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 (MRIs). Our model is a neural network that is both convolutional and deconvolutional, 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 and application. 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 reduces redundant calculations at the overlap of neighboring patches and thereby speeds up the training. 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 500 images (split equally into training and test sets) from a data set from an MS clinical trial, showing that the segmentation performance can be greatly improved by having a representative training set.

Keywords: Multiple sclerosis lesions, segmentation, 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 with pathology that can be observed in vivo by magnetic resonance imaging (MRI). MS is characterized by the formation of lesions, primarily visible in the white matter on conventional MRI. Imaging biomarkers based on the delineation of lesions such as lesion load and lesion count, have established their importance to assess disease progression and treatment affect. However, lesions vary greatly in 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 [1], which can be classified into supervised and unsupervised methods. Unsupervised methods do not require a labeled data set for training. Instead, lesions can be modelled as an outlier of a generative model of image patches [2], or as an outlier class using clustering methods [3]. Current supervised approaches typically start with a large set of features, either predefined by the user 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. While early approaches have used the intensity values from different modalities of a particular voxel as the input features [4], these simple features are sensitive to intensity variations caused by, e.g., different scanners or MRI acquisition protocols. Geremia et. al [5] have shown that carefully chosen context-rich features are more robust to intensity variations, which improves segmentation accuracy. Instead of using a large set of hand-crafted features, Youngjin et. al [6] proposed to learn domain specific features from image patches from an unlabelled data set using unsupervised feature learning. For the automatic segmentation of cell membranes, Ciresan et. al proposed to classify the center of image patches directly using a convolutional neural network without a dedicated feature extraction step [7]. 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 supervised and unsupervised feature learning methods have shown great potential for image segmentation, the time required to train complex patch-based feature extraction methods can make the approach infeasible when the size and the number of patches is large. Ciresan et. al have reported a training time of more than a week to train their patch-based segmentation model using 4 GPUs on 2D images with a resolution of 512×512 pixels [7]. To scale patch-based classification to 3D images with a resolution of $256 \times 256 \times 50$, Youngjin et. al used only a small fraction ($0.1\,\%$) of the possible patches for training, which might limit the ability to learn features that are representative of the entire image.

In this paper, we propose a novel method for segmenting MS lesions that can automatically learn features tuned for lesion segmentation and scales better to large data sets of high-resolution 3D images than previous patch-based feature learning approaches, which allows our model to take advantage of large data sets. Our model is a neural network that is composed of three layers: an input layer composed of the image voxels for different modalities, a convolutional layer [8] that extracts features from the input layer at each voxel location, and a deconvolutional layer [9] that uses the extracted features from the second layer to classify each voxel of the image in a single operation. Both layers are trained at the same time, which facilitates the learning of features that are tuned for lesion segmentation. A key difference to the network of Ciresan et. al [7] is that our model is trained on the entire images instead of multiple patches from the same image, which eliminates redundant calculations at the overlap of neighboring patches and thereby speeds up the training and also eliminates the need to select representative patches. This allows our model to be trained on large data sets in

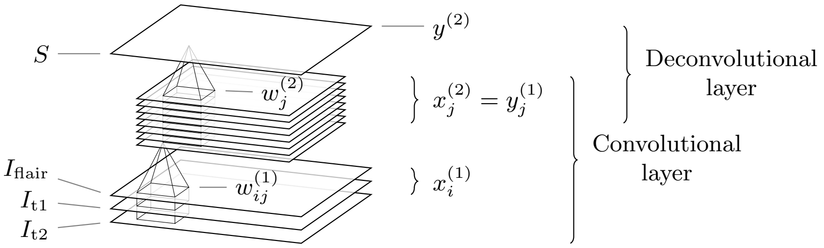

Fig. 1. Convolutional encoder network used to segment MS lesion of images from the MICCAI 2008 lesion segmentation challenge. The first two layers form a convolutional neural network, and the last two layers form a deconvolutional neural network.

order to learn features that cover the broad spectrum of lesion variability. The proposed network is similar in architecture to a convolutional auto-encoder [10] but instead of learning a lower dimensional representation of the input images themselves, the output of your network is the predicted lesion mask. Due to the structural similarity to convolutional auto-encoders, we will call our model a convolutional encoder network (CEN). Traditionally, neural networks are trained by back-propagating the sum of squared differences (SSD) of the predicted and the expected output. However, if one class is much overrepresented, as is the case for lesion segmentation, the algorithm would learn to ignore the minority class completely. To overcome this problem, we propose to use the weighted sum of sensitivity and specificity error as a new objective function for neural networks, which is suitable to deal with very unbalanced classification problems, and we will derive the gradients of our proposed objective function in order train the model using stochastic gradient descent.

needs tuning

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., $I = (I_{\text{flair}}, I_{\text{t1}}, I_{\text{t2}})$, 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

need a better way to write this

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

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 Figure 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 extracts automatically learned 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 jth feature, F is the number of features, w_{ij} and b_j are trainable parameters of the model, $*$ denotes valid convolution, and $\tilde{w}_{ij}$ denotes a flipped version of w_{ij}. The deconvolutional layer uses the extracted features to calculate a probabilistic lesion mask as follows

This equation is almost identical to equation (1) in [9]. I've just added the sigmoid function because our output is binary.

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

where $x_j^{(2)} = y_j^{(1)}$, $w_j^{(2)}$ and $b^{(2)}$ are trainable parameters, $\circledast$ denotes full convolution, and $\operatorname{sigm}(x)$ 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 threshold such that the average 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 (SGD) [8]. Typically, neural networks are trained by minimizing the sum of squared differences (SSD)

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

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

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

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

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

where M^3 is the number of voxels of a feature map and $\mathbb{I}(z)$ denotes the indicator function, which is 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 than the other class, 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 are two measures that are suitable to measure classification performance even for vastly unbalanced classification problems. More precisely, we calculate the mean squared difference for lesion and non-lesion voxels separately and then calculate the weighted sum of the two terms to form the final error measure

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

where the first term captures the squared sensitivity error and the second term captures the squared specificity error. We formulate the sensitivity and specificity error as a squared error 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 higher than the sensitivity is preferable. We found that the sensitivity ratio mostly affects the optimal lesion threshold, but has only a minor impact on the actual segmentation quality. On all our experiments, a sensitivity ratio between 0.1 and 0.01 yields very similar results.

To train our model, we have to derive 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 $\alpha = 2r_{\text{sen}}(\sum_{\boldsymbol{p}} S(\boldsymbol{p}))^{-1}$ and $\beta = 2(1 - r_{\text{sen}})(\sum_{\boldsymbol{p}}(1 - S(\boldsymbol{p})))^{-1}$ we can rewrite E as

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

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

Our objective function is similar to the SSD, with the difference that an additional term is multiplied to the squared differences. The additional factor does not depend on $y^{(2)}$ and is therefore constant with respect to the model parameters. Consequently, $\delta^{(2)}$ can be derived analogously to the SSD case, where the new factor is carried over:

needs tuning

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

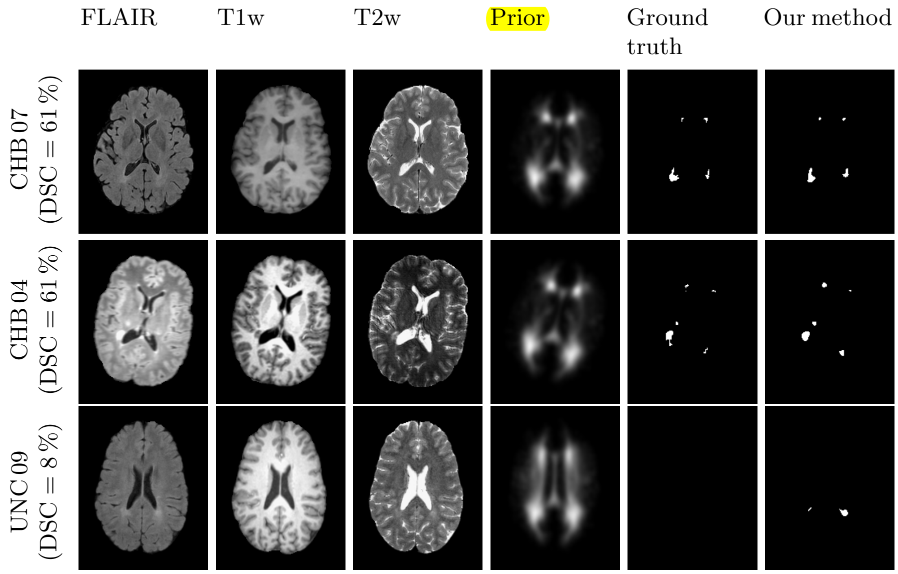

Fig. 2. Example segmentation using our method. Comment why the method performed poorly. Maybe it just was a very difficult case. CHB07, FLAIR, T1, T2, prior, ground truth, predicted segmentation. CHB04, UNC09. Robust to different contrast, but missclassifies diffusely abnormal white matter as MS lesions.

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 [11] ~~after resampling them to a~~ voxel size of 1 mm^3. In addition, we evaluated our method on an in-house dataset from an MS clinical trial of 500 subjects split equally into training and test set. For each subject, the data set contains T2- and PD-weighted MRIs with a voxel size of 1 mm × 1 mm × 3 mm. The main preprocessing steps included rigid intra-subject registration, brain extraction, intensity normalization, and background cropping. To ~~account for the spatial distribution of lesions,~~ we ~~aligned~~ the test subjects of our in-house data set to MNI space ~~using affine registration~~ and calculated the average lesion mask. We used the square root of the average lesion mask as a ~~lesion~~ prior, to counterbalance large differences in lesion probability and added ~~the aligned~~ lesion prior as an additional input channel to both data sets. ~~For the experiments on the challenge and in-house data sets, we used a CEN with 32 filters and a filter size of 9 × 9 × 9 and 9 × 9 × 5 voxels, respectively.~~

We evaluated our method on the challenge data set using 5-fold cross-validation and calculated the true positive rate (TPR), the positive predictive value (PPV) and the ~~dice~~ similarity coefficient (DSC) between the predicted segmentation and the resampled ground truth. Figure 2 shows a comparison of three subjects

Table 1. Comparison of our method with state-of-the-art lesion segmentation methods. Our method performs comparable to the current state of the art, despite learning the features solely from a relatively small training set.

Method	TPR	PPV	DSC
Souplet et al. [3]	20.65	30.00	—
Weiss et al. [2]	33.00	36.85	29.05
Geremia et al. [5]	39.85	40.35	—
Our method	39.71	41.38	35.52

from the challenge data set. The first two rows show FLAIR, T1w, T2w, lesion prior, ground truth segmentation and predicted segmentation of the two subjects with the highest DSC (DSC = 61 %). Despite the large contrast differences between the two subjects, our method ~~is able to segment MS lesions with high accuracy~~, which indicates that our model was able to learn features that are robust to a large range of intensity variations. The last row shows ~~the~~ subject with a DSC of 9 %, one of the lowest DSC from the data set. ~~Our method picks up parts as lesions which are not classified as focal lesions by the rater, which results a much higher than average number of false positives.~~ A 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. [3]) of the MS Lesion segmentation challenge 2008 ~~in terms of average TPR and PPV~~ and the currently best unsupervised method on that data set (Weiss et al. [2]). Our method performs ~~comparable~~ to ~~the current state-of-the-art~~ method that uses a carefully designed set of features specifically tuned for lesion segmentation, despite ~~learning~~ the features solely from a relatively small training set.

Kind of forgot what we really had to say. Just remember that we wanted to say that the data leaves much to interpretation which hurts the evaluation

To evaluate the impact of the training set size on the segmentation performance, we ~~have~~ trained our model on our in-house data set with varying number of training samples and calculated the ~~average~~ DSC on the training and test set as illustrated in Fig. 3. For small ~~numbers of training samples~~, there is a large difference between the DSC on the training and test set, which indicates that the training set ~~size~~ is ~~to~~ small to learn a representative set of features. At around 100 images, the model becomes stable ~~which is indicated by the fast that the test DSC does not improve significantly and there is also not overfitting occurring~~. ~~At around~~ 100 training subjects, our method achieves a DSC on the test set of 58 %, which shows that the segmentation accuracy can be greatly improved compared to the results on the challenge data set, when a representative data set is available.

4 Conclusions

We have introduced a new method for the automatic segmentation of MS lesions based on convolutional auto encoders. The joint training of the feature extraction and prediction layers allows for the automatic learning of features

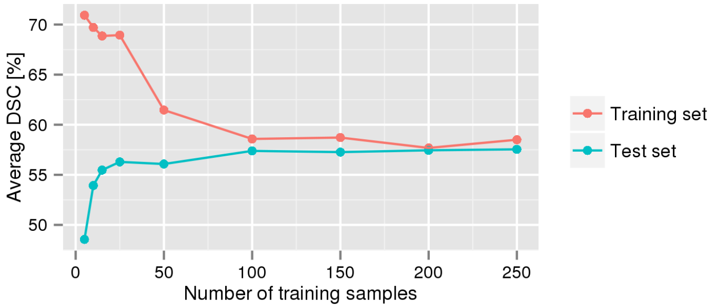

Fig. 3. Comparison of DSC from the training and test set for varying number of training samples. The model ~~is overfitting if only a few training cases are given,~~ indicated by the discrepancy between DSC on the training and test set. The method becomes stable at roughly 100 training samples.

that are tuned for a given combination of image types and ~~lesion segmentation~~. We have evaluated our method on two data sets showing that approximately 100 images are required to train the model without overfitting ~~and~~ that the method performs ~~comparable to the state of the art~~, even when only a relatively small data set is used for training. For future work, we ~~want~~ to investigate ~~the training of deeper networks~~, which would allow the learning of a set of hierarchical features, ~~which could potentially~~ improve segmentation accuracy, but ~~might~~ require larger training sets ~~to be trained without overfitting~~. We would also like to investigate the use of different objective functions for training ~~like, e.g., the DSC or a combination of TPR and PPV~~.

Acknowledgements ****
****

References

1. García-Lorenzo, D., Francis, S., Narayanan, S., Arnold, D.L., Collins, D.L.: Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Medical image analysis **17**(1) (2013) 1–18
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. 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)
4. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in mr images: method and validation. Medical Imaging, IEEE Transactions on **13**(4) (1994) 716–724

5. 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
6. 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
7. 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
8. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11) (1998) 2278–2324
9. 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
10. 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
11. 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