NeuroImage: Clinical 21 (2019) 101638

Contents lists available at ScienceDirect

NeuroImage: Clinical

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

One-shot domain adaptation in multiple sclerosis lesion segmentation using convolutional neural networks

Sergi Valverde[a,*], Mostafa Salem[a,b], Mariano Cabezas[a], Deborah Pareto[c], Joan C. Vilanova[d], Lluís Ramió-Torrentà[e], Àlex Rovira[c], Joaquim Salvi[a], Arnau Oliver[a], Xavier Lladó[a]

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

ARTICLE INFO

Keywords:
Brain
MRI
Multiple sclerosis
Automatic lesion segmentation
Convolutional neural networks

ABSTRACT

In recent years, several convolutional neural network (CNN) methods have been proposed for the automated white matter lesion segmentation of multiple sclerosis (MS) patient images, due to their superior performance compared with those of other state-of-the-art methods. However, the accuracies of CNN methods tend to decrease significantly when evaluated on different image domains compared with those used for training, which demonstrates the lack of adaptability of CNNs to unseen imaging data. In this study, we analyzed the effect of intensity domain adaptation on our recently proposed CNN-based MS lesion segmentation method. Given a source model trained on two public MS datasets, we investigated the transferability of the CNN model when applied to other MRI scanners and protocols, evaluating the minimum number of annotated images needed from the new domain and the minimum number of layers needed to re-train to obtain comparable accuracy. Our analysis comprised MS patient data from both a clinical center and the public ISBI2015 challenge database, which permitted us to compare the domain adaptation capability of our model to that of other state-of-the-art methods. In both datasets, our results showed the effectiveness of the proposed model in adapting previously acquired knowledge to new image domains, even when a reduced number of training samples was available in the target dataset. For the ISBI2015 challenge, our one-shot domain adaptation model trained using only a single case showed a performance similar to that of other CNN methods that were fully trained using the entire available training set, yielding a comparable human expert rater performance. We believe that our experiments will encourage the MS community to incorporate its use in different clinical settings with reduced amounts of annotated data. This approach could be meaningful not only in terms of the accuracy in delineating MS lesions but also in the related reductions in time and economic costs derived from manual lesion labeling.

1. Introduction

Currently, magnetic resonance imaging (MRI) is extensively used in the diagnosis and monitoring of multiple sclerosis (MS), due to the sensitivity of structural MRI to disseminate focal white matter (WM) lesions in time and space (Rovira et al., 2015). With different modifications of MRI criteria over time, the presence of new lesions on MRI scans is considered a prognostic and predictive biomarker for the disease (Filippi et al., 2016). Although visual lesion inspection is feasible in practice, this task is time-consuming, prone to manual errors and variable for different expert raters, which has lead to the development of a wide number of automated strategies in recent years (Lladó et al., 2012).

Although there is a wide range of methods proposed, convolutional neural network (CNN) strategies are being increasingly introduced. In contrast to previously supervised learning methods, CNNs do not require manual feature engineering or prior guidance, which along with the increase in computing power makes them a very interesting alternative for automated lesion segmentation, as seen by their top ranking performance on all of the international MS lesion challenges (Styner et al., 2008; Carass et al., 2017; Commowick et al., 2018). The proposed network architectures and training pipelines include three-dimensional (3D) encoder networks with shortcut connections (Brosch et al., 2016), multi-view image architectures (Birenbaum and Greenspan, 2017), cascaded 3D pipelines (Valverde et al., 2017), multi-dimensional

* Corresponding author.
E-mail address: svalverde@eia.udg.edu (S. Valverde).

https://doi.org/10.1016/j.nicl.2018.101638
Received 25 June 2018; Received in revised form 30 November 2018; Accepted 9 December 2018
Available online 10 December 2018

recurrent gated units (Andermatt et al., 2017) and fully convolutional architectures (Roy et al., 2018; Hashemi et al., 2018).

However, CNN architectures applied in MRI tend to not generalize well on unseen image domains, which is mostly due to variations in image acquisition, MRI scanner, contrast, noise level or resolution between image datasets. As a result, manual expert labeling must be performed on the new image domain, which is very-time consuming and not always possible. In this aspect, only a few papers have analyzed the CNN domain adaptation problem on brain MRI. Recently, Kamnitsas et al. (2017) proposed an unsupervised domain adaptation CNN model for the segmentation of traumatic brain injuries, where adversarial training was applied to adapt two related image domains with distinct types of image sequences. Similarly, Ghafoorian et al. (2017) investigated the transferability of the acquired knowledge of a CNN model that was initially trained for WM hyper-intensity segmentation on legacy low-resolution data when applied to new data from the same scanner but with higher image resolution, showing the minimum amount of supervision required in terms of high-resolution training samples and re-trained network layers. Nevertheless, to the best of our knowledge, still any study has been focused on the domain adaptation between completely unrelated MS image domains in terms of the image acquisition (scanner), resolution and contrast, which can be very interesting in evaluating the usability of these CNN models in different clinical scenarios.

In this paper, we analyzed the effectiveness of supervised image domain adaptation between completely unrelated MS databases. To do so, we first trained a slightly modified version of our already proposed cascaded architecture (Valverde et al., 2017) entirely using two public MS databases from the Medical Image Computing and Computer Assisted Intervention (MICCAI) society, MICCAI2008 (Styner et al., 2008) and MICCAI2016 (Commowick et al., 2018), which was considered the source model. Then, we analyzed the transferring knowledge capability of this model by evaluating its performance on a set of completely unseen images from other target image domains, partly re-training a different number of layers or no layers. We extended our analysis to investigate the minimum number of unseen images and re-trained layers needed to obtain a similar performance on the domain adapted model, even in one-shot domain scenarios in which only a single training case was available on the target domain. Our evaluation included a clinical dataset and public MS data from the International Symposium on Biomedical Imaging (ISBI) 2015 MS challenge (Carass et al., 2017), comparing the performance of the domain-adapted CNN model with those of the same model fully trained on the target domain and other state-of-the-art methods. To promote the reproducibility and usability of our research, the proposed domain adaptation methodology is available as part of our *nicMSlesions* MS lesion software, which can be downloaded freely from our research website.[1]

2. Materials and methods

2.1. CNN architecture

The CNN MS lesion model follows our recently proposed framework for MS lesion segmentation (Valverde et al., 2017). Within this framework, a cascade of two identical CNNs are optimized, where the first network is trained to be more sensitive to revealing possible candidate lesion voxels, while the second network is trained to reduce the number of false positive outcomes. For a complete description of the details and motivations for the proposed architecture, please refer to the original publication.

The architecture by Valverde et al. (2017) was composed of two stacks of convolution and max-pooling layers with 32 and 64 filters, respectively. The convolutional layers were followed by a fully connected (FC) layer of 256 in size and a softmax FC layer, summing ~200 K parameters. Here, to accommodate more expressive features that arise from the baseline training, we propose to double the number of layers on each convolutional stack (see Fig. 1). Additionally, we also stack two additional FC layers of size 128 and 64, to increase the number of potentially retrained classification layers used to adapt the image domains. The resulting CNN architecture consists of ~470 K network parameters.

The CNN training and inference procedures are identical to those proposed by Valverde et al. (2017). Briefly, training is performed following a two-step approach: first, a CNN model is trained using a balanced set of multi-channel Fluid attenuation inversion recovery (FLAIR) and T1-weighted (T1-w) 3D $11 \times 11 \times 11$ patches extracted from all of the available lesion voxels and a random selection of normal appearing tissue voxels. Then, the error of the first CNN model with the respect to the available training annotations is computed. Finally, the second model is trained using again a balanced set of voxels composed of all of the lesion voxels and a random selection of the misclassified lesion voxels on the previous model. Afterward, inferencing on the unseen images is performed by evaluating all of the input voxels using the first trained CNN, which discards all of the voxels with a low probability of being lesion. The remaining voxels are re-evaluated using the second CNN, obtaining a lesion probabilistic lesion mask. Final binary output masks are computed by linear thresholding of probabilities $\geq t_{bin}$ and a posterior filtering of the resulting binary regions with a lesion size below l_{min}.

2.2. Initial training

The proposed CNN architecture was first fully trained using 35 images from the two publicly available MS lesion segmentation datasets of the MICCAI society. Both the MICCAI2008 (Styner et al., 2008) and MICCAI2016 (Commowick et al., 2018) are currently used as benchmarks to compare the accuracy of novel MS lesion segmentation pipelines. Please note that for each individual challenge, the proposed network architecture performed in the top rank (see Valverde et al. (2017) for the final ranking and comparison with other state-of-the-art methods).

2.2.1. MICCAI 2008 dataset

The MICCAI 2008 MS lesion segmentation challenge was composed of 20 training scans from research subjects, which were acquired at Children's Hospital Boston (CHB, 3 T Siemens) and University of North Carolina (UNC, 3 T Siemens Alegra). For each subject, the original T1-w, T2 weighted (T2-w) and FLAIR image modalities were provided with voxel size = $0.5 \times 0.5 \times 0.5\ mm^3$. The provided FLAIR and T2-w image modalities were already rigidly co-registered to the T1-w space. All of the subjects were provided with manual expert annotations of WM lesions from a CHB and UNC expert rater. As pointed out by Styner et al. (2008), the UNC manual annotations were adapted to closely match those from CHB, and thus, only the CHB annotations were used.

As a previous step, we skull-stripped both the T1-w and FLAIR images using the Brain Extraction Tool (BET) (Smith et al., 2002) and bias corrected using N3 (Sled et al., 1998). In order to keep the same voxel spacing between all the experiment datasets used in the paper, all the training images were then interpolated to ($1 \times 1 \times 1\ mm^3$) using the FSL-FLIRT utility (Greve and Fischl, 2009).

2.2.2. MICCAI 2016 dataset

The MICCAI 2016 MS lesion segmentation challenge (Commowick et al., 2018) was composed of 15 training scans acquired in three different scanner vendors: 5 scans (Philips Ingenia 3 T), 5 scans (Siemens Aera 1.5 T) and 5 scans (Siemens Verio 3 T). For each subject, 3D T1-w MPRAGE, 3D FLAIR, 3D T1-w gadolinium enhanced and 2D T2-w/Proton Density (PD) images were provided, with voxel sizes ranging from ($0.74 \times 0.74 \times 0.7\ mm^3$) to ($0.72 \times 0.72 \times 4.0 mm^3$). Please refer

[1] http://github.com/NIC-VICOROB/nicmslesions.

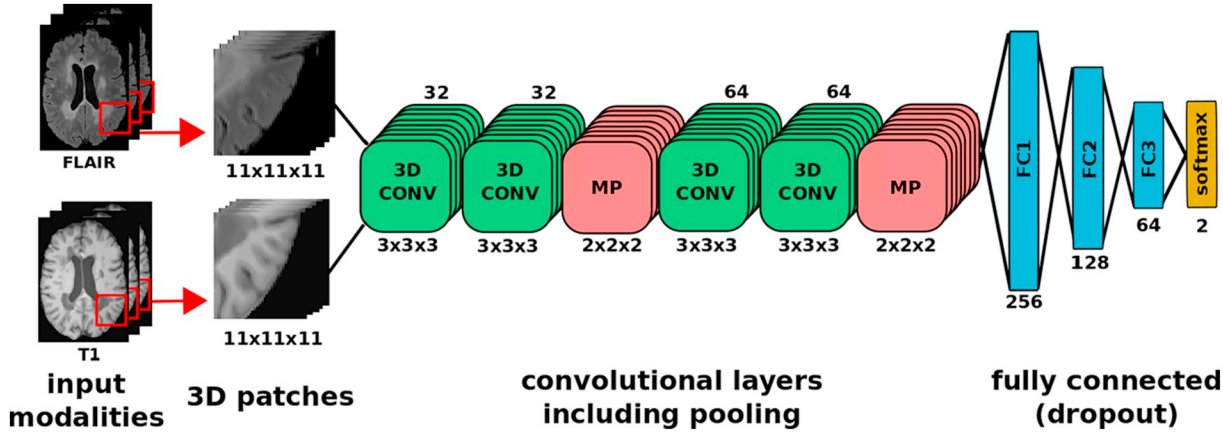

Fig. 1. Eleven-layer CNN model architecture trained using multi-sequence 3D image patches (FLAIR and T1-w) that are $11 \times 11 \times 11$ in size. Compared to the original implementation in Valverde et al. (2017), we double the number of convolutional layers (3DCONV) before each of the two max-pooling layers (MP) and we add two additional fully connected layers of sizes 128 (FC2) and 64 (FC3), before the softmax layer.

to the original publication for more details for the exact details of the acquisition parameters and image resolutions (Commowick et al., 2018). Manual lesion annotations for each training subject were provided as a consensus mask among 7 different human raters.

Pre-processed images were already provided. The pre-processing pipeline consisted of a denoising step with the NL-means algorithm (Coupé et al., 2008) and a rigid registration (Commowick et al., 2012) of all of the modalities against the FLAIR image. Then, each of the modalities were skull-stripped using the volBrain platform (Manjón and Coupé, 2016) and bias corrected using the N4 algorithm (Tustison et al., 2010). Finally, all the training images were also interpolated to ($1 \times 1 \times 1$ mm^3) using the FSL-FLIRT utility (Greve and Fischl, 2009) in order to match the same voxel spacing between all the experiment datasets.

2.2.3. Experiment details

All of the training images were first normalized with a zero mean and a standard deviation of one. The normalized images were used to build a set of 1,200,000 training patches, where 25% was selected for validation and the others were used to optimize the network's weights. We trained each of the two networks for 400 epochs with an early stopping of 50 epochs for each network. This technique permits to prevent over-fitting by stopping training after a number of 50 epochs without a decrease in the validation error. The parametric rectified linear activation function (PReLU) (He et al., 2015) was applied to all layers. The convolutional layers were regularized using batch normalization (Ioffe and Ioffe and Szegedy, 2015), while dropout (Srivastava et al., 2014) was applied to each of the FC layers with ($p = 0.5$). Network optimization was performed using the adaptive learning rate method (ADADELTA) (Zeiler, 2012) with a batch size of 128 and categorical cross-entropy as the loss cost. The post-processing parameters $\geq t_{bin}$ and l_{min} were set to 0.5 and 10, respectively.

2.3. Supervised domain adaptation

Although the convolutional layers can encode domain independent valid image features that describe the location, shape and lesion contrast, these features are then propagated through the FC layers, which learn to classify the lesion voxels based on the training data. However, this process is inherently dependent on the training domain characteristics, such as the intensity ratio between the lesion and the normal appearing tissue, which enables the FC layers to learn to optimize the best correlation between the extracted convolutional layers and the manual labels.

However, the encoded knowledge already present in the source model can be effectively used to adapt it to an unseen target intensity domain because convolutional layers contain related features that can be transferred to unseen data while only re-training the FC layers (see Fig. 2). In our experiments, domain adaptation is performed by re-training all or some of the source FC layers using images from the target domain. Table 1 shows the number of network parameters used in each of the proposed configurations. As a result of reusing part of the implicit knowledge trained on the source model, the number of weights to optimize on the target model is significantly lower, which permits us to train the model with a reduced number of training images without over-fitting the model.

2.4. Implementation

All of the experiments were run on a GNU/Linux machine box running Ubuntu 16.04, with 32GB of RAM memory. CNN training was conducted on a single NVIDIA TITAN-X GPU (NVIDIA Corp, United States) with 12GB of RAM memory. All of the procedures were implemented in the Python language,[2] using the Keras[3] and Tensorflow[4] libraries. The proposed method was integrated as part of our MS lesion segmentation software *nicMSlesions*, which is available for downloading at our research website[1].

3. Experiments

3.1. Clinical MS dataset

3.1.1. Data

A total of 60 patients with a clinically isolated syndrome (Hospital Vall d'Hebron, Barcelona, Spain) were scanned on a 3 T Siemens with a 12-channel phased-array head coil (Trio Tim, Siemens, Germany) with the following acquired sequences: 1) transverse PD and T2-w fast spin-echo (TR = 2500 ms, TE = 16–91 ms, voxel size = $0.78 \times 0.78 \times 3$ mm^3), 2) transverse fast T2-FLAIR (TR = 9000 ms, TE = 93 ms, TI = 2500 ms, flip angle = 120°, voxel size = $0.49 \times 0.49 \times 3$ mm^3), and 3) sagittal 3D T1 magnetization prepared rapid gradient-echo (MPRAGE) (TR = 2300 ms, TE = 2 ms, TI = 900 ms, flip angle = 9°; voxel size = $1 \times 1 \times 1.2$ mm^3). For each patient, WM lesion masks were semi-automatically delineated from either DP or FLAIR masks using JIM software[5] by an expert radiologist of the same hospital center. The T1-w and FLAIR images were first

[2] https://www.python.org/.
[3] https://keras.io.
[4] https://tensorflow.org.
[5] Xinapse Systems, http://www.xinapse.com/home.php.

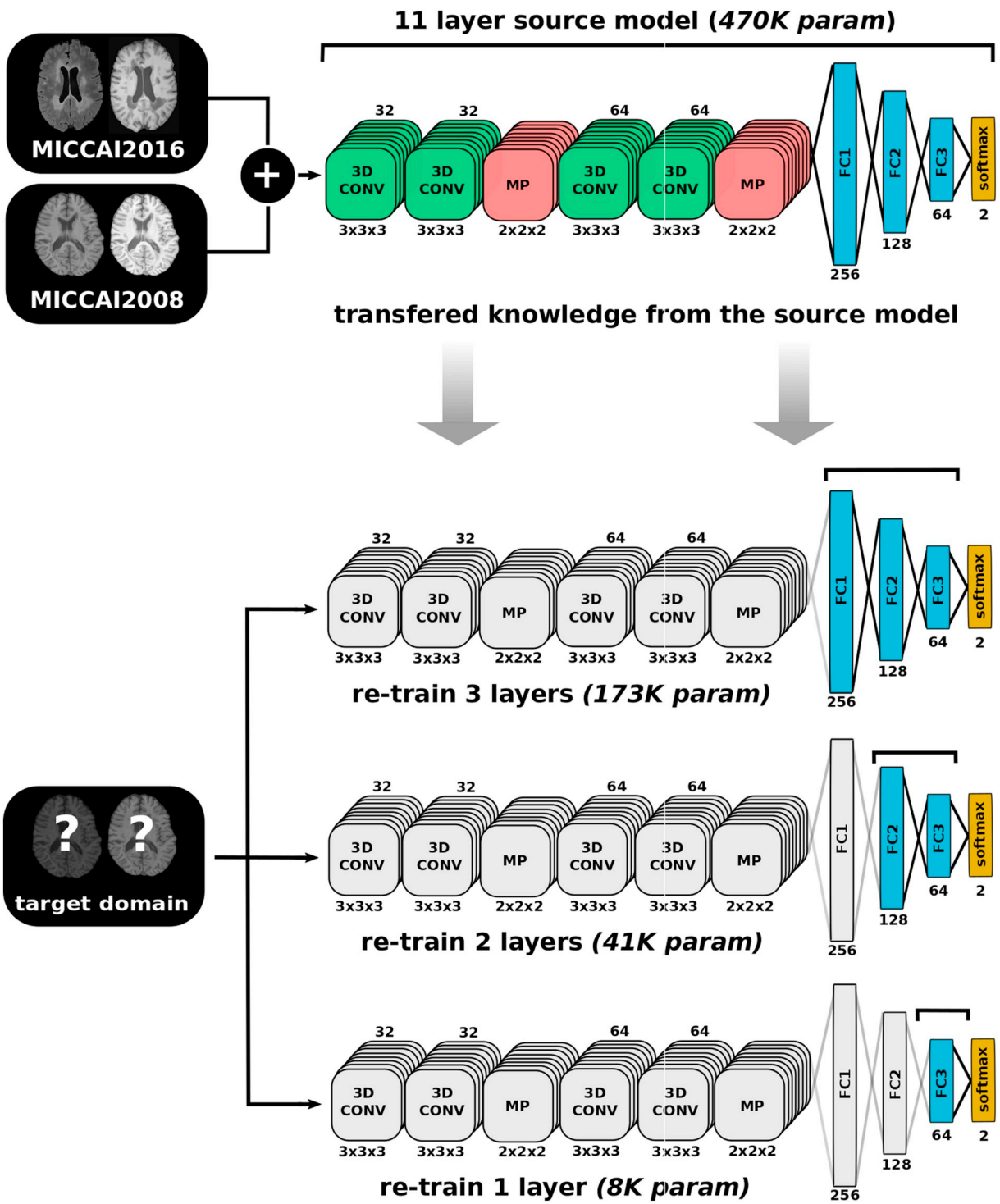

Fig. 2. Supervised intensity domain adaptation framework. From the 11 layer CNN source model trained on two public MS datasets (see Subsection 2.2), we transfer the model knowledge to an unseen target image domain. Domain adaptation is performed via 3 possible configurations by retraining the first FC layer, two FC layers or all FC layers using images and labels from the target intensity domain. In all of the configurations, the layers that are not re-trained are depicted in gray.

skull-stripped using BET (Smith et al., 2002) and bias corrected using N3 (Sled et al., 1998). The FLAIR images were affinely co-registered to the T1-w space using the FSL-FLIRT utility (Greve and Fischl, 2009).

3.1.2. Evaluation

The images were first randomly split into two sets composed of 30 training and testing images. Then, the training data were used to train the different target models while accounting for the following factors:

- The effect of one-shot domain adaptation training. Each proposed domain adaptation configuration was trained using a single training case with a lesion size in the range of [0.5–18] ml.
- The effect of the proposed domain adaptation configurations on the accuracy of the target model (retraining 1, 2 or all of the FC layers, see Table 1).
- The effect of the number of training images used to re-train the target model. Each proposed domain adaptation configuration was trained using 1, 2, 5, 10, 15 or all of the available training images.

Table 1
Training parameters on each of the CNN models used. When training the source model (see Subsection 2.2), all of the network layers are optimized from scratch. On the target models, only the last FC layer (FC3), last two FC layers (F2 + FC3) or all FC layers (FC1 + FC2 + FC3) are optimized, which significantly reduces the number of training parameters.

Model	Trained layers	Network param
Source	All (11 layers)	470,466
Target 3 layers	FC1 + FC2 + FC3	172,928
Target 2 layers	FC2 + FC3	41,344
Target 1 layer	FC3	8320

After training, each of the target models was evaluated on the test set, evaluating the accuracy of the resulting segmentations against the available lesion annotations using the following evaluation metrics:

- The overall % segmentation accuracy in terms of the dice similarity coefficient (*DSC*) between the manual lesion annotations and the output segmentation masks:

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

where TP_s and FP_s denote the number of voxels correctly and incorrectly classified as a lesion, respectively, and *FN* denotes the number of voxels incorrectly classified as a non-lesion.

- Sensitivity of the method in detecting lesions between manual lesion annotations and output segmentation masks, expressed in %:

$$\text{sensitivity} = \frac{TP_d}{TP_d + FN_d} \times 100$$

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

- Precision of the method in detecting lesions between manual lesion annotations and output segmentation masks, also expressed in %:

$$\text{precision} = \frac{TP_d}{TP_d + FP_d} \times 100$$

where TP_d and FP_d denote the number of correctly and incorrectly classified lesion region candidates, respectively.

To evaluate the effectiveness of the proposed framework, the obtained results were compared against the source model without retraining and the same target model fully trained using all of the available training images. For comparison, the segmentation accuracies of two state-of-the-art MS lesion segmentation pipelines LST (Schmidt et al., 2012) and SLS (Roura et al., 2015), were also reported.

3.1.3. Experiment details

All of the training images were first normalized with a zero mean and standard deviation of one. Each of the trained models was run with the exact parameters used to train the source model (see Subsection 2.2.3). The number of lesion voxels was equal during all of the training epochs. Normal appearing tissue voxels were re-sampled every 10 epochs to augment the tissue variability during the training. As in the source model, the post-processing parameters $\geq t_{bin}$ and l_{min} were set to 0.5 and 10, respectively. In the LST, the parameters κ and *lgm* were optimized for the current dataset with the values $\kappa = 0.15$ and $lgm = gm$, respectively. In the SLS, the parameters α, λ_{ts} and λ_{ns} were also optimized for this particular dataset with the values $\alpha = 3$, $\lambda_{ts} = 0.6$ and $\lambda_{nb} = 0.6$ for both iterations.

3.1.4. Results

First, we evaluated the models under a one-shot domain adaptation scenario, by training them again several times using only a single case from the training set with lesion burdens equal to 0.5, 1.2, 3.1, 8.3 and 18 ml. Table 2 shows the DSC, sensitivity and precision coefficients of each of the re-trained models under different one-shot training sets. The same evaluation is also shown for LST, SLS, and the cascaded CNN architecture without fine-tunning (source) and fully trained using the entire training dataset. As expected, the model without domain adaptation reported the worst accuracy by the lack of adaptability of the source knowledge. In contrast, the models performance increased with the number of annotated lesions on the target domain, showing better DSC with the manual annotations than LST and SLS, even in extreme cases in which only 9 lesions are manually annotated on the target domain (0.5 ml).

Table 2
Clinical MS dataset: DSC, sensitivity and precision coefficients for each of the models re-trained using a single case with varying degree of lesion load. For comparison, the obtained values for SLS (Roura et al., 2015), LST (Schmidt et al., 2012) and the same cascaded CNN method fully trained using the 30 available training cases (Valverde et al., 2017) are also shown. For each coefficient, the reported values are the mean (standard deviation) when evaluated on the 30 testing cases.

llesion vol (num lesions)	DSC	Sensitivity	Precision
1 layer (FC3)			
0.5 ml (9 lesions)	0.30 (0.19)	0.44 (0.23)	0.49 (0.30)
1.2 ml (11 lesions)	0.39 (0.19)	0.44 (0.19)	0.67 (0.23)
3.1 ml (17 lesions)	0.38 (0.22)	0.46 (0.20)	0.54 (0.25)
8.3 ml (90 lesions)	0.44 (0.17)	0.58 (0.19)	0.58 (0.26)
18 ml (78 lesions)	0.47 (0.18)	0.59 (0.18)	0.58 (0.23)
2 layers (FC2 + FC3)			
0.5 ml (9 lesions)	0.30 (0.17)	0.52 (0.23)	0.54 (0.28)
1.2 ml (11 lesions)	0.39 (0.18)	0.49 (0.21)	0.72 (0.29)
3.1 ml (17 lesions)	0.36 (0.22)	0.42 (0.20)	0.54 (0.27)
8.3 ml (90 lesions)	0.45 (0.15)	0.55 (0.18)	0.66 (0.24)
18 ml (78 lesions)	0.44 (0.19)	0.62 (0.20)	0.52 (0.25)
3 layers (FC1 + FC2 + FC3)			
0.5 ml (9 lesions)	0.28 (0.17)	0.48 (0.22)	0.48 (0.28)
1.2 ml (11 lesions)	0.38 (0.17)	0.52 (0.22)	0.72 (0.26)
3.1 ml (17 lesions)	0.38 (0.21)	0.46 (0.21)	0.55 (0.25)
8.3 ml (90 lesions)	0.44 (0.17)	0.61 (0.17)	0.57 (0.26)
18 ml (78 lesions)	0.45 (0.18)	0.60 (0.21)	0.55 (0.23)
Source (0 lesions)	0.23 (0.22)	0.42 (0.43)	0.45 (0.34)
SLS	0.25 (0.17)	0.34 (0.25)	0.51 (0.30)
LST	0.28 (0.23)	0.31 (0.21)	0.59 (0.27)
CNN	0.53 (0.16)	0.60 (0.21)	0.75 (0.21)

As a second experiment, we evaluated the effect of adding more training data on the accuracy of the domain adapted models. Fig. 3 shows the DSC, sensitivity and precision coefficients of each of the re-trained models using different number of training image patients which ranged from 1 to 30. The number of training samples was $\sim 18K$, $\sim 36k$, $\sim 48k$, $\sim 60K$, $\sim 70K$, $\sim 95K$ and $\sim 130K$ for 1, 2, 5, 10, 15, 20 and 30 images, respectively. When more training data on the target space were available, the performances of the re-trained models were similar to that of the fully trained CNN pipeline, especially those of the models in which the last two or all of the FC layers were re-trained. In contrast, in the sensitivity and precision plots, the re-trained models were in general more sensitive to inferring WM lesions but at the cost of increasing also the number of false-positive outcomes.

3.2. ISBI 2015 dataset

3.2.1. Data

The ISBI2015 MS lesion challenge (Carass et al., 2017) was composed of 5 training and 14 testing subjects with 4 or 5 different image time-points per subject. All of the data were acquired on a 3.0 Tesla MRI scanner (Philips Medical Systems, Best, The Netherlands) with T1-w MPRAGE (voxel size = $0.82 \times 0.82 \times 1.17 mm^3$), FLAIR, T2-w and PD (voxel size = $0.82 \times 0.82 \times 2.2 mm^3$), DP and FLAIR sequences. A complete description of the image protocol and pre-processing details is

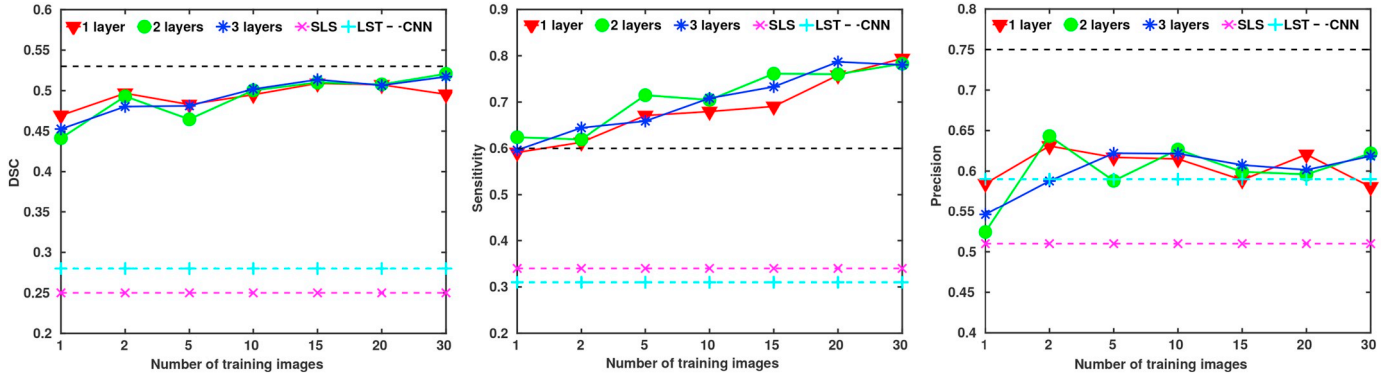

Fig. 3. Effect of the number of re-trained FC layers and training images on the DSC, sensitivity and precision coefficients when evaluated on the clinical MS dataset. The represented value for each configuration is computed as the mean DSC, sensitivity and precision scores over the 30 testing images. For comparison, the obtained values for the lesion segmentation methods SLS (Roura et al., 2015) (× pink line), LST (Schmidt et al., 2012) (+ cyan line) and the same cascaded CNN method fully trained using all of the available training data (Valverde et al., 2017) (− black line) are shown.

available on the organizer's website.[6] On the challenge competition, each subject image was evaluated independently, which led to a final training and testing sets composed of 21 and 61 images, respectively. Additionally, manual delineations of MS lesions performed by two experts were included for each of the 21 training images.

The evaluation of the ISBI 2015 challenge is performed blind for the teams by submitting the segmentation masks of the 61 testing cases to the challenge website evaluation platform.[7] Different metrics are computed as part of an overall performance score (Carass et al., 2017), where values above 90 are considered to be comparable to human performance.

3.2.2. Evaluation

Here, we analyzed the effect of one-shot domain adaptation on the overall performance of the testing set. To do so, we retrained all of the model configurations (1, 2 or all FC layers) with the first training case from each training subject, which led to 5 different training sets with varying number of lesions and a total lesion volume in the range [2.3–26.8 ml]. Then, each of the resulting trained models was evaluated on the blind test set. Based on that approach, we evaluated the following experiments:

- The effect of the number lesions and lesion volume on the performance of each of the one-shot domain adaptation models. We considered the segmentation masks of the same cascaded architecture fully trained using the 21 training images (Carass et al., 2017) as silver mask annotations, given that this particular model already reported human-like accuracy (score 91.44) when submitted to the challenge platform. We evaluated the performance of each of the one-shot models again while computing the DSC, sensitivity and precision coefficients between the one-shot segmentation masks and the silver masks.
- The performance of the best one-shot domain adaptation model on the blind test set. The best performing model from the previous experiment was sent to the challenge's evaluation platform, comparing its accuracy to those of the other submitted MS lesion segmentation pipelines fully trained using the entire available training set. Among the set of evaluated coefficients computed in the challenge, only the DSC, sensitivity and precision metrics are shown for comparison.

3.2.3. Experimental details

Like in the clinical MS dataset, all of the training images were first normalized with a zero mean and a standard deviation of one. Each of the trained models was run with the exact parameters used to train the source model (see Subsection 2.2.3). The number of lesion voxels was equal during all of the training epochs. Normal appearing tissue voxels were re-sampled every 10 epochs to augment the tissue variability during the training. The post-processing parameters $\geq t_{bin}$ and l_{min} were set also to 0.5 and 10, respectively.

3.2.4. Results

Table 3 shows the performance of each of the one-shot domain adaptation models when trained on different images with varying degrees of lesion size. For comparison, the results for the source model without re-training on the target domain are also depicted. The performance of the source model pre-trained only on the MICCAI2008 and MICCAI2016 datasets shows the lack of accuracy of the method in delineating WM lesions on the unseen target domain. Following the same pattern seen on the clinical MS dataset, the best performance with respect to the silver masks was obtained when re-training all of the FC layers with the maximum number of available voxels (ISBI02, 26.8 ml.). Interestingly, the performance of the model re-trained using just 26 lesions (ISBI03, 5.9 ml.) was remarkably higher than that of the other trained models, especially when only the last two or one FC layers were re-trained. Fig. 4 depicts the effect of the available number of lesion voxels on the resulting number of true-positive, false-positive and false-negative outcomes when re-training only the last FC layer.

Table 4 depicts the performance of the best domain adaptation model (ISBI02 with 3 re-trained layers) against different top rank participant challenge strategies. From the list of compared methods, the best five strategies were based on CNN models (Andermatt et al., 2017; Hashemi et al., 2018; Valverde et al., 2017; Birenbaum and Greenspan, 2017; Roy et al., 2018), while the others were based on either other supervised learning techniques (Valcarcel et al., 2018; Deshpande et al., 2015; Sudre et al., 2015) or unsupervised intensity models (Shiee et al., 2010; Jain et al., 2015). The accuracy of the one-shot domain model was similar to those of other recently fully trained submitted CNN models (Roy et al., 2018), yielding a performance that was comparable to human performance (score 90.3), even when trained it with a single training case. Furthermore, the proposed one-shot method reported a performance similar to that of the same fully trained cascaded CNN architecture (score 91.44) (Valverde et al., 2017), which shows the capability of the model to adapt the source knowledge into the target domain using a reduced training dataset.

4. Discussion

In this paper, we have studied the effect of intensity domain adaptation on our recently published CNN-based MS lesion segmentation

[6] http://iacl.ece.jhu.edu/index.php/MSChallenge/data.
[7] https://smart-stats-tools.org/node/26.

Table 3
ISBI dataset: DSC, sensitivity and precision coefficients for each of the models re-trained using a single case of the training dataset against the silver masks. For comparison, the obtained values for the same source CNN method without domain adaptation (see Subsection 2.2) are also shown. For each coefficient, the reported values are the mean (standard deviation) when evaluated on the 61 testing images.

lesion vol (num lesions)	DSC	Sensitivity	Precision
1 layer (FC3)			
ISBI01 (17.4 ml, 29 lesions)	0.56 (0.14)	0.80 (0.11)	0.62 (0.07)
ISBI02 (26.8 ml, 45 lesions)	0.51 (0.21)	0.83 (0.13)	0.55 (0.07)
ISBI03 (5.9 ml, 26 lesions)	0.65 (0.11)	0.60 (0.17)	0.80 (0.14)
ISBI04 (2.3 ml, 20 lesions)	0.33 (0.12)	0.41 (0.16)	0.81 (0.14)
ISBI05 (4.3 ml, 22 lesions)	0.54 (0.11)	0.56 (0.16)	0.84 (0.12)
2 layers (FC2 + FC3)			
ISBI01 (17.4 ml, 29 lesions)	0.56 (0.14)	0.74 (0.11)	0.59 (0.06)
ISBI02 (26.8 ml, 45 lesions)	0.53 (0.21)	0.87 (0.11)	0.56 (0.06)
ISBI03 (5.9 ml, 26 lesions)	0.65 (0.11)	0.66 (0.15)	0.79 (0.13)
ISBI04 (2.3 ml, 20 lesions)	0.47 (0.12)	0.48 (0.18)	0.83 (0.11)
ISBI05 (4.3 ml, 22 lesions)	0.56 (0.11)	0.54 (0.16)	0.82 (0.13)
3 layers (FC1 + FC2 + FC3)			
ISBI01 (17.4 ml,29 lesions)	0.66 (0.10)	0.73 (0.11)	0.78 (0.10)
ISBI02 (26.8 ml,45 lesions)	0.69 (0.13)	0.70 (0.18)	0.77 (0.10)
ISBI03 (5.9 ml, 26 lesions)	0.65 (0.11)	0.63 (0.13)	0.79 (0.14)
ISBI04 (2.3 ml, 20 lesions)	0.47 (0.14)	0.40 (0.16)	0.84 (0.08)
ISBI05 (4.3 ml, 22 lesions)	0.46 (0.12)	0.46 (0.17)	0.87 (0.13)
Source (0 lesions)	0.33 (0.12)	0.40 (0.16)	0.72 (0.14)

Table 4
ISBI challenge: DSC, sensitivity, precision and overall score coefficients for the best one-shot domain adaptation model (ISBI02 with 3 layers) after submitting the segmentation masks for blind evaluation. The obtained results are compared with different top rank participant strategies and the same model fully trained on all the available data. For each method, the reported values are extracted from the challenge results board. The reported values are the mean (standard deviation) when evaluated on the 61 testing images. The performance of the methods with an overall score ≥ 90 is considered to be similar to human performance.

Method	DSC	Sensitivity	Precision	Score
Andermatt et al. (2017)	0.63 (0.14)	0.54 (0.19)	0.84 (0.10)	92.07
Hashemi et al. (2018)	0.66 (0.11)	0.67 (0.20)	0.71 (0.16)	91.52
Valverde et al. (2017)	0.64 (0.12)	0.57 (0.17)	0.79 (0.15)	91.44
Birenbaum and Greenspan (2017)	0.63 (0.14)	0.55 (0.18)	0.80 (0.15)	91.26
Roy et al. (2018)[a]	0.52 (− −)	- - (− −)	0.86 (− −)	90.48
Deshpande et al. (2015)	0.60 (0.13)	0.55 (0.17)	0.73 (0.18)	89.81
Jain et al. (2015)	0.55 (0.14)	0.47 (0.15)	0.73 (0.20)	88.74
Shiee et al. (2010)	0.55 (0.19)	0.54 (0.15)	0.70 (0.29)	88.46
Valcarcel et al. (2018)	0.57 (0.13)	0.57 (0.18)	0.61 (0.16)	87.71
Sudre et al. (2015)	0.52 (0.14)	0.46 (0.15)	0.66 (0.18)	86.44
Full train	0.63 (0.13)	0.55 (0.16)	0.79 (0.14)	91.33
One-shot (3 layers, 26.8 ml.)	0.58 (0.16)	0.48 (0.19)	0.84 (0.13)	90.32

[a] Obtained results for Roy et al. (2018) were extracted from the related publication.

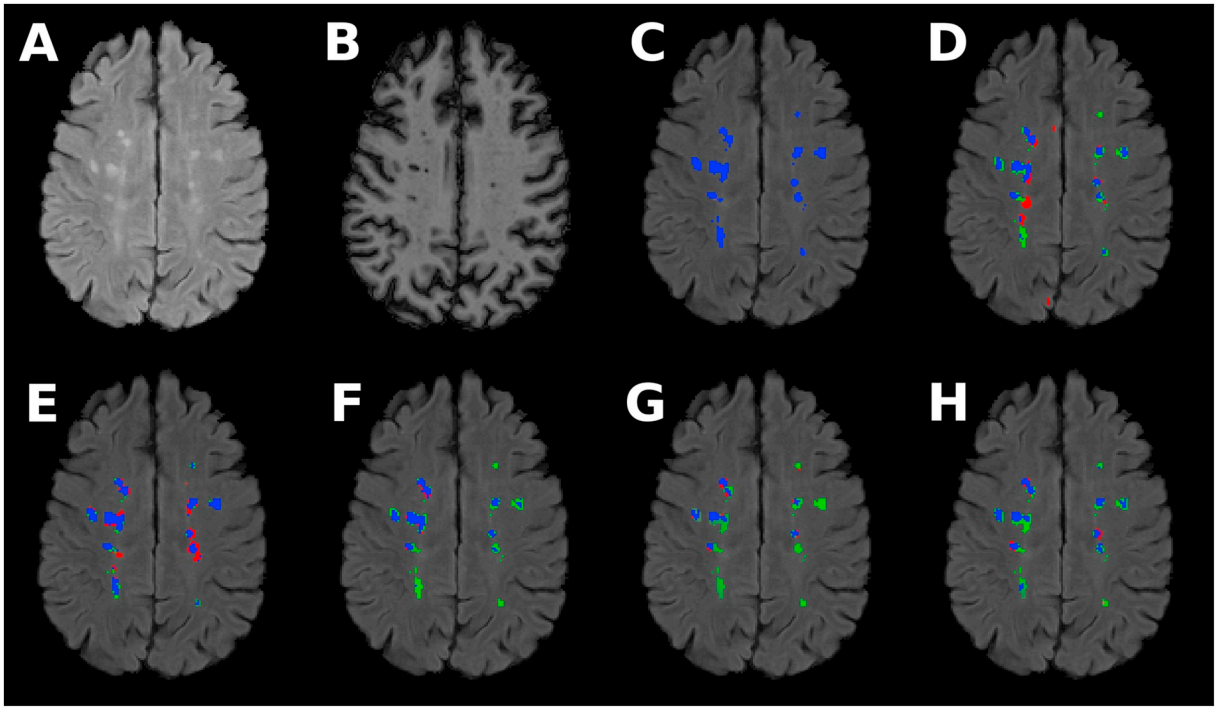

Fig. 4. Output segmentation masks for the first image of the ISBI testing set. (A) FLAIR and (B) T1-w input masks. Silver mask (C) obtained based on the same CNN method fully trained on the entire training dataset (Valverde et al., 2017). The other panels show the output masks for the one-shot domain adaptation model re-trained only for the last FC layer using the images (D) ISBI01 (17.4 ml), (E) ISBI02 (26.8 ml), (F) ISBI03 (5.9 ml), (G) ISBI04 (2.3 ml), and (H) ISBI05 (4.3 ml). The blue regions depict the overlapped lesion voxels between the silver mask and each of the models. The red and green regions depict false-positive and false-negative lesion voxels, respectively, with respect to the silver masks.

method. The model was fully trained on two public MS lesion datasets (MICCAI2008, MICCAI2016), analyzing its capability to transfer the acquired knowledge to two completely unrelated datasets. For this particular architecture, we evaluated the number of necessary layers that must be retrained and the minimum number of annotated images from the unseen domain that is required to obtain a similar fully trained performance. Our results highlighted the effectiveness of the proposed domain adaptation model in transferring previously acquired knowledge to new image domains even if only a single training case was available on the target dataset. Furthermore, in some scenarios such as the ISBI2015 dataset, the performance of these one-shot models was similar to fully trained models and comparable to human rate performance.

Our experiments on the ISBI dataset show a similar score between the original architecture presented in Valverde et al. (2017) and the

proposed method when both were fully trained on the 21 training images, suggesting a similar performance when enough training data is available. However, compared to the original architecture, the proposed model doubles the number of network parameters in order to incorporate more expressive features and potential retrained layers, increasing the probability of overfitting if the model is fully trained on small datasets. In this regard, the performance of the models in which only the FC layers were re-trained were very similar to that of the same model fully trained for both the convolutional and FC layers. This result suggests that there is an inherent capability of the convolutional layers to encode useful image features that can be used across different image domains without re-adaptation. As shown in Table 1, by re-using some of the network layers we drastically reduce the number of parameters to optimize on the target domain, and thus, the domain-adapted networks can be fitted using a small number of training samples without overfitting the model.

Our experiments highlight the relationship between the number of available lesion samples used to re-train the model and the resulting accuracy. As seen in the first experiment, the incorporation of additional training samples increases the segmentation DSC coefficient on all of re-trained models. Domain adaptation was progressively more effective with increasing training cases, since the additional characteristics of the target dataset could be fine-tuned on the FC layers. As expected, the number of false-positive lesion voxels was reduced also with the addition of more lesion examples with contextual information of the target dataset. More interestingly, the models still yielded a remarkably high performance on reduced training sets, such as a single training case. In some cases, one-shot models trained with extremely low lesion load showed a similar or better accuracy than those with higher lesion volume, suggesting that lesion location may be also an important factor. Related to that, one-shot models tended in general to perform better on images with higher number of lesions but not necessarily higher lesion load, which suggests that the addition of different lesion locations may help to increase the variability of the target patches extracted.

On the clinical MS dataset, the performances of the one-shot adapted models were significantly higher than those of the LST and SLS, even when trained using a single case with a 3.1 ml. lesion load and 17 manual annotated regions. Although the SLS and LST methods were unsupervised models that did not require strict training, their parameters were optimized for the target image domain using a time consuming grid-search. In the ISBI2015 challenge, the same cascaded CNN model fully trained on the 21 training images performed in the top rank (4th position/46 participants), yielding comparable human-like accuracy. When compared with this fully trained model, the accuracy of the one-shot domain-adapted model trained with only one of the 21 training images was still remarkably higher than those of most of the participant strategies, which was very similar to other CNN methods and still yielded a comparable human-like accuracy. This finding is relevant, and it shows the potential applicability of our cascaded CNN method on very reduced datasets with a limited loss in the accuracy.

In general, none of the hyper-parameters optimized for the source model were fine-tuned on any of the domain-adapted models, which kept them fixed along of all the experiments conducted in this study. As previously observed, for a training dataset that contained at least 3000 lesion voxels (3 ml. on a isotropic $1mm^3$), the best results were obtained when the last two or all of the FC layers were re-adapted. In contrast, on extremely small datasets of < 3 ml., re-training only the last layer appeared to be more indicative in order reducing the over-fitting of the model. Given that these parameters appeared to work well in most of the datasets, we propose using them as a rule of thumb on future settings.

5. Conclusions

In this study, we analyzed the effect of intensity domain adaptation on a recent CNN-based MS lesion segmentation method. Given a source model trained on two public MS datasets, we studied how transferable the acquired knowledge was when applied to a private dataset and the ISBI2015 challenge dataset, upon evaluating the minimum number of annotated images needed from the new domain and the minimum number of layers needed to re-train to obtain a comparable accuracy.

Our experiments showed the effectiveness of the proposed domain adaptation model in transferring previously acquired knowledge to new image domains even if only a single training case was available on the target dataset. On the ISBI2015, the accuracy of our one-shot domain-adapted model was comparable to that of a human expert rater and similar to those of other CNN methods trained on a wide set of training data. In this aspect, we believe that the performance shown by our domain adapted models will encourage the MS community to incorporate its use in different clinical settings with reduced amounts of annotated data. This finding could be meaningful not only in terms of the accuracy in delineating MS lesions but also in the related reductions in time and economic costs derived from manual lesion labeling.

Acknowledgements

Mariano Cabezas holds a Juan de la Cierva - Incorporación grant from the Spanish Government with reference number IJCI-2016-29240. This work has been partially supported by La Fundació la Marató de TV3, Spain; by Retos de Investigación TIN2014-55710-R, TIN2015-73563-JIN and DPI2017-86696-R from the Ministerio de Ciencia y Tecnología, Spain. The authors gratefully acknowledge the support of the NVIDIA Corporation with their donation of the TITAN-X PASCAL GPU used in this research.

References

Andermatt, Simon, Pezold, Simon, Cattin, Philippe, 2017. Automated segmentation of multiple sclerosis lesions using multi-dimensional gated recurrent units. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer.

Birenbaum, Ariel, Greenspan, Hayit, 2017. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Eng. Appl. Artif. Intell. 65, 111–118. https://doi.org/10.1016/j.engappai.2017.06.006.

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

Carass, Aaron, Roy, Snehashis, Jog, Amod, Cuzzocreo, Jennifer L., Magrath, Elizabeth, Gherman, Adrian, Button, Julia, et al., 2017. Longitudinal multiple sclerosis lesion segmentation: resource and challenge. NeuroImage 148 (March), 77–102. https://doi.org/10.1016/j.neuroimage.2016.12.064.

Commowick, Olivier, Wiest-Daessle, Nicolas, Prima, Sylvain, 2012. Block-matching strategies for rigid registration of multimodal medical images. In: Proceedings - International Symposium on Biomedical Imaging, pp. 700–703. https://doi.org/10.1109/ISBI.2012.6235644.

Commowick, Olivier, Istace, Audrey, Kain, Michaël, Laurent, Baptiste, Leray, Florent, Simon, Mathieu, Camarasu Pop, Sorina, et al., 2018. Objective evaluation of multiple sclerosis lesion segmentation using a data management and processing infrastructure. Sci. Rep. 8 (1), 13650. https://doi.org/10.1038/s41598-018-31911-7.

Coupé, Pierrick, Yger, Pierre, Prima, Sylvain, Hellier, Pierre, Kervrann, Charles, Barillot, Christian, 2008. An optimized blockwise nonlocal means denoising filter for 3-D magnetic resonance images. IEEE Trans. Med. Imaging 27 (4), 425–441. https://doi.org/10.1109/TMI.2007.906087.

Deshpande, H., Maurel, P., Barillot, C., 2015. Classification of multiple sclerosis lesions using adaptive dictionary learning. Comput. Med. Imaging Graph. 46, 2–10. https://doi.org/10.1016/j.compmedimag.2015.05.003.

Filippi, Massimo, Rocca, Maria A., Ciccarelli, Olga, De, StefanoNicola, Evangelou, Nikos, Kappos, Ludwig, Rovira, Alex, et al., 2016. MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines. Lancet Neurol. 15 (3), 292–303. https://doi.org/10.1016/S1474-4422(15)00393-2.

Ghafoorian, M., Mehrtash, A., Kapur, T., Karssemeijer, N., Marchiori, E., Pesteie, M., Guttmann, C.R.G., et al., 2017. Transfer learning for domain adaptation in MRI: Application in brain lesion segmentation. In: Lecture Notes in Computer Science. 10435. pp. 516–524. https://doi.org/10.1007/978-3-319-66179-7_59.

Greve, Douglas N., Fischl, Bruce, 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48 (1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060.

Hashemi, Seyed Raein, Mohseni Salehi, Seyed Sadegh, Erdogmus, Deniz, Prabhu, Sanjay P., Warfield, Simon K., Gholipour, Ali, 2018. Tversky as a loss function for highly unbalanced image segmentation using 3D fully convolutional deep networks. ArXiv. https://arxiv.org/abs/1803.11078v1 Preprint 1803.11078v1.

He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, Sun, Jian, 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1026–1034. https://doi.org/10.1109/ICCV.2015.123.

Ioffe, Sergey, Szegedy, Christian, 2015. Batch normalization: accelerating deep network training by reducing internal covariate shift. J. Mach. Learn. Res. 37 (February). http://arxiv.org/abs/1502.03167.

Jain, Saurabh, Sima, Diana M., Ribbens, Annemie, Cambron, Melissa, Maertens, Anke, Heckewim, Van, De, MeyJohan, et al., 2015. Automatic segmentation and volumetry of multiple sclerosis brain lesions from MR images. NeuroImage 8, 367–375. https://doi.org/10.1016/j.nicl.2015.05.003.

Kamnitsas, Konstantinos, Baumgartner, Christian, Ledig, Christian, Newcombe, Virginia, Simpson, Joanna, Kane, Andrew, Menon, David, et al., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Lecture Notes in Computer Science, pp. 597–609. https://doi.org/10.1007/978-3-319-59050-9_47. 10265 LNCS.

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

Manjón, José V., Coupé, Pierrick, 2016. volBrain: an online MRI brain volumetry system. Front. Neuroinform. 10 (30). https://doi.org/10.3389/fninf.2016.00030.

Roura, E., Oliver, A., Cabezas, M., Valverde, S., Pareto, D., Vilanova, J.C., Ramió-Torrentà, L., Rovira, A., Lladó, X., 2015. A toolbox for multiple sclerosis lesion segmentation. Neuroradiology 57 (10), 1031–1043. https://doi.org/10.1007/s00234-015-1552-2.

Rovira, Àlex, Wattjes, Mike P., Tintoré, Mar, Tur, Carmen, Yousry, Tarek a., Sormani, Maria P., De, StefanoNicola, et al., 2015. Evidence-based guidelines: MAGNIMS consensus guidelines on the use of MRI in multiple sclerosis—clinical implementation in the diagnostic process. Nat. Rev. Neurol. 11, 1–12. https://doi.org/10.1038/nrneurol.2015.106.

Roy, Snehashis, Butman, John A., Reich, Daniel S., Calabresi, Peter A., Pham, Dzung L., 2018. Multiple sclerosis lesion segmentation from brain MRI via fully convolutional neural networks. ArXiv Preprint 1803.09172 (in press). http://arxiv.org/abs/1803.09172.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., et al., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in Multiple Sclerosis. NeuroImage 59 (4), 3774–3783.

Shiee, Navid, Bazin, Pierre-Louis, Ozturk, Arzu, Reich, Daniel S., Calabresi, Peter A., Pham, Dzung L., 2010. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage 49 (2), 1524–1535. https://doi.org/10.1016/j.neuroimage.2009.09.005.

Sled, J.G., Zijdenbos, a P., Evans, a C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97.

Smith, Stephen M., Zhang, Yongyue, Jenkinson, Mark, Chen, Jacqueline, Matthews, P.M., Federico, Antonio, StefanoNicola, De, 2002. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. NeuroImage 17 (1), 479–489.

Srivastava, Nitish, Hinton, Geoffrey E., Krizhevsky, Alex, Sutskever, Ilya, Salakhutdinov, Ruslan, 2014. Dropout: a simple way to prevent neural networks from overfitting. J. Machine Learn. Res. 15, 1929–1958. https://doi.org/10.1214/12-AOS1000.

Styner, Martin, Joohwi Lee, B. Chin, Chin, M., 2008. 3D segmentation in the clinic: A grand challenge II: MS lesion segmentation. Midas 1–6. http://grand-challenge2008.bigr.nl/proceedings/pdfs/msls08/Styner.pdf.

Sudre, Carole H., Cardoso, M. Jorge, Bouvy, Willem H., Biessels, Geert Jan, Barnes, Josephine, Ourselin, Sebastien, 2015. Bayesian model selection for pathological neuroimaging data applied to white matter lesion segmentation. IEEE Trans. Med. Imaging 34 (10), 2079–2102. https://doi.org/10.1109/TMI.2015.2419072.

Tustison, Nicholas J., Avants, Brian B., Cook, Philip A., Zheng, Yuanjie, Egan, Alexander, Yushkevich, Paul A., Gee, James C., 2010. N4ITK: improved N3 bias correction. IEEE Trans. Med. Imaging 29 (6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908.

Valcarcel, A.M., Linn, K.A., Vandekar, S.N., Satterthwaite, T.D., Muschelli, J., Calabresi, P.A., Pham, D.L., Martin, M.L., Shinohara, R.T., 2018. MIMoSA: an automated method for intermodal segmentation analysis of multiple sclerosis brain lesions. J. Neuroimaging 00, 1–10. https://doi.org/10.1111/jon.12506.

Valverde, Sergi, Cabezas, Mariano, Roura, Eloy, González-Villà, Sandra, Pareto, Deborah, Vilanova, Joan C., Ramió-Torrentà, Lluís, Rovira, Àlex, Oliver, Arnau, Lladó, Xavier, 2017. Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach. NeuroImage 155, 159–168. https://doi.org/10.1016/j.neuroimage.2017.04.034.

Zeiler, Matthew D., 2012. ADADELTA: an adaptive learning rate method. ArXiv Preprint 1212.5701. http://arxiv.org/abs/1212.5701.