Location Sensitive Deep Convolutional Neural Networks for Segmentation of White Matter Hyperintensities

Mohsen Ghafoorian[1a,b], Nico Karssemeijer[b], Tom Heskes[a], Inge W.M. van Uden[c], Clara I. Sanchez[b], Geert Litjens[b], Frank-Erik de Leeuw[c], Bram van Ginneken[b], Elena Marchiori[a], Bram Platel[b]

[a]*Institute for Computing and Information Sciences, Radboud University, Nijmegen, The Netherlands*
[b]*Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, Nijmegen, The Netherlands*
[c]*Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, The Netherlands*

Abstract

The anatomical location of imaging features is of crucial importance for accurate diagnosis in many medical tasks. Convolutional neural networks (CNN) have had huge successes in computer vision, but they lack the natural ability to incorporate the anatomical location in their decision making process, hindering success in some medical image analysis tasks.

In this paper, to integrate the anatomical location information into the network, we propose several deep CNN architectures that consider multi-scale patches or take explicit location features while training. We apply and compare the proposed architectures for segmentation of white matter hyperintensities in brain MR images on a large dataset. As a result, we observe that the CNNs that incorporate location information substantially outperform a conventional segmentation method with hand-crafted features as well as CNNs that do not integrate location information. On a test set of 46 scans, the best configuration of our networks obtained a Dice score of 0.791, compared to 0.797 for an independent human observer. Performance levels of the machine and the independent human observer were not statistically significantly different (p-value=0.17).

Keywords: white matter hyperintensities, white matter lesions, small vessel disease, automated segmentation, deep learning, convolutional neural networks

1. Introduction

White matter hyperintensities (WMH), also known as leukoaraiosis or white matter lesions are a common finding on brain MR images of patients diagnosed with small vessel disease (SVD) [1], multiple sclerosis [2], Parkinsonism [3], stroke [4], Alzheimer's disease [5] and Dementia [6]. WMHs often represent areas of demyelination found in the white matter of the brain, but they can also be caused by other mechanisms such as edema. WMHs are best observable in fluid-attenuated inversion recovery (FLAIR) MR images, as high value signals [7]. The prevalence of WMHs among SVD patients has been reported to reach up to 95% depending on the population studied and the imaging technique used [8]. Studies have reported a relationship between WMH severity and other neurological disturbances and symptoms including cognitive decline [9, 10], gait dysfunction [11], hypertension [12] as well as depression [13] and mood disturbances [14]. It has been shown that using a more accurate WMH volumetric assessment, a better association with clinical measures of physical performance and cognition is achieved [15].

Accurate quantification of WMHs in terms of total volume and distribution is believed to be of clinical importance for prognosis, tracking of disease progression and assessment of the treatment effectiveness [16]. However, manual segmentation of WMHs is a laborious time consuming task that makes it infeasible for larger datasets and in clinical practice. Furthermore, manual segmentation is subject to considerable inter- and intra-rater variability [17].

In the last decade, many automated and semi-automated algorithms have been proposed that can be classified into two general categories. Some methods use supervised machine learning algorithms, often using hand-crafted features [18–28] or more recently with learned representations [29–33]. This is while other methods use unsupervised approaches [34–41] to cluster WMHs as outliers or model them with additional classes. Although a multitude of approaches has been suggested for this problem, a truly reliable fully automated method that performs as good as human readers has not been identified [42, 43].

Deep neural networks [44, 45] are biologically plausible

[1]Email: mohsen.ghafoorian@radboudumc.nl, phone: +31 243655793, fax: +31 24 3652728

learning structures, inspired by early neuroscience-related work [46, 47] and have so far claimed human level or super-human performances in several different domains [48–52]. Convolutional neural networks (CNN) [53], perhaps the most popular form of deep neural networks, have attracted enormous attention from the computer vision community since Alex Krizhevsky's network [54] won the Imagenet competition [55] by a large margin. Although the initial focus of CNN methods was concentrated on image classification, soon the framework was extended to cover segmentation as well. A natural way to apply CNNs to segmentation tasks is to train a network in a sliding-window setup to predict the label of each pixel/voxel considering a local neighborhood, which is usually referred to as a patch [51, 56–58]. Later fully convolutional neural networks were proposed to computationally optimize the segmentation process [59, 60].

Deep neural networks have recently been widely used in many medical image analysis domains including lesion detection, image segmentation, shape modeling and image registration [61]. In particular on neuroimaging, several studies are proposed using CNNs for brain extraction [62], tissue and anatomical region segmentation [63–68], tumor segmentation [69–72], microbleed detection [73, 74], and brain lesion segmentation [30–33].

In many bio-medical segmentation applications, including the segmentation of WMHs [18, 23, 42, 43, 75], anatomical location information plays an important role for an accurate classification of voxels (see Figure 1). In contrast, in commonly used segmentation benchmarks in the computer vision community, such as general scene labeling and crowd segmentation, it is normally not a valid assumption to consider pixel/voxel spatial location as an important piece of information. This explains why the literature lacks enough studies investigating ways to integrate spatial information into CNNs.

In this study, we train a number of CNNs to build systems for an accurate fully-automated segmentation of WMHs. We train, validate and evaluate our networks with a large dataset of more than 500 patients, that enables us to learn optimal values for millions of weights in our deep networks. In order to feed the CNN with location information, it is possible to incorporate multi-scale patches or add an explicit set of spatial features to the network. We evaluate and compare three different strategies and network architectures for providing the networks with more context/spatial location information. Experimental results suggest not only our best performing network outperforms a conventional segmentation method with hand-crafted features with a considerable margin, but also its performance does not significantly differ from an independent human observer.
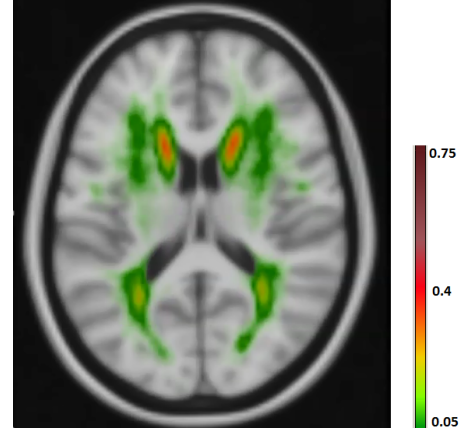

Figure 1: A pattern is observable in WMHs occurrence probability map.

2. Materials

2.1. Data

The research presented in this paper uses data from a longitudinal study called the Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC) [1]. Baseline scanning was performed in 2006. The patients were rescanned in 2011/2012 and currently a third follow-up is being acquired.

2.1.1. Subjects

Subjects for the RUN DMC study were selected at baseline based on the following inclusion criteria [1]: (a) aged between 50 and 85 years (b) cerebral SVD on neuroimaging (appearance of WMHs and/or lacunes). Exclusion criteria comprised: presence of (a) dementia (b) parkinson(-ism) (c) intracranial hemorrhage (d) life expectancy less than six months (e) intracranial space occupying lesion (f) (psychiatric) disease interfering with cognitive testing or follow-up (g) recent or current use of acetylcholine-esterase inhibitors, neuroleptic agents, L-dopa or dopa-a(nta)gonists (h) non-SVD related WMH (e.g. MS) (i) prominent visual or hearing impairment (j) language barrier and (k) MRI contraindications. Based on these criteria, MRI scans of 503 patients were taken at baseline.

2.1.2. Magnetic resonance imaging

The machine used for the baseline was a single 1.5 Tesla scanner (Magnetom Sonata, Siements Medical Solution, Erlangen, Germany). Details of the imaging protocol are listed in Table 1.

2.1.3. Reference annotations

Reference annotations were created in a slice by slice manner by two experienced raters, manually contouring hy-

Table 1: MR imaging protocol specification for the T1 and FLAIR modalities.

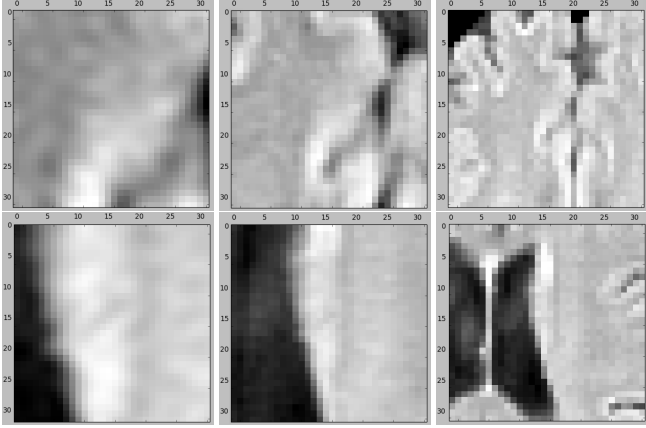| Modality	TR/TE/TI	Flip angle	Voxel size	Interslice gap
T1	2250/3.68/850 ms	15°	1.0×1.0×1.0	0
FLAIR	9000/84/2200 ms	15°	1.2×1.0×5.0	1 mm

Figure 2: An example of negative (top row) and positive (bottom row) samples in three scales (from left to right) 32×32, 64×64 and 128×128 on the FLAIR image. The two larger scales are down sampled to 32×32.

perintense lesions on FLAIR MRI that did not show corresponding cerebrospinal fluid like hypo-intense lesions on the T1 weighted image. Gliosis surrounding lacunes and territorial infarcts were not considered to be WMH related to SVD [76]. One of the observers (observer 1) manually annotated all of the cases. 50 of these 503 images were selected at random and were annotated also by another human observer (observer 2).

2.2. Preprocessing

Before supplying the data to our networks, we first pre-processed the data with the following four steps:

2.2.1. Multi-modal registration

Due to possible movement of patients during scanning, the image coordinates of the T1 and FLAIR modalities might not represent the same location. Thus we transformed the T1 image to align with the FLAIR image in the native space using FSL-FLIRT [77] implementation of rigid registration with trilinear interpolation and mutual information optimization criteria. Also to obtain a mapping between patient space and an atlas space, all subjects were non-linearly registered to the ICBM152 atlas [78] using FSL-FNIRT [79].

2.2.2. Brain extraction

In order to extract the brain and exclude other structures, such as skull, eyes, etc., we apply FSL-BET [80] on T1 images, because this modality has the highest resolution. The resulting mask is then transformed using registration transformation and is applied to the FLAIR images.

2.2.3. Bias field correction

Bias field correction is another necessary step due to magnetic field inhomogeneity. We apply FSL-FAST [81], which uses a hidden Markov random field and an associated expectation-maximization algorithm to correct for spatial intensity variations caused by RF inhomogeneities.

2.2.4. Intensity normalization

Apart from intensity variations caused by the bias field, intensities can also vary between patients. Thus we normalize the intensities per patient to be within the range of [0, 1].

2.3. Training, validation and test sets

From the 503 RUN DMC cases, we removed a number of cases that were extremely noisy or had failed in some of the preprocessing steps including brain extraction and registration, which left us with 420 out of 453 cases with single annotations and 46 cases out of 50 with double annotations. From 420 cases annotated by one human observer, we select 378 cases for training the model and the remaining 42 cases for validation and parameter tuning purposes. We use the 46 cases that were annotated by both human observers as independent test set. All the four cases that we left out from the test set were because of presence of severe noise as a result of head movement during image acquisition.

Medical datasets usually suffer from the fact that pathological observations are significantly less frequent compared to healthy observations, which also holds for our dataset. Given this, a simple uniform sampling may cause serious problems for the learning process [82], as a classifier that labels all of the samples as normal, would achieve a high accuracy. To handle this, we undersample the negative samples to create a balanced dataset. We randomly select 50% of positive and select an equal number of negative samples from normal voxels of all cases. This sampling procedure resulted in datasets consisting of 3.88 million
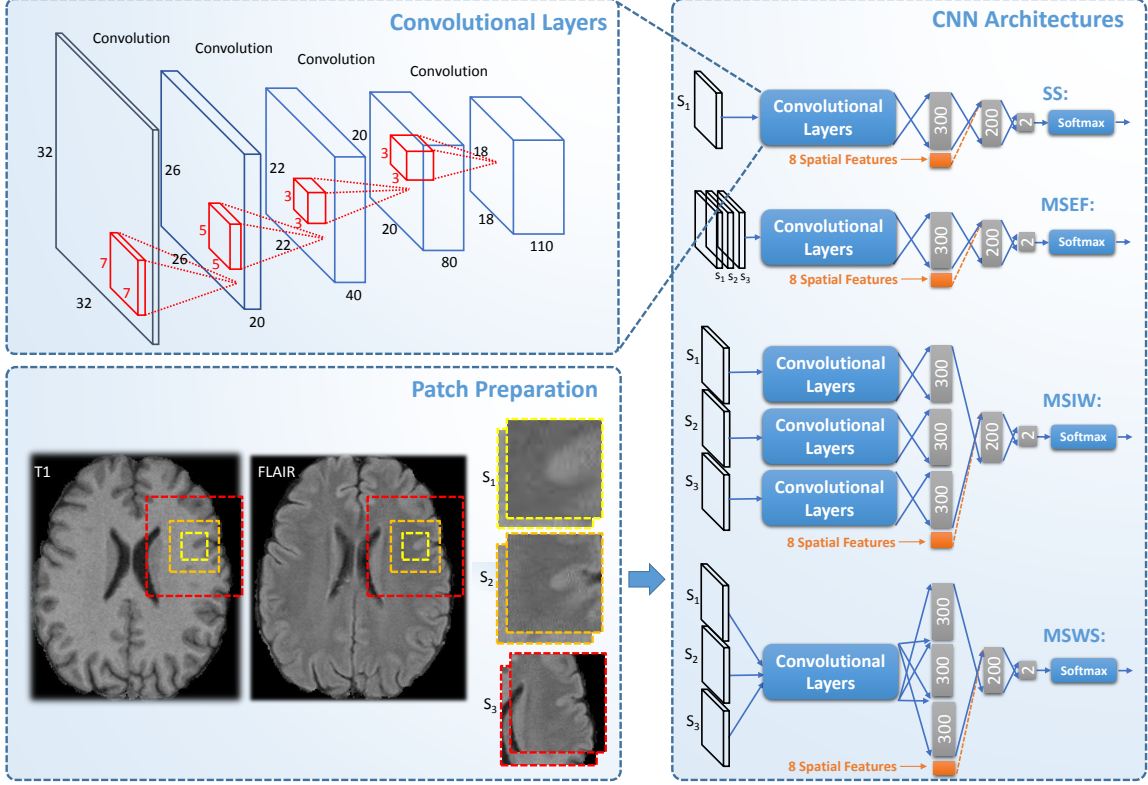
Figure 3: Patch preparation process and different proposed CNN architectures.

and 430 thousand samples for training and validation sets respectively.

3. Methods

3.1. Patch preparation

From each voxel neighborhood, we extract patches with three different sizes: 32×32, 64×64 and 128×128. To reduce the computational costs, we down sample the larger two scales to 32×32. Resulting patches for this procedure are demonstrated in Figure 2, for a negative and a positive sample, obtained from a FLAIR image. We included these three patches for both the T1 and FLAIR modalities for each sample. This results in a set of patches in three scales s_1, s_2 and s_3, each consisting of two patches from T1 and FLAIR, as depicted in Figure 3.

3.2. Network architectures

3.2.1. Single-scale (SS) model

The simplest CNN model we applied to our dataset was a CNN trained on patches from a single scale (with patches of 32×32). The top architecture in Figure 3 shows the architecture of our single-scale deep CNN. This network, which is a basis for the other location sensitive architectures, consists of four convolutional layers that have 20, 40,

80 and 110 filters of size 7×7, 5×5, 3×3, 3×3 respectively. We do not use pooling since it results in a shift-invariance property [83], which is not desired in segmentation tasks. Then we apply three layers of fully connected neurons of size 300, 200 and 2. Finally the resulting responses are turned into probability values using a softmax classifier.

3.2.2. Multi-scale early fusion (MSEF)

In many cases, it is impossible to correctly classify a 32×32 patch just from its appearance. For instance, only looking at the small scale positive patch in Figure 2, it is hard to distinguish it from cortex tissue. In contrast, given the two larger scale patches, it is fairly easy to identify it as WMH tissue near the ventricles. Furthermore there is a trade-off between context capturing and localization accuracy. Although more context information might be captured with a larger patch-size, the ability of the classifier to accurately localize the structure in the center of the patch is decreased [60]. This motivates a multi-scale approach that has the advantages of the smaller and larger size patches. A simple and intuitive way to train a multi-scale network is to accumulate the different scales as different channels of the input. This is possible since the larger scale patches were down sampled to 32×32. The second top network in Figure 3 illustrates this.

3.2.3. Multi-scale late fusion with independent weights (MSIW)

Another possibility to create a model with multi-scale patches is to train independent convolutional layers for each scale, fusing the representations of each scale and taking them into more fully connected layers. As can be observed in Figure 3, in this architecture each scale has its own fully connected layer. These are concatenated and fed into the joint fully connected layers. The main rationale behind giving each scale stream its own fully connected layer is that this incurs less weights compared to the approach that firsts merges the feature maps and then fully connects it to the first layer of neurons.

3.2.4. Multi-scale late fusion with weight sharing (MSWS)

The first convolutional layers of a CNN typically detect various forms of edges, corners and other basic structuring elements. Since we do not expect that these basic building blocks differ much among the different scale patches, a considerable number of filters might be very similar in the three separate convolutional layers learned for different scales. Thus a potentially efficient strategy to reduce the number of weights and consequently to reduce the overfitting, is to share the convolutional filters among the different scales. As illustrated in Figure 3, each of the scales from the different patches are separately passed through a unified set of convolutional layers and each get described with separate feature maps. These feature maps are then connected to separate fully connected layers and are merged later, similar to the MSIW approach.

3.2.5. Integrating explicit spatial location features

The main aim for considering patches at different scales is to let the network learn about the spatial location of the samples it is observing. Alternatively we can provide the network with such information, by adding explicit features describing the spatial location. One possible place to add the location information is the first fully connected layer after the convolutional layers. All the location features are normalized per case to be within the range of [0, 1]. As the response of other neurons in the same layer that the location features are integrated with might have a different scale, all the eight features are scaled with a coefficient α as a parameter of the method. We tuned the best value for α as a parameter by validation. The possibility to add spatial location features is not restricted to the single-scale architecture. It is also feasible to integrate these features into the three possible architectures for multi-scale approaches. The orange parts in Figure 3 illustrate this procedure.

There are eight features that we utilize to describe the spatial location: the x, y and z-coordinates of the corresponding voxel in the MNI atlas space, in-plane distances from the left ventricle, right ventricle, brain cortex and midsagittal brain surface as well as the prior probability of WMH occurring in that location [23].

3.3. Training procedure

For learning the network weights, we use the stochastic gradient descent algorithm [84], with mini-batch size of 128 and a cross-entropy cost. We also utilize the RM-SPROP algorithm [85] to speed up the learning process by adaptively changing the learning rate for each parameter. The non-linearity applied to neurons is a rectified linear unit to prevent the vanishing gradient problem [86]. As random weight initialization is important to break the symmetry between the units in the same hidden layer [87], the initial weights are drawn at random using the Glorot method [88]. Since CNNs are complex architectures, they are prone to overfit the data very early. Therefore we use drop-out regularization [89] with 0.3 probability on all fully connected layers of the networks. We pick the resulting network from an epoch with the highest validation A_z as the final model.

4. Experimental Evaluation

For characterization of WMHs, several different methods have been proposed in this study, some of which only use patch appearance features, while others use multi-scale patches or explicit location features to the network or both. In order to obtain segmentations, we apply the trained networks to classify all the voxels inside the brain mask in a sliding window fashion. A comparison between the performance of the mentioned methods, together with a comparison to performance of an independent human observer and a conventional method with hand-crafted features would be insightful.

Integrating the location information into the first fully connected layer, as depicted in the architectures Figure 3, is only one of the possibilities. We can alternatively add the spatial location features to one layer before or after, i.e. to the responses from the last convolutional layer and to the second fully connected layer. To evaluate the relative performance of each possibility, we also train single-scale networks with the two other possibilities and compare them to each other. In order to provide information on how much effect the dataset size has on the performance of the trained network, we present and compare the results of a MSWS+Loc network trained with 100%, 50%, 25%, 12.5% and 6.25% of the total training images.

4.1. Metrics

The Dice similarity index, also known as the Dice score, is the most widely used measure for evaluating the agreement between different segmentation methods and their

Table 2: Performance comparison of different CNN architectures based on validation set A$_z$ and test set Dice score considering observer 1 and observer 2 as the reference standard.

| Method | Without location features | | | With location features | | |
	Validation set A$_z$	Test set Dice (obs1)	Test set Dice (obs2)	Validation set A$_z$	Test set Dice (obs1)	Test set Dice (obs2)
SS	0.9939	0.730	0.729	0.9972	0.783	0.778
MSEF	0.9947	0.758	0.747	0.9966	0.783	0.771
MSIW	0.9966	0.775	0.762	0.9972	0.791	0.781
MSWS	0.9965	0.773	0.759	0.9973	0.791	0.780

reference standard segmentations. [42, 43]. It is computed as

$$Dice = \frac{2 \times TP}{FP + FN + 2 \times TP} \qquad (1)$$

where the value varies between 0 for complete disagreement, and 1 representing complete agreement between the reference standard and the evaluated segmentation. A Dice similarity index of 0.7 or higher is usually considered a good segmentation in the literature [42]. To create binary masks out of probability maps resulting from CNNs, we find an optimal value as a threshold that maximizes the overall Dice score on the validation set. The optimal thresholds are computed separately for each method. We also present test set receiver operating characteristic (ROC) curves and validation set area under the ROC curve (A$_z$). For computing each of these measures, we only consider the voxels inside the brain mask, to avoid taking easy voxels belonging to the background into account.

For the statistical significance test, we created a 100 bootstraps by sampling 46 instances with replacement. Then the Dice scores were computed on each bootstrap for each of the two compared methods. Empirical p-values were reported as the proportion of bootstraps where the Dice score for method B was higher than A, when the null-hypothesis to reject was "method A is no better than B". If no such bootstrap existed, the p-value<0.01 was reported, representing a significant difference.

4.2. Conventional segmentation system

In order to evaluate the relative performance of the proposed deep learning systems, we also train a conventional segmentation system, using hand-crafted features [23]. The set of hand-crafted features consists of 22 features in total: intensity features including FLAIR and T1 intensities, second order derivative features including multi-scale Laplacian of Gaussian (σ=1,2,4 mm), multi-scale determinant of Hessian (t=1,2,4 mm), vesselness filter (σ=1 mm), a multi-scale annular filter (t=1,2,4 mm), FLAIR intensity mean and standard deviation in a 16×16 neighborhood, as well as the same 8 location features that were used in the previous subsection. We use a random forest classifier with 50 subtrees to train the model.

Table 5: A performance comparison of the single-scale architecture with different possible locations to add the spatial location information. Abbreviations: last convolutional layer (LCL), first fully connected layer (FFCL), second fully connected layer (SFCL).

Method	Validation set A_z	Test set Dice
LCL	0.9964	0.759
FFCL	0.9971	0.783
SFCL	0.9967	0.771

5. Experimental Results

Table 2 represents a comparison on validation set A_z and test set Dice score, for each of the methods, once without and another time with addition of spatial location features, considering observer 1 as the reference standard. Table 3 compares the performance of the conventional segmentation method, our late fusion multi-scale architecture with weight sharing and location information (MSWS+Loc), and the two human observers on the independent test set, with each observer as the reference standard.

P-values were computed as a result of patient-level bootstrapping on the test set and are presented in Table 4.

Regarding the different options for integration of the location information in the network, Table 5 compares the performance of these options on the validation and training sets. Adding the spatial location information to the first fully connected layer results in a significantly better Dice score compared to the other two possibilities (p-value < 0.01).

Figure 4a shows the ROC curves for some of the trained CNN architectures and compares them to the conventional segmentation method and the independent human observer. The ROC curves have been cut to show only low false positive rates that are of interest for practical use. In order to preserve readability of the figures, we only compare the most informative methods. Figure 4b shows the Dice similarity scores as a function of the binary masking threshold. It also compares them to the Dice similarity measure between the two human observers. 95% confidence intervals are depicted for each curve, as a result of bootstrapping on patients. The effect of the training

Table 3: A performance comparison between conventional method, MSWS+Loc architecture, and human observers.

Method	Dice (obs1)	Dice (obs2)
Conventional	0.710	0.685
MSWS+Loc	0.791	0.780
observer 1	-	0.797
observer 2	0.797	-

Table 4: Statistical significance test for pairwise comparison of the methods Dice score. p_{ij} indicates the p-value for the null hypothesis that method i is better than method j.

Method	MSWS	SS+Loc	MSWS+Loc	Ind. Obs.
SS	<0.01	<0.01	<0.01	<0.01
MSWS	-	<0.01	<0.01	<0.01
SS+Loc	-	-	0.23	0.15
MSWS+Loc	-	-	-	0.17
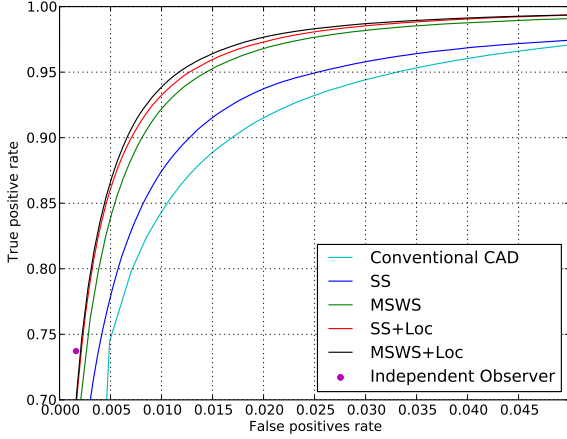
(a) An ROC comparison of different CNN methods, a conventional segmentation method and independent human observer, considering observer 1 as the reference standard.
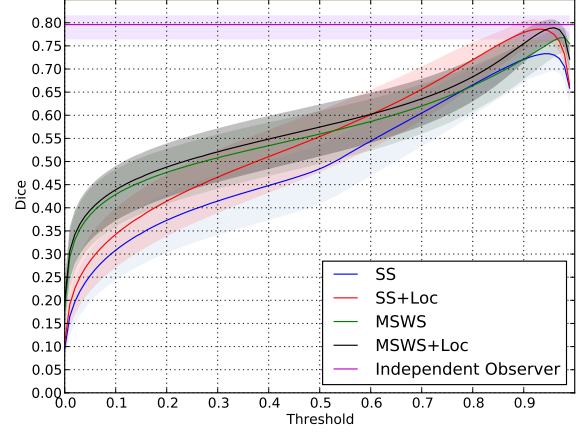
(b) A comparison of different methods on Dice score as a function of binary masking threshold. The light shades around the curves indicate 95% confidence intervals with bootstrapping on patients.

Figure 4: Integration of spatial location information fills the gap between performance of a normal CNN and human observer.

dataset size can be observed in Table 6 and Figure 5.

6. Discussion

6.1. Contribution of larger context and location information

Comparing the performance of the SS and SS+Loc approaches, as presented in the first row of Table 2, a significant difference in Dice score is observable (p-value < 0.01). This points us to the fact that a knowledge about where the input patch is located can substantially improve WMH segmentation quality of a CNN. A similar significant difference is observable when comparing performance measures of SS and MSWS methods (p-value < 0.01). This implies that by using a multi-scale approach, a CNN can learn about context information quite well. Considering the better performance of SS+Loc compared to MSWS, we can infer that the learning of location and large scale context from multi-scale patches is not as good as adding explicit location information to the architecture.
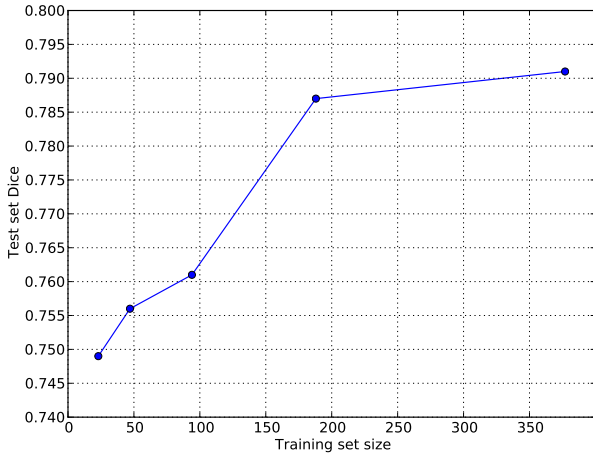
Figure 5: Test Dice as a function of training set size.

Table 6: Test Dice as a function of training set size.

Training set size	23	47	94	189	378
Test set Dice	0.749	0.756	0.761	0.787	0.791

6.2. Early fusion vs. late fusion, independent weights vs. weight sharing

As the experimental results suggest, among the different multi-scale fusion architectures, early fusion shows the least improvement over the single-scale approach. The related patch voxels of different scales, do not have a meaningful correspondence. Given the fact that the convolution operation in the first convolutional layer sums up the responses on each scale, we assume that the useful information provided by different scales is washed out too early in the network. In contrast, the two late fusion architectures show comparable good performance, however in general, since the late fusion architecture with weight sharing is a simpler model with less parameters to be learned, one might prefer to use this model.

6.3. Comparison to human observer and a conventional method

Shown by Table 3, MSWS+Loc substantially outperforms a conventional segmentation method, with Dice score of 0.79 compared to 0.71 (p-value<0.01). Furthermore, the Dice score of MSWS+Loc method closely resembles the inter-observer variability, which implies that the segmentation provided by MSWS+Loc approach is as good as the two human observers. Also the statistical test does not show a significant advantage of the independent observer compared to this method (p-value = 0.17).

6.4. A visual look into the results

Figures 6-8 show some qualitative examples. Figure 6 contains two sample cases, where the location and larger context information leads to a better segmentation. As evident from the first sample, the single-scale CNN falsely segments an area on septum pellucidum, which also appears as hyperintense tissue. These false positives can be avoided by considering location information. A second sample shows improvements on FNs of the single-scale method.

Figure 7 illustrates an instance of a prevalent class of false positives of the system, which are the hyperintense voxels around the lacunes. Since the model has not been trained on so many negative samples similar to this, the distinction between WMH and hyperintensities around lacunes is not well learned by the system. An obvious solution is to extensively include the lacunes surrounding voxels as negative samples in the training dataset.

As an example of missed lesions by human observers, Figure 8 shows a small lesion on the right temporal lobe, missed by both human observers, where it is detected by MSWS+Loc method. Another sample of such missed lesions can be observed in the second sample of Figure 6, on the right hemisphere frontal lobe. Based on similar observations, we can assume that some of the false positives are possibly small lesions missed by one or both of the observers. Therefore there may be a chance that the real performance of the system is better than reported, but it would require more research to investigate this.

6.5. Integration of location features

For integration of explicit spatial location information into the CNN, there are several possibilities that were investigated in this study. The results as represented in Table 3, suggest that adding the spatial location features to the first fully connected layer results in a significantly better performance. Adding them to around 35K features as the responses of the last convolutional layer, almost makes the eight location features insignificant among so many representation features. At the other extreme, although integrating the location features into the second fully connected layer does not suffer from this problem, but leaves less flexibility for the network to consider location features for the discrimination to be learned. The first fully connected layer seems to be the best option, where the appearance features provided by the last convolutional layer are already considerably reduced, and at same time the more fully connected layer provides more flexibility for an optimal discrimination.

6.6. Two-stage vs. single-stage model

As shown in the results, integrating location information into a CNN can play an important role in obtaining an
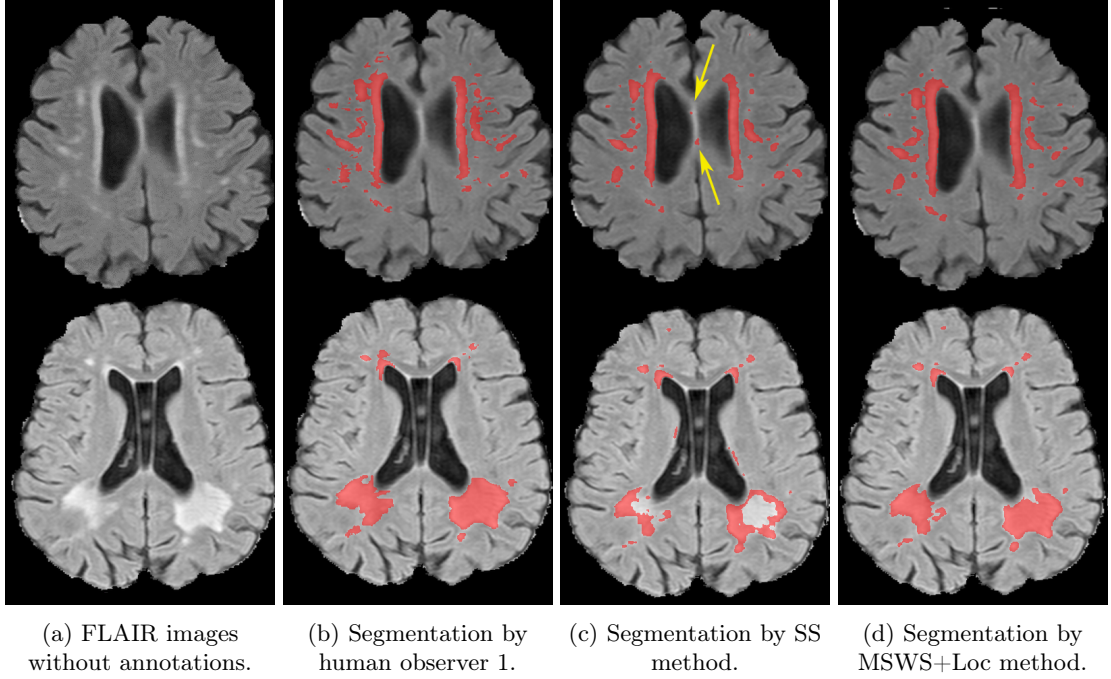
(a) FLAIR images without annotations.	(b) Segmentation by human observer 1.	(c) Segmentation by SS method.	(d) Segmentation by MSWS+Loc method.

Figure 6: Two sample cases of segmentation improvement by adding location information to the network.

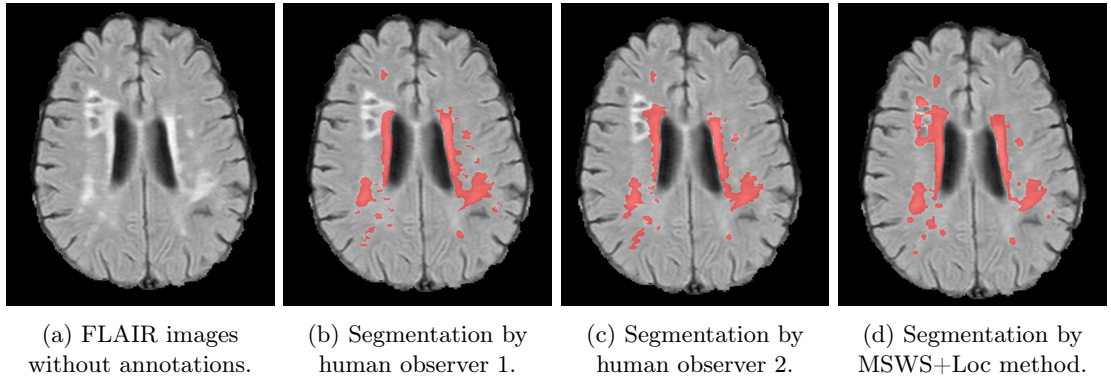|			
(a) FLAIR images without annotations.	(b) Segmentation by human observer 1.	(c) Segmentation by human observer 2.	(d) Segmentation by MSWS+Loc method.

Figure 7: Gliosis around the lacunes is a prevalent type of false positive segmentation.

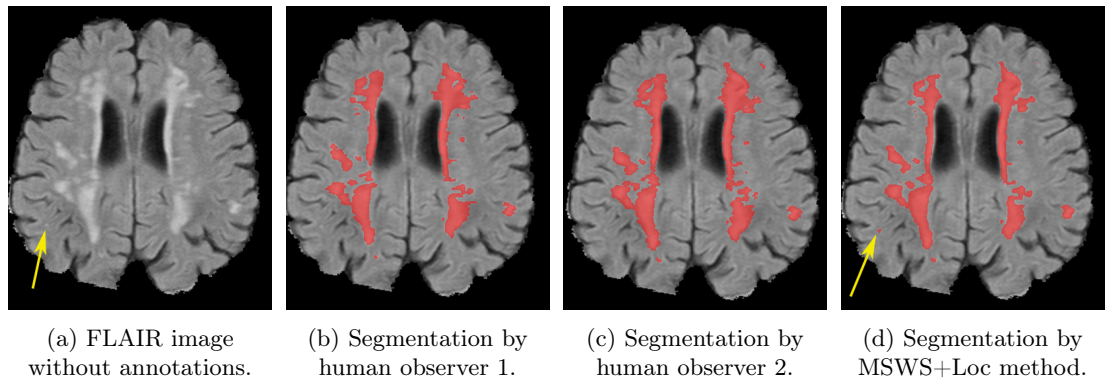|			
(a) FLAIR image without annotations.	(b) Segmentation by human observer 1.	(c) Segmentation by human observer 2.	(d) Segmentation by MSWS+Loc method.

Figure 8: A sample case with a small lesion missed by the two human observers.

accurate segmentation. We integrate the features while we train our network to learn the representations. Another approach is to perform this task in two stages; first training an independent network that learns the represen-

tations, and later training a second classifier that takes the output features of the first network, integrated with location or other external features (as followed in [90] for instance). The first approach, which is followed in this study, seems more reasonable as the set of learned filters without location information could differ from the optimal set of filters given the location information. The two-stage system lacks this information and might devote some of the filters for capturing of location that are redundant given the location features.

6.7. 2D vs. 3D patches

In this research, we sample 2D patches from each of the two modalities (T1 and FLAIR), while one might argue that considering consecutive slices and sampling 3D patches from each image modality could provide useful information. Given the slice thickness of 5 mm with a 1 mm inter-slice gap in our dataset, the consecutive slices do not highly correspond to each other. Furthermore incorporation of 3D patches extensively increases the computational costs at both the training and the segmentation time. These motivated us to use 2D patches. In contrast, for datasets with isotropic or thin slice FLAIR images, 3D patches might be very useful.

6.8. Fully convolutional segmentation network

Fully convolutional networks replace the fully connected layers with $1{\times}1$ convolutions that perform exactly the same functionality as the fully connected do, however implemented with convolutions [59]. This would speed the segmentation up, since convolutions can get larger input images, make dense predictions for the whole input image and avoid repetitive computations. While we have trained our networks in a patch-based manner, it does not restrict us from reforming the fully connected layers of the trained network into convolutional layer counterparts at the segmentation time. The current implementation uses a patch-based segmentation, as we found it fast enough in the current experimental setup ($\sim$3 minutes for the multi-scale and $\sim$1.5 minutes for the single-scale architectures per case on a Titan X card).

7. Conclusions

In this study we showed that location information can have a significant added value when using CNNs for WMH segmentation. While for this task, making use of CNNs, not only a better performance compared to conventional segmentation method was achieved, we approached the performance level of an independent human observer with incorporation of location information.

9. References

[1] A. G. van Norden, K. F. de Laat, R. A. Gons, I. W. van Uden, E. J. van Dijk, L. J. van Oudheusden, R. A. Esselink, B. R. Bloem, B. G. van Engelen, M. J. Zwarts, I. Tendolkar, M. G. Olde-Rikkert, M. J. van der Vlugt, M. P. Zwiers, D. G. Norris, F. E. de Leeuw, Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol, BMC Neurol 11 (2011) 29.

[2] M. M. Schoonheim, R. M. Vigeveno, F. C. R. Lopes, P. J. Pouwels, C. H. Polman, F. Barkhof, J. J. Geurts, Sex-specific extent and severity of white matter damage in multiple sclerosis: Implications for cognitive decline, Human Brain Mapping 35 (5) (2014) 2348–2358.

[3] G. Marshall, E. Shchelchkov, D. Kaufer, L. Ivanco, N. Bohnen, White matter hyperintensities and cortical acetylcholinesterase activity in parkinsonian dementia, Acta Neurologica Scandinavica 113 (2) (2006) 87–91.

[4] G. Weinstein, A. S. Beiser, C. DeCarli, R. Au, P. A. Wolf, S. Seshadri, Brain imaging and cognitive predictors of stroke and alzheimer disease in the framingham heart study, Stroke 44 (10) (2013) 2787–2794.

[5] N. Hirono, H. Kitagaki, H. Kazui, M. Hashimoto, E. Mori, Impact of white matter changes on clinical manifestation of alzheimer's disease a quantitative study, Stroke 31 (9) (2000) 2182–2188.

[6] C. D. Smith, D. A. Snowdon, H. Wang, W. R. Markesbery, White matter volumes and periventricular white matter hyperintensities in aging and dementia, Neurology 54 (4) (2000) 838–842.

[7] J. M. Wardlaw, E. E. Smith, G. J. Biessels, C. Cordonnier, F. Fazekas, R. Frayne, R. I. Lindley, J. T. O'Brien, F. Barkhof, O. R. Benavente, S. E. Black, C. Brayne, M. Breteler, H. Chabriat, C. Decarli, F. E. de Leeuw, F. Doubal, M. Duering, N. C. Fox, S. Greenberg, V. Hachinski, I. Kilimann, V. Mok, R. v. Oostenbrugge, L. Pantoni, O. Speck, B. C. Stephan, S. Teipel, A. Viswanathan, D. Werring, C. Chen, C. Smith, M. van Buchem, B. Norrving, P. B. Gorelick, M. Dichgans, Neuroimaging standards for research into small vessel disease and its contribution to ageing and neurodegeneration, The Lancet Neurology 12 (8) (2013) 822–838.

[8] F. De Leeuw, J. C. de Groot, E. Achten, M. Oudkerk, L. Ramos, R. Heijboer, A. Hofman, J. Jolles, J. Van Gijn, M. Breteler, Prevalence of cerebral white matter lesions in elderly people: a population based magnetic resonance imaging study. the rotterdam scan study, Journal of Neurology, Neurosurgery & Psychiatry 70 (1) (2001) 9–14.

[9] J. C. de Groot, M. Oudkerk, J. v. Gijn, A. Hofman, J. Jolles, M. Breteler, Cerebral white matter lesions and cognitive function: the rotterdam scan study, Annals of Neurology 47 (2) (2000) 145–151.

[10] R. Au, J. M. Massaro, P. A. Wolf, M. E. Young, A. Beiser, S. Seshadri, R. B. D'Agostino, C. DeCarli, Association of white matter hyperintensity volume with decreased cognitive functioning: the framingham heart study, Archives of Neurology 63 (2) (2006) 246–250.

[11] G. Whitman, T. Tang, A. Lin, R. Baloh, A prospective study of cerebral white matter abnormalities in older people with gait dysfunction, Neurology 57 (6) (2001) 990–994.

[12] M. J. Firbank, R. M. Wiseman, E. J. Burton, B. K. Saxby, J. T. O'Brien, G. A. Ford, Brain atrophy and white matter hyperintensity change in older adults and relationship to blood pressure, Journal of Neurology 254 (6) (2007) 713–721.

[13] L. L. Herrmann, M. Le Masurier, K. P. Ebmeier, White matter hyperintensities in late life depression: a systematic review, Journal of Neurology, Neurosurgery & Psychiatry 79 (6) (2008) 619–624.

[14] I. W. van Uden, A. M. Tuladhar, K. F. de Laat, A. G. van Norden, D. G. Norris, E. J. van Dijk, I. Tendolkar, F.-E. de Leeuw, White matter integrity and depressive symptoms in cerebral small vessel disease: The run dmc study, The American Journal of Geriatric Psychiatry 23 (5) (2015) 525–535.

[15] E. C. Van Straaten, F. Fazekas, E. Rostrup, P. Scheltens, R. Schmidt, L. Pantoni, D. Inzitari, G. Waldemar, T. Erkinjuntti, R. Mäntylä, et al., Impact of white matter hyperintensities scoring method on correlations with clinical data the ladis study, Stroke 37 (3) (2006) 836–840.

[16] C. H. Polman, S. C. Reingold, G. Edan, M. Filippi, H. P. Hartung, L. Kappos, F. D. Lublin, L. M. Metz, H. F. McFarland, P. W. O'Connor, M. Sandberg-Wollheim, A. J. Thompson, B. G. Weinshenker, J. S. Wolinsky, Diagnostic criteria for multiple sclerosis: 2005 revisions to the "mcdonald criteria", Annals of Neurology 58 (6) (2005) 840–846.

[17] J. Grimaud, M. Lai, J. Thorpe, P. Adeleine, L. Wang, G. Barker, D. Plummer, P. Tofts, W. McDonald, D. Miller, Quantification of mri lesion load in multiple sclerosis: a comparison of three computer-assisted techniques, Magnetic Resonance Imaging 14 (5) (1996) 495–505.

[18] P. Anbeek, K. L. Vincken, M. J. van Osch, R. H. Bisschops, J. van der Grond, Probabilistic segmentation of white matter lesions in mr imaging, NeuroImage 21 (3) (2004) 1037–1044.

[19] Z. Lao, D. Shen, D. Liu, A. F. Jawad, E. R. Melhem, L. J. Launer, R. N. Bryan, C. Davatzikos, Computer-assisted segmentation of white matter lesions in 3d mr images using support vector machine, Academic Radiology 15 (3) (2008) 300–313.

[20] E. Herskovits, R. Bryan, F. Yang, Automated bayesian segmentation of microvascular white-matter lesions in the accord-mind study, Advances in Medical Sciences 53 (2) (2008) 182–190.

[21] R. Simões, C. Mönninghoff, M. Dlugaj, C. Weimar, I. Wanke, A.-M. v. C. van Walsum, C. Slump, Automatic segmentation of cerebral white matter hyperintensities using only 3d flair images, Magnetic Resonance Imaging 31 (7) (2013) 1182–1189.

[22] V. Ithapu, V. Singh, C. Lindner, B. P. Austin, C. Hinrichs, C. M. Carlsson, B. B. Bendlin, S. C. Johnson, Extracting and summarizing white matter hyperintensities using supervised segmentation methods in alzheimer's disease risk and aging studies, Human Brain Mapping 35 (8) (2014) 4219–4235.

[23] M. Ghafoorian, N. Karssemeijer, I. van Uden, F. E. de Leeuw, T. Heskes, E. Marchiori, B. Platel, Small white matter lesion detection in cerebral small vessel disease, in: SPIE Medical Imaging, Vol. 9414, 2015, pp. 941411–941411.

[24] S. Klöppel, A. Abdulkadir, S. Hadjidemetriou, S. Issleib, L. Frings, T. N. Thanh, I. Mader, S. J. Teipel, M. Hüll, O. Ronneberger, A comparison of different automated methods for the detection of white matter lesions in mri data, NeuroImage 57 (2) (2011) 416–422.

[25] A. P. Zijdenbos, R. Forghani, A. C. Evans, Automatic" pipeline" analysis of 3-d mri data for clinical trials: application to multiple sclerosis, Medical Imaging, IEEE Transactions on 21 (10) (2002) 1280–1291.

[26] T. B. Dyrby, E. Rostrup, W. F. Baare, E. C. van Straaten, F. Barkhof, H. Vrenken, S. Ropele, R. Schmidt, T. Erkinjuntti, L. O. Wahlund, L. Pantoni, D. Inzitari, O. B. Paulson, L. K. Hansen, G. Waldemar, T. Erkinjuntti, T. Pohjasvaara, P. Pihanen, R. Ylikoski, H. Jokinen, M. M. Somerkoski, F. Fazekas, R. Schmidt, S. Ropele, A. Seewann, K. Petrovic, U. Garmehi, J. M. Ferro, A. Verdelho, S. Madureira, P. Scheltens, I. van Straaten, A. Gouw, W. van de Flier, F. Barkhof, A. Wallin, M. Jonsson, K. Lind, A. Nordlund, S. Rolstad, K. Gustavsson, L. O. Wahlund, M. Crisby, A. Pet-

tersson, K. Amberla, H. Chabriat, L. Benoit, K. Hernandez, S. Pointeau, A. Kurtz, D. Reizine, M. Hennerici, C. Blahak, H. Baezner, M. Wiarda, S. Seip, G. Waldemar, E. Rostrup, C. Ryberg, T. B. Dyrby, O. B. Paulson, J. O'Brien, S. Pakrasi, T. Minnet, M. Firbank, J. Dean, P. Harrison, P. English, D. Inzitari, L. Pantoni, A. M. Basile, M. Simoni, G. Pracucci, M. Martini, E. Magnani, A. Poggesi, L. Bartolini, E. Salvadori, M. Moretti, M. Mascalchi, D. Inzitari, T. Erkinjuntti, P. Scheltens, M. Visser, P. Langhorne, Segmentation of age-related white matter changes in a clinical multi-center study, Neuroimage 41 (2) (2008) 335–345.

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

[28] M. Ghafoorian, N. Karssemeijer, I. van Uden, F. E. de Leeuw, T. Heskes, E. Marchiori, B. Platel, Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease, Medical Physics 43.

[29] K. Vijverberg, M. Ghafoorian, I. W. van Uden, F.-E. de Leeuw, B. Platel, T. Heskes, A single-layer network unsupervised feature learning method for white matter hyperintensity segmentation, in: SPIE Medical Imaging, International Society for Optics and Photonics, 2016, pp. 97851C–97851C.

[30] T. Brosch, L. Y. Tang, Y. Yoo, D. K. Li, A. Traboulsee, R. Tam, Deep 3d convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation, IEEE transactions on medical imaging 35 (5) (2016) 1229–1239.

[31] T. Brosch, Y. Yoo, L. Y. Tang, D. K. Li, A. Traboulsee, R. Tam, Deep convolutional encoder networks for multiple sclerosis lesion segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015, Springer, 2015, pp. 3–11.

[32] M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, F.-E. de Leeuw, E. Marchiori, B. van Ginneken, B. Platel, Nonuniform patch sampling with deep convolutional neural networks for white matter hyperintensity segmentation, in: International Symposium on Biomedical Imaging (ISBI), IEEE, 2016, pp. 1414–1417.

[33] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, B. Glocker, Efficient multiscale 3d cnn with fully connected crf for accurate brain lesion segmentation, arXiv preprint arXiv:1603.05959.

[34] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, P. Suetens, Automated segmentation of multiple sclerosis lesions by model outlier detection, Medical Imaging, IEEE Transactions on 20 (8) (2001) 677–688.

[35] L. Shi, D. Wang, S. Liu, Y. Pu, Y. Wang, W. C. Chu, A. T. Ahuja, Y. Wang, Automated quantification of white matter lesion in magnetic resonance imaging of patients with acute infarction, Journal of Neuroscience Methods 213 (1) (2013) 138–146.

[36] A. Khademi, A. Venetsanopoulos, A. R. Moody, Robust white matter lesion segmentation in flair mri, Biomedical Engineering, IEEE Transactions on 59 (3) (2012) 860–871.

[37] F. Admiraal-Behloul, D. Van Den Heuvel, H. Olofsen, M. J. van Osch, J. van der Grond, M. Van Buchem, J. Reiber, Fully automatic segmentation of white matter hyperintensities in mr images of the elderly, Neuroimage 28 (3) (2005) 607–617.

[38] R. de Boer, H. A. Vrooman, F. van der Lijn, M. W. Vernooij, M. A. Ikram, A. van der Lugt, M. M. Breteler, W. J. Niessen, White matter lesion extension to automatic brain tissue segmentation on mri, Neuroimage 45 (4) (2009) 1151–1161.

[39] S. Jain, D. M. Sima, A. Ribbens, M. Cambron, A. Maertens, W. Van Hecke, J. De Mey, F. Barkhof, M. D. Steenwijk, M. Daams, F. Maes, S. Van Huffel, H. Vrenken, D. Smeets, Automatic segmentation and volumetry of multiple sclerosis brain lesions from mr images, NeuroImage: Clinical 8 (2015) 367–375.

[40] N. Shiee, P.-L. Bazin, A. Ozturk, D. S. Reich, P. A. Calabresi, D. L. Pham, A topology-preserving approach to the segmenta-

tion of brain images with multiple sclerosis lesions, NeuroImage 49 (2) (2010) 1524–1535.

[41] P. Schmidt, C. Gaser, M. Arsic, D. Buck, A. Förschler, A. Berthele, M. Hoshi, R. Ilg, V. J. Schmid, C. Zimmer, et al., An automated tool for detection of flair-hyperintense white-matter lesions in multiple sclerosis, Neuroimage 59 (4) (2012) 3774–3783.

[42] M. E. Caligiuri, P. Perrotta, A. Augimeri, F. Rocca, A. Quattrone, A. Cherubini, Automatic detection of white matter hyperintensities in healthy aging and pathology using magnetic resonance imaging: A review, Neuroinformatics 13 (3) (2015) 1–16.

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

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

[45] J. Schmidhuber, Deep learning in neural networks: An overview, Neural Networks 61 (2015) 85–117.

[46] D. H. Hubel, T. N. Wiesel, Receptive fields, binocular interaction and functional architecture in the cat's visual cortex, The Journal of Physiology 160 (1) (1962) 106.

[47] K. Fukushima, Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position, Biological cybernetics 36 (4) (1980) 193–202.

[48] D. Cireşan, U. Meier, J. Masci, J. Schmidhuber, Multi-column deep neural network for traffic sign classification, Neural Networks 32 (2012) 333–338.

[49] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, arXiv preprint arXiv:1502.01852.

[50] Y. Taigman, M. Yang, M. Ranzato, L. Wolf, Deepface: Closing the gap to human-level performance in face verification, in: Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on, 2014, pp. 1701–1708.

[51] D. Ciresan, A. Giusti, L. M. Gambardella, J. Schmidhuber, Deep neural networks segment neuronal membranes in electron microscopy images, in: Advances in Neural Information Processing Systems, 2012, pp. 2843–2851.

[52] D. Cireşan, J. Schmidhuber, Multi-column deep neural networks for offline handwritten chinese character classification, arXiv preprint arXiv:1309.0261.

[53] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

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

[55] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE, 2009, pp. 248–255.

[56] C. Farabet, C. Couprie, L. Najman, Y. LeCun, Learning hierarchical features for scene labeling, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (8) (2013) 1915–1929.

[57] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from rgb-d images for object detection and segmentation, in: Computer Vision–ECCV 2014, Lecture Notes in Computer Science (LNCS 8695), Springer, 2014, pp. 345–360.

[58] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: Computer Vision–ECCV 2014, Lecture Notes in Computer Science (LNCS 8695), Springer, 2014, pp. 297–312.

[59] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, arXiv preprint arXiv:1411.4038.

[60] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, arXiv preprint arXiv:1505.04597.

[61] H. Greenspan, B. van Ginneken, R. M. Summers, Guest editorial deep learning in medical imaging: Overview and future

promise of an exciting new technique, IEEE Transactions on Medical Imaging 35 (5) (2016) 1153–1159.

[62] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, A. Biller, Deep mri brain extraction: a 3d convolutional neural network for skull stripping, NeuroImage 129 (2016) 460–469.

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

[64] P. Moeskops, M. A. Viergever, A. M. Mendrik, L. S. de Vries, M. J. Benders, I. Išgum, Automatic segmentation of mr brain images with a convolutional neural network, IEEE transactions on medical imaging 35 (5) (2016) 1252–1261.

[65] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, et al., Hough-cnn: Deep learning for segmentation of deep brain regions in mri and ultrasound, arXiv preprint arXiv:1601.07014.

[66] H. Chen, Q. Dou, L. Yu, P.-A. Heng, Voxresnet: Deep voxelwise residual networks for volumetric brain segmentation, arXiv preprint arXiv:1608.05895.

[67] D. Nie, L. Wang, Y. Gao, D. Sken, Fully convolutional networks for multi-modality isointense infant brain image segmentation, in: Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on, IEEE, 2016, pp. 1342–1345.

[68] M. Shakeri, S. Tsogkas, E. Ferrante, S. Lippe, S. Kadoury, N. Paragios, I. Kokkinos, Sub-cortical brain structure segmentation using f-cnn's, arXiv preprint arXiv:1602.02130.

[69] S. Pereira, A. Pinto, V. Alves, C. A. Silva, Brain tumor segmentation using convolutional neural networks in mri images, IEEE transactions on medical imaging 35 (5) (2016) 1240–1251.

[70] M. Havaei, A. Davy, D. Warde-Farley, A. Biard, A. Courville, Y. Bengio, C. Pal, P.-M. Jodoin, H. Larochelle, Brain tumor segmentation with deep neural networks, Medical Image Analysis.

[71] M. Havaei, N. Guizard, N. Chapados, Y. Bengio, Hemis: Hetero-modal image segmentation, arXiv preprint arXiv:1607.05194.

[72] L. Zhao, K. Jia, Multiscale cnns for brain tumor segmentation and diagnosis, Computational and mathematical methods in medicine 2016.

[73] Q. Dou, H. Chen, L. Yu, L. Shi, D. Wang, V. C. Mok, P. A. Heng, Automatic cerebral microbleeds detection from mr images via independent subspace analysis based hierarchical features, in: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2015, pp. 7933–7936.

[74] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, P.-A. Heng, Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks, IEEE transactions on medical imaging 35 (5) (2016) 1182–1195.

[75] M. Kamber, R. Shinghal, D. L. Collins, G. S. Francis, A. C. Evans, Model-based 3-d segmentation of multiple sclerosis lesions in magnetic resonance brain images, Medical Imaging, IEEE Transactions on 14 (3) (1995) 442–453.

[76] D. Hervé, J.-F. Mangin, N. Molko, M.-G. Bousser, H. Chabriat, Shape and volume of lacunar infarcts a 3d mri study in cerebral autosomal dominant arteriopathy with subcortical infarcts and leukoencephalopathy, Stroke 36 (11) (2005) 2384–2388.

[77] M. Jenkinson, S. Smith, A global optimisation method for robust affine registration of brain images, Medical Image Analysis 5 (2) (2001) 143–156.

[78] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma, H. Hulshoff Pol, T. Cannon, R. Kawashima, B. Mazoyer, A four-dimensional probabilistic atlas of the human brain, Journal of the American Medical Informatics Association 8 (5) (2001) 401–430.

[79] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, S. M. Smith, Fsl, Neuroimage 62 (2) (2012) 782–790.

[80] S. M. Smith, Fast robust automated brain extraction, Human Brain Mapping 17 (3) (2002) 143–155.

[81] Y. Zhang, M. Brady, S. Smith, Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm, Medical Imaging, IEEE Transactions on 20 (1) (2001) 45–57.

[82] J. Pastor-Pellicer, F. Zamora-Martínez, S. España-Boquera, M. J. Castro-Bleda, F-measure as the error function to train neural networks, in: Advances in Computational Intelligence, Lecture Notes in Computer Science (LNCS 7902), Springer, 2013, pp. 376–384.

[83] D. Scherer, A. Müller, S. Behnke, Evaluation of pooling operations in convolutional architectures for object recognition, in: Artificial Neural Networks–ICANN 2010, Lecture Notes in Computer Science (LNCS 6354), Springer, 2010, pp. 92–101.

[84] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.

[85] Y. N. Dauphin, H. de Vries, J. Chung, Y. Bengio, Rmsprop and equilibrated adaptive learning rates for non-convex optimization, arXiv preprint arXiv:1502.04390.

[86] A. L. Maas, A. Y. Hannun, A. Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML, Vol. 30, 2013.

[87] L. Bottou, Stochastic gradient descent tricks, in: Neural Networks: Tricks of the Trade, Lecture Notes in Computer Science (LNCS 7700), Springer, 2012, pp. 421–436.

[88] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks., in: Aistats, Vol. 9, 2010, pp. 249–256.

[89] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, The Journal of Machine Learning Research 15 (1) (2014) 1929–1958.

[90] T. Kooi, G. Litjens, B. van Ginneken, A. Gubern-Mérida, C. I. Sánchez, R. Mann, A. den Heeten, N. Karssemeijer, Large scale deep learning for computer aided detection of mammographic lesions, Medical Image Analysis 35 (2017) 303–312.