Deep Convolutional Encoder Networks for Multiple Sclerosis Lesion Segmentation

***

[1] ***
[2] ***
[3] ***

Abstract. We propose a novel segmentation approach based on deep convolutional encoder networks and apply it to the segmentation of multiple sclerosis (MS) lesions in magnetic resonance images. Our model is a neural network that has both convolutional and deconvolutional layers, and combines feature extraction and segmentation prediction in a single model. The joint training of the feature extraction and prediction layers allows the model to automatically learn features that are optimized for accuracy for any given combination of image types. In contrast to existing automatic feature learning approaches, which are typically patch-based, our model learns features from entire images, which eliminates patch selection and redundant calculations at the overlap of neighboring patches and thereby speeds up the training. Our network also uses a novel objective function that works well for segmenting underrepresented classes, such as MS lesions. We have evaluated our method on the publicly available labeled cases from the MS lesion segmentation challenge 2008 data set, showing that our method performs comparably to the state-of-the-art. In addition, we have evaluated our method on the images of 500 subjects from an MS clinical trial and varied the number of training samples from 5 to 250 to show that the segmentation performance can be greatly improved by having a representative data set.

Keywords: Segmentation, multiple sclerosis lesions, MRI, machine learning, unbalanced classification, deep learning, convolutional neural nets

1 Introduction

Multiple sclerosis (MS) is an inflammatory and demyelinating disease of the central nervous system, and is characterized by the formation of lesions, primarily visible in the white matter on conventional magnetic resonance images (MRIs). Imaging biomarkers based on the delineation of lesions, such as lesion load and lesion count, have established their importance for assessing disease progression and treatment effect. However, lesions vary greatly in size, shape, intensity and location, which makes their automatic and accurate segmentation challenging. Many automatic methods have been proposed for the segmentation of MS lesions over the last two decades, which can be classified into unsupervised and supervised methods. Unsupervised methods do not require a labeled data set for training. Instead, lesions are identified as an outlier class using, e.g., clustering methods [1] or dictionary learning and sparse coding to model healthy tissue [2]. Current supervised approaches typically start with a large set of features, either predefined by the user [3] or gathered in a feature extraction step, which is followed by a separate training step with labeled data to determine which set of features are the most important for segmentation in the particular domain. For example, Yoo et al. [4] proposed performing unsupervised learning of domain-specific features from image patches from unlabelled data using deep learning. The most closely related methodology to our currently proposed one comes from the domain of cell membrane segmentation, in which Ciresan et al. [5] proposed to classify the centers of image patches directly using a convolutional neural network [6] without a dedicated feature extraction step. Instead, features are learned indirectly within the lower layers of the neural network during training, while the higher layers can be regarded as performing the classification. In contrast to unsupervised feature learning, this approach allows the learning of features that are specifically tuned to the segmentation task. Although deep network-based feature learning methods have shown great potential for image segmentation, the time required to train complex patch-based methods can make the approach infeasible when the size and number of patches are large.

We propose a new method for segmenting MS lesions that processes entire MRI volumes through a neural network with a novel objective function to automatically learn features tuned for lesion segmentation. By processing entire volumes instead of patches, our model removes the need to select representative patches, eliminates redundant calculations where patches overlap, and therefore scales up more efficiently with image resolution. This speeds up training and allows our model to take advantage of large data sets. Our neural network is composed of three layers: an input layer composed of the image voxels of different modalities, a convolutional layer [6] that extracts features from the input layer at each voxel location, and a deconvolutional layer [7] that uses the extracted features to predict a lesion mask and thereby classify each voxel of the image in a single operation. The entire network is trained at the same time, which enables feature learning to be driven by segmentation performance. The proposed network is similar in architecture to a convolutional auto-encoder [8], which produces a lower dimensional encoding of the input images and uses the decoder output to measure the reconstruction error needed for training, while our network uses the decoder to predict lesion masks of the input images. Due to the structural similarity to convolutional auto-encoders, we call our model a convolutional encoder network (CEN). Traditionally, neural networks are trained by back-propagating the sum of squared differences of the predicted and expected outputs. However, if one class is greatly underrepresented, as is the case for lesions, which typically comprise less than 1 % of the image voxels, the algorithm would learn to ignore the minority class completely. To overcome this problem, we propose a new objective function based on a weighted combination of sensitivity and specificity, designed to deal with unbalanced classes and formulated to allow stable gradient computations.

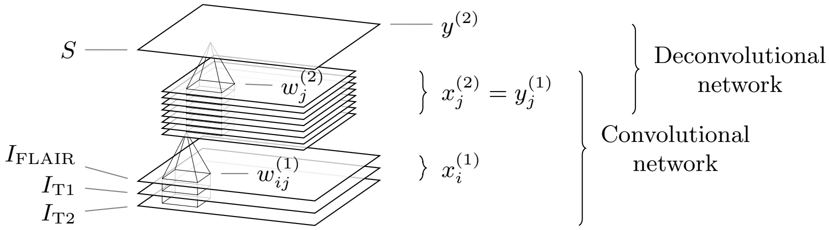

Fig. 1. Convolutional encoder network used to produce a lesion segmentation, S, from multi-modal images, $I = (I_{\mathrm{FLAIR}}, I_{\mathrm{T1}}, I_{\mathrm{T2}})$. The first two layers form a convolutional neural network with trainable filter kernels $w_{ij}^{(1)}$, and the last two layers form a deconvolutional neural network with trainable filter kernels $w_j^{(2)}$.

2 Methods

In this paper, the task of segmenting MS lesions is defined as finding a function s that maps multi-modal images I, e.g., $I = (I_{\mathrm{FLAIR}}, I_{\mathrm{T1}}, I_{\mathrm{T2}})$, to corresponding lesion masks S. Given a set of training images I_n, $n \in \mathbb{N}$, and corresponding segmentations S_n, we model finding an appropriate function for segmenting MS lesions as an optimization problem of the following form

$$\hat{s} = \arg\min_{s \in \mathcal{S}} \sum_n E(S_n, s(I_n)). \tag{1}$$

where $\mathcal{S}$ is the set of possible segmentation functions, and E is an error measure that calculates the dissimilarity between ground truth segmentations and predicted segmentations.

The set of possible segmentation functions is modeled by the convolutional encoder network illustrated in Fig. 1. Our network consists of three layers: an input layer, a convolutional layer, and a deconvolutional layer. The input layer is composed of the image voxels $x_i^{(1)}(\boldsymbol{p})$, $i \in [1, C], C \in \mathbb{N}$, where i indexes the modality, C is the number of modalities, and $\boldsymbol{p} \in \mathbb{R}^3$ are the coordinates of a particular voxel. The convolutional layer automatically learns features from the input images. It is a deterministic function of the following form

$$y_j^{(1)} = \max\left(0, \sum_{i=1}^{C} \tilde{w}_{ij}^{(1)} * x_i^{(1)} + b_j^{(1)}\right) \tag{2}$$

where $y_j^{(1)}, j \in [1, F], F \in \mathbb{N}$, denotes the feature map corresponding to the trainable convolution filter $w_{ij}^{(1)}$, F is the number of filters, b_j is a trainable bias term, $*$ denotes valid convolution, and $\tilde{w}$ denotes a flipped version of w. The deconvolutional layer uses the extracted features to calculate a probabilistic lesion mask as follows

$$y^{(2)} = \mathrm{sigm}\left(\sum_{j=1}^{F} w_j^{(2)} \circledast x_j^{(2)} + b^{(2)}\right) \tag{3}$$

where $x_j^{(2)} = y_j^{(1)}$, $w_j^{(2)}$ and $b^{(2)}$ are trainable parameters, $\circledast$ denotes full convolution, and $\operatorname{sigm}(z)$ denotes the sigmoid function defined as $\operatorname{sigm}(z) = (1 + \exp(-z))^{-1}, z \in \mathbb{R}$. To obtain a binary lesion mask from the probabilistic output of our model, we chose a fixed threshold such that the mean Dice similarity coefficient is maximized on the training set.

The parameters of the model can be efficiently learned by minimizing the error E on the training set using stochastic gradient descent [6]. Typically, neural networks are trained by minimizing the sum of squared differences (SSD)

$$E = \frac{1}{2} \sum_{\boldsymbol{p}} \Big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p}) \Big)^2 . \tag{4}$$

The partial derivatives of the error with respect to the model parameters can be calculated using the delta rule and are given by

$$\frac{\partial E}{\partial w_j^{(2)}} = \delta^{(2)} * \tilde{x}_j^{(2)}, \qquad \frac{\partial E}{\partial b^{(2)}} = \frac{1}{N^3} \sum_{\boldsymbol{p}} \delta^{(2)}(\boldsymbol{p}) \tag{5}$$

with

$$\delta^{(2)} = \big(y^{(2)} - S \big) y^{(2)} \big(1 - y^{(2)} \big) \tag{6}$$

where N^3 is the number of voxels of a single input channel. The derivatives of the error with respect to the first layer parameters can be calculated by applying the chain rule of partial derivatives and is given by

$$\frac{\partial E}{\partial w_{ij}^{(1)}} = x_i^{(1)} * \tilde{\delta}_j^{(1)}, \qquad \frac{\partial E}{\partial b_j^{(1)}} = \frac{1}{M^3} \sum_{\boldsymbol{q}} \delta_j^{(1)}(\boldsymbol{q}) \tag{7}$$

with

$$\delta_j^{(1)} = \big(w_j^{(2)} \circledast \delta^{(2)} \big) \mathbb{I} \big(y_j^{(1)} > 0 \big) \tag{8}$$

where M^3 is the number of voxels of a feature map, $\boldsymbol{q} \in \mathbb{R}^3$, and $\mathbb{I}(z)$ denotes the indicator function defined as 1 if the predicate z is true and 0 otherwise.

The sum of squared differences is a good measure of classification accuracy, if the two classes are fairly balanced. However, if one class contains vastly more samples, as is the case for lesion segmentation, the error measure is dominated by the majority class and consequently, the neural network would learn to completely ignore the minority class. To overcome this problem, we use a combination of sensitivity and specificity, which can be used together to measure classification performance even for vastly unbalanced problems. More precisely, the final error measure is a weighted sum of the mean squared difference of the lesion voxels (sensitivity) and non-lesion voxels (specificity), reformulated to be error terms:

$$E = r \frac{\sum_{\boldsymbol{p}} \big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p}) \big)^2 S(\boldsymbol{p})}{\sum_{\boldsymbol{p}} S(\boldsymbol{p})} + (1-r) \frac{\sum_{\boldsymbol{p}} \big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p}) \big)^2 \big(1 - S(\boldsymbol{p}) \big)}{\sum_{\boldsymbol{p}} \big(1 - S(\boldsymbol{p}) \big)} \tag{9}$$

where the first term captures the squared sensitivity error and the second term captures the squared specificity error. We formulate the sensitivity and specificity errors as squared errors in order to yield smooth gradients, which makes the optimization more robust. The sensitivity ratio r can be used to assign different weights to the two terms. Due to the large number of non-lesion voxels, weighting the specificity error higher is important, but the algorithm is stable with respect to changes in r, which largely affects the threshold used to binarize the probabilistic output. In all our experiments, a sensitivity ratio between 0.10 and 0.01 yields very similar results.

To train our model, we must compute the derivatives of the modified objective function with respect to the model parameters. Equations (5), (7), and (8) are a consequence of the chain rule of derivatives and independent of the chosen similarity measure. Hence, we only need to derive the update rule for $\delta^{(2)}$. With $\alpha = 2r(\sum_{\boldsymbol{p}} S(\boldsymbol{p}))^{-1}$ and $\beta = 2(1-r)(\sum_{\boldsymbol{p}}(1 - S(\boldsymbol{p})))^{-1}$ we can rewrite E as

$$E = \frac{1}{2}\sum_{\boldsymbol{p}} \Big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p})\Big)^2 \alpha S(\boldsymbol{p}) + \frac{1}{2}\sum_{\boldsymbol{p}} \Big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p})\Big)^2 \beta\big(1 - S(\boldsymbol{p})\big) \tag{10}$$

$$= \frac{1}{2}\sum_{\boldsymbol{p}} \big(\alpha S(\boldsymbol{p}) + \beta(1 - S(\boldsymbol{p}))\big) \Big(S(\boldsymbol{p}) - y^{(2)}(\boldsymbol{p})\Big)^2 \tag{11}$$

Our objective function is similar to the SSD, with an additional multiplicative term applied to the squared differences. The additional factor is constant with respect to the model parameters. Consequently, $\delta^{(2)}$ can be derived analogously to the SSD case, and the new factor is simply carried over:

$$\delta^{(2)} = \big(\alpha S + \beta(1 - S)\big)\big(y^{(2)} - S\big)y^{(2)}\big(1 - y^{(2)}\big) \tag{12}$$

3 Experiments and Results

To allow for a direct comparison with state-of-the-art lesion segmentation methods, we evaluated our method on the FLAIR, T1-, and T2-weighted MRIs of the 20 publicly available labeled cases from the MS lesion segmentation challenge 2008 [9], which we downsampled from the original isotropic voxel size of $0.5\,\text{mm}^3$ to an isotropic voxel size of $1.0\,\text{mm}^3$. In addition, we evaluated our method on an in-house data set from an MS clinical trial of 500 subjects split equally into training and test sets. The images were acquired from 45 different scanning sites. For each subject, the data set contains T2- and PD-weighted MRIs with a voxel size of $0.937\,\text{mm} \times 0.937\,\text{mm} \times 3.000\,\text{mm}$. The main preprocessing steps included rigid intra-subject registration, brain extraction, intensity normalization, and background cropping. We used a CEN with 32 filters and filter sizes of $9 \times 9 \times 9$ and $9 \times 9 \times 5$ voxels for the challenge and in-house data sets, respectively. Training on a single GeForce GTX 780 graphics card took between 6 and 32 hours per model depending on the training set size. However, once the network is trained, segmentation of trimodal 3D volumes with a resolution of, e.g., $159 \times 201 \times 155$ voxels can be performed in less than one second. As a rough[1]

[1] The comparison is imprecise due to differing experimental conditions.

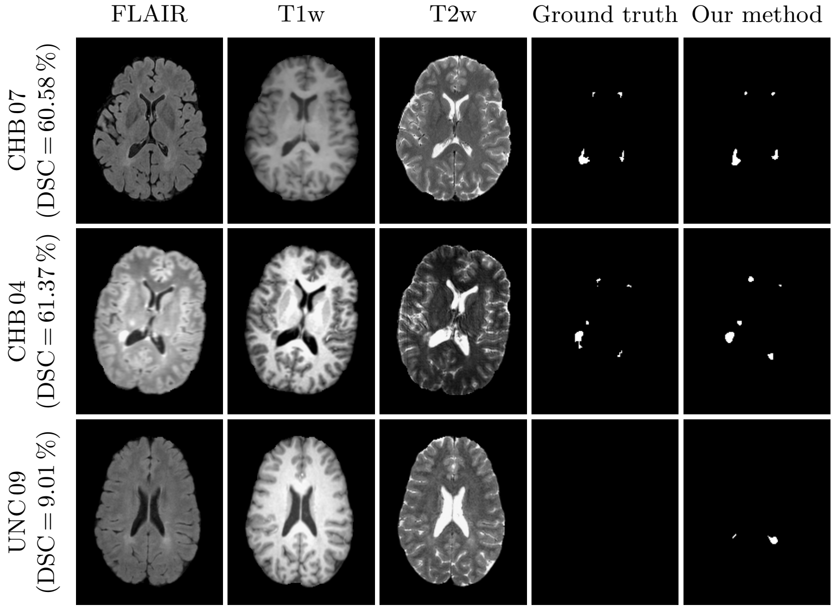

Fig. 2. Example segmentations of our method for three different subjects from the challenge data set. Our method performed well and consistently despite the large contrast differences seen between the first two rows. In the third row, our method also segmented lesions that have similar contrast, but these regions had not been identified as lesions by the manual rater, which highlights the difficulty in distinguishing focal lesions from diffuse damage, even for experts.

comparison, Ciresan et al. [5] reported that their patch-based method required 10 to 30 minutes to segment a single 2D image with a resolution of 512 × 512 voxels using four graphics cards, which demonstrates the large speed-ups gained by processing entire volumes.

We evaluated our method on the challenge data set using 5-fold cross-validation and calculated the true positive rate (TPR), positive predictive value (PPV), and Dice similarity coefficient (DSC) between the predicted segmentations and the resampled ground truth. Figure 2 shows a comparison of three subjects from the challenge data set. The first two rows show the FLAIR, T1w, T2w, ground truth segmentations, and predicted segmentations of two subjects with a DSC of 60.58 % and 61.37 %. Despite the large contrast differences between the two subjects, our method performed well and consistently, which indicates that our model was able to learn features that are robust to a large range of intensity variations. The last row shows a subject with a DSC of 9.01 %, one of the lowest DSC scores from the data set. Our method segmented lesions that have similar contrast to the other two subjects, but these regions were not classified as lesions by the manual rater. This highlights the difficulty of manual lesion segmentation, as the difference between diffuse white matter pathology and focal lesions is

Table 1. Comparison of our method with state-of-the-art lesion segmentation methods in terms of mean TPR, PPV, and DSC. Our method performs comparably to the best methods reported on the MS lesion segmentation challenge data set.

Method	TPR	PPV	DSC
Souplet et al. [1]	20.65	30.00	—
Weiss et al. [2]	33.00	36.85	29.05
Geremia et al. [3]	39.85	40.35	—
Our method	39.71	41.38	35.52

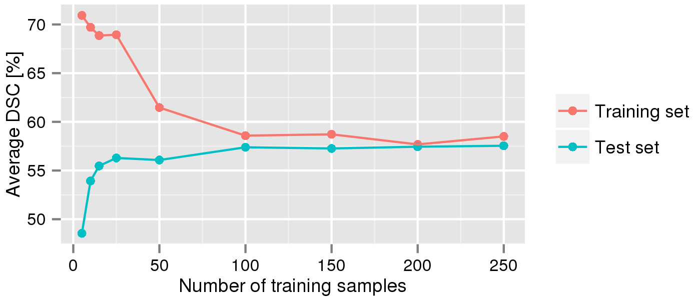

Fig. 3. Comparison of DSC scores calculated on the training and test sets for varying numbers of training samples. At around 100 samples, the model becomes stable in terms of test performance and the small difference between training and test DSCs, indicating that overfitting of the training data no longer occurs.

often indistinct. A quantitative comparison of our method with other state-of-the-art methods is summarized in Table 1. Our method outperforms the winning method (Souplet et al. [1]) of the MS lesion segmentation challenge 2008 and the currently best unsupervised method reported on that data set (Weiss et al. [2]) in terms of mean TPR and PPV. Our method performs comparably to a current method (Geremia et al. [3]) that uses a carefully designed set of features specifically designed for lesion segmentation, despite our method having learned its features solely from a relatively small training set.

To evaluate the impact of the training set size on the segmentation performance, we trained our model on our in-house data set with a varying number of training samples and calculated the mean DSC on the training and test sets as illustrated in Fig. 3. For small training sets, there is a large difference between the DSCs on the training and test sets, which indicates that the training set is too small to learn a representative set of features. At around 100 samples, the model becomes stable in terms of test performance and the small difference between training and test DSCs, indicating that overfitting of the training data is no longer occurring. With 100 training subjects, our method achieves a mean DSC on the test set of 57.38 %, which shows that the segmentation accuracy can be greatly improved compared to the results on the challenge data set, when a representative training set is available.

4 Conclusions

We have introduced a new method for the automatic segmentation of MS lesions based on convolutional encoder networks. The joint training of the feature extraction and prediction layers with a novel objective function allows for the automatic learning of features that are tuned for a given combination of image types and a segmentation task with very unbalanced classes. We have evaluated our method on two data sets showing that approximately 100 images are required to train the model without overfitting but even when only a relatively small training set is available, the method still performs comparably to the state-of-the-art algorithms. For future work, we plan to increase the depth of the network, which would allow the learning of a set of hierarchical features. This could further improve segmentation accuracy, but may require larger training sets. We would also like to investigate the use of different objective functions for training based on other measures of segmentation performance.

Acknowledgements ****
****

References

1. Souplet, J.C., Lebrun, C., Ayache, N., Malandain, G.: An automatic segmentation of T2-FLAIR multiple sclerosis lesions. In: The MIDAS Journal-MS Lesion Segmentation (MICCAI 2008 Workshop). (2008)
2. Weiss, N., Rueckert, D., Rao, A.: Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer (2013) 735–742
3. Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer (2010) 111–118
4. Yoo, Y., Brosch, T., Traboulsee, A., Li, D.K., Tam, R.: Deep learning of image features from unlabeled data for multiple sclerosis lesion segmentation. In: Machine Learning in Medical Imaging. Springer (2014) 117–124
5. Ciresan, D., Giusti, A., Schmidhuber, J.: Deep neural networks segment neuronal membranes in electron microscopy images. Advances in Neural Information Processing Systems (2012) 1–9
6. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
7. Zeiler, M.D., Taylor, G.W., Fergus, R.: Adaptive deconvolutional networks for mid and high level feature learning. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 2018–2025
8. Masci, J., Meier, U., Cireşan, D., Schmidhuber, J.: Stacked convolutional auto-encoders for hierarchical feature extraction. In: Artificial Neural Networks and Machine Learning–ICANN 2011. Springer (2011) 52–59
9. Styner, M., Lee, J., Chin, B., Chin, M., Commowick, O., Tran, H., Markovic-Plese, S., Jewells, V., Warfield, S.: 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. MIDAS Journal **2008** (2008) 1–6