Transfer Learning for Domain Adaptation in MRI: Application in Brain Lesion Segmentation

Mohsen Ghafoorian[1,2,3], Alireza Mehrtash[2,4] *, Tina Kapur[2], Nico Karssemeijer[1], Elena Marchiori[3], Mehran Pesteie[4], Charles R. G. Guttmann[2], Frank-Erik de Leeuw[5], Clare M. Tempany[2], Bram van Ginneken[1], Andriy Fedorov[2], Purang Abolmaesumi[4], Bram Platel[1], and William M. Wells III[2]

[1] Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen, the Netherlands
mohsen.ghafoorian@radboudumc.nl
[2] Radiology Department, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA
mehrtash@bwh.harvard.edu
[3] Institute for Computing and Information Sciences, Radboud University, Nijmegen, the Netherlands
[4] University of British Columbia, Vancouver, BC, Canada
[5] Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands

Abstract. Magnetic Resonance Imaging (MRI) is widely used in routine clinical diagnosis and treatment. However, variations in MRI acquisition protocols result in different appearances of normal and diseased tissue in the images. Convolutional neural networks (CNNs), which have shown to be successful in many medical image analysis tasks, are typically sensitive to the variations in imaging protocols. Therefore, in many cases, networks trained on data acquired with one MRI protocol, do not perform satisfactorily on data acquired with different protocols. This limits the use of models trained with large annotated legacy datasets on a new dataset with a different domain which is often a recurring situation in clinical settings. In this study, we aim to answer the following central questions regarding domain adaptation in medical image analysis: Given a fitted legacy model, 1) How much data from the new domain is required for a decent adaptation of the original network?; and, 2) What portion of the pre-trained model parameters should be retrained given a certain number of the new domain training samples? To address these questions, we conducted extensive experiments in white matter hyperintensity segmentation task. We trained a CNN on legacy MR images of brain and evaluated the performance of the domain-adapted network on the same task with images from a different domain. We then compared the performance of the model to the surrogate scenarios where either the same trained network is used or a new network is trained from scratch on the new dataset. The domain-adapted network tuned only by two training examples achieved a Dice score of 0.63 substantially outperforming a similar network trained on the same set of examples from scratch.

* Mohsen Ghafoorian and Alireza Mehrtash contributed equally to this work.

1 Introduction

Deep neural networks have been extensively used in medical image analysis and have outperformed the conventional methods for specific tasks such as segmentation, classification and detection [1]. For instance on brain MR analysis, convolutional neural networks (CNN) have been shown to achieve outstanding performance for various tasks including white matter hyperintensities (WMH) segmentation [2], tumor segmentation [3], microbleed detection [4], and lacune detection [5]. Although many studies report excellent results on specific domains and image acquisition protocols, the generalizability of these models on test data with different distributions are often not investigated and evaluated. Therefore, to ensure the usability of the trained models in real world practice, which involves imaging data from various scanners and protocols, domain adaptation remains a valuable field of study. This becomes even more important when dealing with Magnetic Resonance Imaging (MRI), which demonstrates high variations in soft tissue appearances and contrasts among different protocols and settings.

Mathematically, a domain D can be expressed by a feature space χ and a marginal probability distribution $P(X)$, where $X = \{x_1, ..., x_n\} \in \chi$ [6]. A supervised learning task on a specific domain $D = \{\chi, P(X)\}$, consists of a pair of a label space Y and an objective predictive function $f(.)$ (denoted by $T = \{Y, f(.)\}$). The objective function $f(.)$ can be learned from the training data, which consists of pairs $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in Y$. After the training process, the learned model denoted by $\tilde{f}(.)$ is used to predict the label for a new instance x. Given a source domain D_S with a learning task T_S and a target domain D_T with learning task T_T, transfer learning is defined as the process of improving the learning of the target predictive function $f_T(.)$ in D_T using the information in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$ [6]. We denote $\tilde{f}_{ST}(.)$ as the predictive model initially trained on the source domain D_S, and domain-adapted to the target domain D_T.

In the medical image analysis literature, transfer classifiers such as adaptive SVM and transfer AdaBoost, are shown to outperform the common supervised learning approaches in segmenting brain MRI, trained only on a small set of target domain images [7]. In another study a machine learning based sample weighting strategy was shown to be capable of handling multi-center chronic obstructive pulmonary disease images [8]. Recently, also several studies have investigated transfer learning methodologies on deep neural networks applied to medical image analysis tasks. A number of studies used networks pre-trained on natural images to extract features and followed by another classifier, such as a Support Vector Machine (SVM) or a random forest [9]. Other studies [10,11] performed layer fine-tuning on the pre-trained networks for adapting the learned features to the target domain.

Considering the hierarchical feature learning fashion in CNN, we expect the first few layers to learn features for general simple visual building blocks, such as edges, corners and simple blob-like structures, while the deeper layers learn more complicated abstract task-dependent features. In general, the ability to learn domain-dependent high-level representations is an advantage enabling CNNs to

achieve great recognition capabilities. However, it is not obvious how these qualities are preserved during the transfer learning process for domain adaptation. For example, it would be practically important to determine how much data on the target domain is required for domain adaptation with sufficient accuracy for a given task, or how many layers from a model fitted on the source domain can be effectively transferred to the target domain. Or more interestingly, given a number of available samples on the target domain, what layer types and how many of those can we afford to fine-tune. Moreover, there is a common scenario in which a large set of annotated legacy data is available, often collected in a time-consuming and costly process. Upgrades in the scanners, acquisition protocols, etc., as we will show, might make the direct application of models trained on the legacy data unsuccessful. To what extent these legacy data can contribute to a better analysis of new datasets, or vice versa, is another question worth investigating.

In this study, we aim towards answering the questions discussed above. We use transfer learning methodology for domain adaptation of models trained on legacy MRI data on brain WMH segmentation.

2 Materials and Method

2.1 Dataset

Radboud University Nijmegen Diffusion tensor and Magnetic resonance imaging Cohort (RUN DMC) [12] is a longitudinal study of patients diagnosed with small vessel disease. The baseline scans acquired in 2006 consisted of fluid-attenuated inversion recovery (FLAIR) images with voxel size of $1.0\times1.2\times5.0$ mm and an inter-slice gap of 1.0 mm, scanned with a 1.5 T Siemens scanner. However, the follow-up scans in 2011 were acquired differently with a voxel size of $1.0\times1.2\times3.0$ mm, including a slice gap of 0.5 mm. The follow-up scans demonstrate a higher contrast as the partial volume effect is less of an issue due to thinner slices. For each subject, we also used 3D T1 magnetization-prepared rapid gradient-echo (MPRAGE) with voxel size of $1.0\times1.0\times1.0$ mm which is the same among the two datasets. Reference WMH annotations on both datasets were provided semi-automatically, by manually editing segmentations provided by a WMH segmentation method [13] wherever needed.

The T1 images were linearly registered to FLAIR scans, followed by brain extraction and bias-filed correction operations. We then normalized the image intensities to be within the range of [0, 1].

In this study, we used 280 patient acquisitions with WMH annotations from the baseline as the source domain, and 159 scans from all the patients that were rescanned in the follow-up as the target domain. Table 1 shows the data split into the training, validation and test sets. It should be noted that the same patient-level partitioning which was used on the baseline, was respected on the follow-up dataset to prevent potential label leakages.

Table 1. Number of patients for the domain adaptation experiments.

	Source Domain			Target Domain		
Set	Train	Validation	Test	Train	Validation	Test
Size	200	30	50	100	26	33

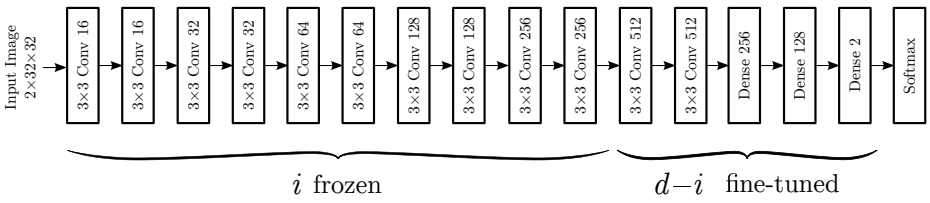


Fig. 1. Arcitecture of the convolutional neural network used in our experiments. The shallowest i layers are frozen and the rest $d - i$ layers are fine-tuned. d is the depth of the network which was 15 in our experiments.

2.2 Sampling

We sampled 32×32 patches to capture local neighborhoods around WMH and normal voxels from both FLAIR and T1 images. We assigned each patch with the label of the corresponding central voxel. To be more precise, we randomly selected 25% of all voxels within the WMH masks, and randomly selected the same number of negative samples from the normal appearing voxels inside the brain mask. We augmented the dataset by flipping the patches along the y axis. This procedure resulted in training and validation datasets of size ∼1.2m and ∼150k on the baseline, and ∼1.75m and ∼200k on the followup.

2.3 Network Architecture and Training

We stacked the FLAIR and T1 patches as the input channels and used a 15-layer architecture consisting of 12 convolutional layers of 3×3 filters and 3 dense layers of 256, 128 and 2 neurons, and a final softmax layer. We avoided using pooling layers as they would result in a shift-invariance property that is not desirable in segmentation tasks, where the spatial information of the features are important to be preserved. The network architecture is illustrated in Figure 1.

To tune the weights in the network, we used the Adam update rule [14] with a mini-batch size of 128 and a binary cross-entropy loss function. We used the Rectified Linear Unit (ReLU) activation function as the non-linearity and the He method [15] that randomly initializes the weights drawn from a $\mathcal{N}(0, \sqrt{\frac{2}{m}})$ distribution, where m is the number of inputs to a neuron. Activations of all layers were batch-normalized to speed up the convergence [**?**]. A decaying learning rate was used with a starting value of 0.0001 for the optimization process.

To avoid over-fitting, we regularized our networks with a drop-out rate of 0.3 as well as the L_2 weight decay with λ_2=0.0001. We trained our networks for a maximum of 100 epochs with an early stopping policy. For each experiment, we picked the model with the highest area under the curve on the validation set.

We trained our networks with a patch-based approach. At segmentation time, however, we converted the dense layers to their equivalent convolutional counterparts to form a fully convolutional network (FCN). FCNs are much more efficient as they avoid the repetitive computations on neighboring patches by feeding the whole image into the network. We prefer the conceptual distinction between dense and convolutional layers at the training time, to keep the generality of experiments for classification problems as well (e.g., testing the benefits of fine-tuning the convolutional layers in addition to the dense layers). Patch-based training allows class-specific data augmentation to handle domains with hugely imbalanced class ratios (e.g., WMH segmentation domain).

2.4 Domain Adaptation

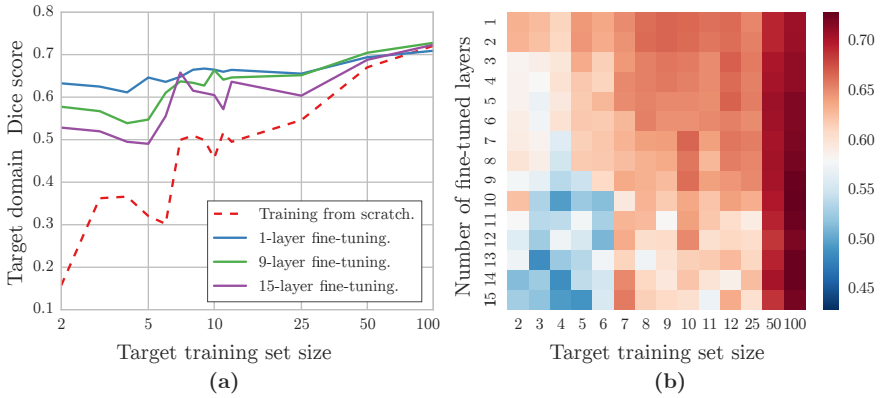
To build the model $\tilde{f}_{ST}(.)$, we transferred the learned weights from $\tilde{f}_S$, then we froze shallowest i layers and fine-tuned the remaining $d - i$ deeper layers with the training data from D_T, where d is the depth of the trained CNN. This is illustrated in Figure 1. We used the same optimization update-rule, loss function, and regularization techniques as described in Section 2.3.

2.5 Experiments

On the WMH segmentation domain, we investigated and compared three different scenarios: 1) Training a model on the source domain and directly applying it on the target domain; 2) Training networks on the target domain data from scratch; and 3) Transferring model learned on the source domain onto the target domain with fine-tuning. In order to identify the target domain dataset sizes where transfer learning is most useful, the second and third scenarios were explored with different training set sizes of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 25, 50 and 100 cases. We extensively expanded the third scenario investigating the best freezing/tuning cut-off for each of the mentioned target domain training set sizes. We used the same network architecture and training procedure among the different experiments. The reported metric for the segmentation quality assessment is the Dice score.

3 Results

The model trained on the set of images from the source domain ($\tilde{f}_S$), achieved a Dice score of 0.76. The same model, without fine-tuning, failed on the target domain with a Dice score of 0.005. Figure 2(a) demonstrates and compares the Dice scores obtained with three domain-adapted models to a network trained from scratch on different target training set sizes. Figure 2(b) illustrates the



(a) (b)

Fig. 2. (a) The comparison of Dice scores on the target domain with and without transfer learning. A logarithmic scale is used on the x axis. **(b)** Given a deep CNN with d=15 layers, transfer learning was performed by freezing the i initial layers and fine-tuning the last $d - i$ layers. The Dice scores on the test set are illustrated with the color-coded heatmap. On the map, the number of fine-tuned layers are shown horizontally, whereas the target domain training set size is shown vertically.

target domain test set Dice scores as a function of target domain training set size and the number of abstract layers that were fine-tuned. Figure 3 presents and compares qualitative results of WMH segmentation of several different models of a single sample slice.

4 Discussion and Conclusions

We observed that while $\tilde{f}_S$ demonstrated a decent performance on D_S, it totally failed on D_T. Although the same set of learned representations is expected to be useful for both as the two tasks are similar, the failure comes to no surprise as the distribution of the responses to these features are different. Observing the comparisons presented by Figure 2(a), it turns out that given only a small set of training examples on D_T, the domain adapted model substantially outperforms the model trained from scratch with the same size of training data. For instance, given only two training images, $\tilde{f}_{ST}$ achieved a Dice score of 0.63 on a test set of 33 target domain test images, while $\tilde{f}_T$ resulted in a dice of 0.15. As Figure 2(b) suggests, with only a few D_T training cases available, best results can be achieved by fine-tuning only the last dense layers, otherwise enormous number of parameters compared to the training sample size would result in over-fitting. As soon as more training data becomes available, it makes more sense to fine-tune the shallower representations (e.g., the last convolutional layers). It

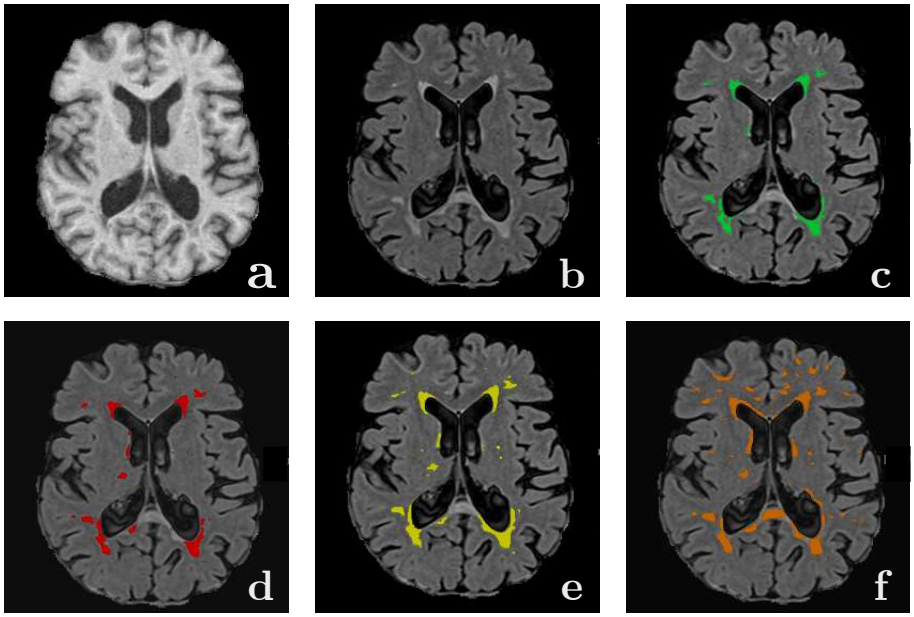


Fig. 3. Examples of the brain WMH MRI segmentations. **(a)** Axial T1-weighted image. **(b)** FLAIR image. **(c-f)** FLAIR images with WMH segmented labels: **(c)** reference (green) WMH. **(d)** WMH (red) from a domain adapted model ($\tilde{f}_{ST}(.)$) fine-tuned on five target training samples. **(e)** WMH (yellow) from model trained from scratch ($\tilde{f}_{ST}(.)$) on 100 target training samples. **(f)** WMH (orange) from model trained from scratch ($\tilde{f}_{ST}(.)$) on 5 target training samples.

is also interesting to note that tuning the first few convolutional layers is rarely useful considering their domain-independent characteristics.

References

1. G. Litjens, T. Kooi, B. Ehteshami Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, Jeroen A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *arXiv preprint arXiv:1702.05747*, 2017.
2. M. Ghafoorian, N. Karssemeijer, T. Heskes, I. van Uden, C. Sanchez, G. Litjens, F. de Leeuw, B. van Ginneken, E. Marchiori, and B. Platel. Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities. *arXiv preprint arXiv:1610.04834*, 2016.
3. K. Kamnitsas, C. Ledig, V. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, 2017.

4. Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. CT Mok, L. Shi, and P. A. Heng. Automatic detection of cerebral microbleeds from mr images via 3d convolutional neural networks. *IEEE transactions on medical imaging*, 35(5):1182–1195, 2016.
5. M. Ghafoorian, N. Karssemeijer, T. Heskes, M. Bergkamp, J. Wissink, J. Obels, K. Keizer, de Leeuw F.E, B. van Ginneken, E. Marchiori, and B. Platel. Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin. *NeuroImage: Clinical*, feb 2017.
6. S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.
7. A. Van Opbroek, M A. Ikram, M. W Vernooij, and M. De Bruijne. Transfer learning improves supervised image segmentation across imaging protocols. *IEEE transactions on medical imaging*, 34(5):1018–1030, 2015.
8. V. Cheplygina, I. P. Pena, J. H. Pedersen, D. A Lynch, L. Sørensen, and M. de Bruijne. Transfer learning for multi-center classification of chronic obstructive pulmonary disease. *arXiv preprint arXiv:1701.05013*, 2017.
9. A. Esteva, B. Kuprel, R. A Novoa, J. Ko, S. M Swetter, H. M Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
10. N. Tajbakhsh, J. Y Shin, S. R Gurudu, R Todd Hurst, Christopher B Kendall, Michael B Gotway, and Jianming Liang. Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE transactions on medical imaging*, 35(5):1299–1312, 2016.
11. H. C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers. Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning. *IEEE Transactions on Medical Imaging*, 35(5):1285–1298, 2016.
12. A. G. van Norden, K. F. de Laat, R. A. Gons, I. W. van Uden, E. J. van Dijk, L. J. van Oudheusden, R. A. Esselink, B. R. Bloem, B. G. van Engelen, M. J. Zwarts, I. Tendolkar, M. G. Olde-Rikkert, M. J. van der Vlugt, M. P. Zwiers, D. G. Norris, and F. E. de Leeuw. Causes and consequences of cerebral small vessel disease. The RUN DMC study: a prospective cohort study. Study rationale and protocol. *BMC Neurol*, 11:29, 2011.
13. M. Ghafoorian, N. Karssemeijer, I. WM van Uden, F.E de Leeuw, T. Heskes, E. Marchiori, and B. Platel. Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease. *Medical Physics*, 43(12):6246–6258, 2016.
14. D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
15. K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1026–1034, 2015.