ameli, R., 2016. MSSEG Challenge Proceedings: Multiple Sclerosis Lesions Segmentation Challenge Using a Data Management and Processing Infrastructure. France.

Compston, A., Coles, A., 2008. Multiple sclerosis. Lancet 372 (9648), 1502–1517.

Fartaria, M.J., Bonnier, G., Roche, A., Kober, T., Meuli, R., Rotzinger, D., Frackowiak, R., Schluep, M., Du Pasquier, R., Thiran, J.-P., Krueger, G., Bach Cuadra, M., Granziera, C., 2016. Automated detection of white matter and cortical lesions in early stages of multiple sclerosis. J. Magn. Reson. Imaging 43, 1445–1454.

Filippi, M., Rocca, M.A., Ciccarelli, O., De Stefano, N., Evangelou, N., Kappos, L., Rovira, A., Sastre-Garriga, J., Tintorè, M., Frederiksen, J.L., Gasperini, C., Palace, J., Reich, D.S., Banwell, B., Montalban, X., Barkhof, F., 2016. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines.

García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D., Collins, D., 2013. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med. Image Anal. 17 (1), 1–18.

Geremia, E., Clatz, O., Menze, B., Konukoglu, E., Criminisi, A., Ayache, N., 2011. Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images. NeuroImage 57 (2), 378–390.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS), 9: pp. 249–256.

Griffanti, L., Zamboni, G., Khan, A., Li, L., Bonifacio, G., Sundaresan, V., Schulz, U.G., Kuker, W., Battaglini, M., Rothwell, P.M., Jenkinson, M., 2016. BIANCA (Brain Intensity AbNormality Classification Algorithm): A new tool for automated segmentation of white matter hyperintensities. NeuroImage 141, 191–205.

Guizard, N., Coupé, P., Fonov, V., Manjón, J., Arnold, D., Collins, D., 2015. Rotation-invariant multi-contrast non-local means for MS lesion segmentation. NeuroImage: Clin. 8, 376–389.

Harmouche, R., Subbanna, N., Collins, D., Arnold, D., Arbel, T., 2015. Probabilistic multiple sclerosis lesion classification based on modeling regional intensity variability and local neighborhood information. IEEE Trans. Bio-Med. Eng. 62 (5), 1281–1292.

Havaei, M., Guizard, N., Chapados, N., Bengio, Y., 2016. HeMIS: Hetero-Modal Image Segmentation. Springer International Publishing, 469–477.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2017. Brain tumor segmentation with Deep Neural Networks. Med. Image Anal. 35, 18–31.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.

Ioffe, S., Szegedy, C., 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. J. Mach. Learn. Res., 37.

Jerman, T., Galimzianova, A., Pernuš, F., Likar, B., Špiclin, Ž., 2016. Combining Unsupervised and Supervised Methods for Lesion Segmentation. Springer International Publishing, Cham, 45–56.

Jesson, A., Arbel, T., 2015. Hierarchical MRF and random forest segmentation of ms lesions and healthy tissues in brain MRI, pp. 1–2.

Kamnitsas, K., Ledig, C., Newcombe, V.F.J., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B., 2016. Efficient multi-scale 3D {CNN} with fully connected {CRF} for accurate brain lesion segmentation. Med. Image Anal..

Krizhevsky, A., Sutskever, I., Hinton, G., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1106–1114.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444.

Lee, B., Newberg, A., 2005. Neuroimaging in traumatic brain imaging. NeuroRx: J. Am.

Soc. Exp. Neurother. 2 (2), 372–383.

Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, A., 2012. Segmentation of multiple sclerosis lesions in brain MRI: A review of automated approaches. Inf. Sci. 186 (1), 164–185.

Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., Lanczi, L., Gerstner, E., Weber, M.A., Arbel, T., Avants, B.B., Ayache, N., Buendia, P., Collins, D.L., Cordier, N., Corso, J.J., Criminisi, A., Das, T., Delingette, H., Demiralp, a., Durst, C.R., Dojat, M., Doyle, S., Festa, J., Forbes, F., Geremia, E., Glocker, B., Golland, P., Guo, X., Hamamci, A., Iftekharuddin, K.M., Jena, R., John, N.M., Konukoglu, E., Lashkari, D., Mariz, J.A., Meier, R., Pereira, S., Precup, D., Price, S.J., Raviv, T.R., Reza, S.M.S., Ryan, M., Sarikaya, D., Schwartz, L., Shin, H.C., Shotton, J., Silva, C.A., Sousa, N., Subbanna, N.K., Szekely, G., Taylor, T.J., Thomas, O.M., Tustison, N.J., Unal, G., Vasseur, F., Wintermark, M., Ye, D.H., Zhao, L., Zhao, B., Zikic, D., Prastawa, M., Reyes, M., Van Leemput, K., 2015. The multimodal Brain Tumor Image Segmentation Benchmark (BRATS). IEEE Trans. Med. Imaging 34 (10), 1993–2024.

Moeskops, P., Benders, M.J., Chi, S.M., Kersbergen, K.J., Groenendaal, F., de Vries, L.S., Viergever, M.A., Išgum, I., 2015. Automatic segmentation of MR brain images of preterm infants using supervised classification. NeuroImage 118, 628–641.

Moeskops, P., Viergever, M.A., Mendrik, A.M., de Vries, L.S., Benders, M.J.N.L., Isgum, I., (2016a). Automatic segmentation of MR brain images with a convolutional neural network, 35(5): pp. 1252–1261.

Moeskops, P., Wolterink, J., van der Velden, B., Gilhuijs, K., Leiner, T., Viergever, M., Išgum, I., 2016b. Deep learning for multi-task medical image segmentation in multiple modalities, volume 9901 LNCS.

Nair, V., Hinton, G.E., 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. In: Proceedings of the 27th International Conference on Machine Learning, (3): pp. 807–814.

Pereira, S., Pinto, A., Alves, V., Silva, C.A., 2016. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imaging 35 (5), 1240–1251.

Polman, C., Reingold, S., Banwell, B., Clanet, M., Cohen, J.a., Filippi, M., Fujihara, K., Havrdova, E., Hutchinson, M., Kappos, L., Lublin, F., Montalban, X., O'Connor, P., Sandberg-Wollheim, M., Thompson, A., Waubant, E., Weinshenker, B., Wolinsky, J., 2011. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. Ann. Neurol. 69 (2), 292–302.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015a. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57 (10), 1031–1043.

Roura, E., Sarbu, N., Oliver, A., Valverde, S., 2015b. Automated detection of Lupus white matter lesions in MRI images, pp. 1–7.

Rovira, À., Wattjes, M.P., Tintoré, M., Tur, C., Yousry, T.a., Sormani, M.P., de Stefano, N., Filippi, M., Auger, C., Rocca, M.a., Barkhof, F., Fazekas, F., Kappos, L., Polman, C., Miller, D., Montalban, X., 2015. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis clinical implementation in the diagnostic process. Nat. Rev. Neurol. 11, 1–12, (August).

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V., Zimmer, C., Hemmer, B., Mühlau, M., 2012. An automated tool for

detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. NeuroImage 59 (4), 3774–3783.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. ImageNet Chall., 1–10.

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 (1), 87–97.

Smith, S.M., Zhang, Y., Jenkinson, M., Chen, J., Matthews, P., Federico, A., De Stefano, N., 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17 (1), 479–489.

Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R., 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. (JMLR) 15, 1929–1958.

Steenwijk, M.D., Pouwels, P.J.W., Daams, M., van Dalen, J.W., Caan, M.W.a., Richard, E., Barkhof, F., Vrenken, H., 2013. Accurate white matter lesion segmentation by k nearest neighbor classification with tissue type priors (kNN-TTPs). NeuroImage: Clin. 3, 462–469.

Steinman, L., 1996. Multiple sclerosis: a coordinated immunological attack against myelin in the central nervous system. Cell 85 (3), 299–302.

Strumia, M., Schmidt, F., Anastasopoulos, C., Granziera, C., Krueger, G., Brox, T., 2016. White matter MS-lesion segmentation using a geometric brain model. IEEE Trans. Med. imaging 99, 1, (PP).

Styner, M., Lee, J., Chin, B., Chin, M., 2008. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. Midas, 1–6.

Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Adv. Neural Inf. Process. Syst. (NIPS), 3104–3112.

Tomas-Fernandez, X., Warfield, S., 2015. A Model of Population and Subject (MOPS) intensities with application to multiple sclerosis lesion segmentation. IEEE Trans. Med. imaging 0062 (c), 1–15.

Valverde, S., Oliver, A., Lladó, X., 2014. A white matter lesion-filling approach to improve brain tissue volume measurements. NeuroImage: Clin. 6, 86–92.

Valverde, S., Oliver, A., Roura, E., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Sastre-Garriga, J., Montalban, X., Rovira, A., Lladó, X., 2015. Quantifying brain tissue volume in multiple sclerosis with automated lesion segmentation and filling. NeuroImage: Clin. 9, 640–647.

Wang, L., Gao, Y., Shi, F., Li, G., Gilmore, J.H., Lin, W., Shen, D., 2015. LINKS: Learning-based multi-source IntegratioN frameworK for Segmentation of infant brain images. NeuroImage 108, 160–172.

Wolterink, J.M., Leiner, T., de Vos, B.D., van Hamersvelt, R.W., Viergever, M.A., I??gum, I., 2016. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med. Image Anal. 34, 123–136.

Zeiler, M.D., 2012. ADADELTA: An Adaptive Learning Rate Method. ArXiv preprint arXiv:1212.5701, page 6.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. 108, pp. 214–224.

Zheng, W., Chee, M.W.L., Zagorodnov, V., 2009. Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. NeuroImage 48, 73–83.