Improving Pathological Structure Segmentation via Transfer Learning Across Diseases

Barleen Kaur[1,2,4(✉)], Paul Lemaître[2], Raghav Mehta[2], Nazanin Mohammadi Sepahvand[2], Doina Precup[1,4], Douglas Arnold[3,5], and Tal Arbel[2]

[1] School of Computer Science, McGill University, Montreal, Canada
barleen.kaur@mail.mcgill.ca
[2] Centre for Intelligent Machines, McGill University, Montreal, Canada
[3] Montreal Neurological Institute, McGill University, Montreal, Canada
[4] Mila Quebec AI Institute, Montreal, Canada
[5] NeuroRx Research, Montreal, Canada

Abstract. One of the biggest challenges in developing robust machine learning techniques for medical image analysis is the lack of access to large-scale annotated image datasets needed for supervised learning. When the task is to segment pathological structures (e.g. lesions, tumors) from patient images, training on a dataset with few samples is very challenging due to the large class imbalance and inter-subject variability. In this paper, we explore how to best leverage a segmentation model that has been pre-trained on a large dataset of patients images with one disease in order to successfully train a deep learning pathology segmentation model for a different disease, for which only a relatively small patient dataset is available. Specifically, we train a UNet model on a large-scale, proprietary, multi-center, multi-scanner Multiple Sclerosis (MS) clinical trial dataset containing over 3500 multi-modal MRI samples with expert-derived lesion labels. We explore several transfer learning approaches to leverage the learned MS model for the task of multi-class brain tumor segmentation on the BraTS 2018 dataset. Our results indicate that adapting and fine-tuning the encoder and decoder of the network trained on the larger MS dataset leads to improvement in brain tumor segmentation when few instances are available. This type of transfer learning outperforms training and testing the network on the BraTS dataset from scratch as well as several other transfer learning approaches, particularly when only a small subset of the dataset is available.

Keywords: Transfer learning · Brain tumor segmentation · MRI

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-33391-1_11) contains supplementary material, which is available to authorized users.

Q. Wang et al. (Eds.): DART 2019/MIL3ID 2019, LNCS 11795, pp. 90–98, 2019.
https://doi.org/10.1007/978-3-030-33391-1_11

1 Introduction

An important challenge in developing robust pathology segmentation methods in medical imaging is the lack of access to sufficiently large annotated datasets needed for training. Large datasets are required for a number of reasons. First, many of the state-of-art models are based on deep learning methods, which perform well when trained on large datasets [6,13]. Second, pathological structures (e.g. lesions, tumors) tend to be present in only small parts of an image, leading to large class imbalance, and also presents with high variability between patients, exacerbating the need to have annotations for many patients. Unfortunately, larger proprietary datasets cited in the literature are not always available for public use and public labelled pathology segmentation datasets are often relatively small.

To overcome this problem, transfer learning has recently been explored in various medical imaging applications, including classification [9], detection [10] and segmentation [7] tasks (see [4] for a survey). Investigated tasks include using data acquired from different scanners [7] or detection of different types of abnormality in the same set of data [22]. It has also been shown that knowledge could be transferred from both medical and non-medical datasets to improve results in other medical applications [8,15]. Deep networks trained on a larger source dataset have been used as feature extractors [9] or as a starting point for fine-tuning further on target data [20].

This paper explores the hypothesis that transfer learning for the segmentation of pathological structures can be performed across diseases. Specifically, we leverage a deep learning segmentation network pre-trained on a large pathology segmentation dataset, in order to improve segmentation performance on a small dataset, in a scenario in which: (a) the two image datasets are acquired from patients with different diseases, (b) the pathological structures are different in the two datasets (lesions vs. tumors), and (c) the inference tasks themselves differ (binary vs. multi-class segmentation). We explore several fine-tuning strategies to see how to best leverage the source model and adapt it to the target dataset, including: freezing the network and only retraining the last few layers, fine-tuning only the decoder, or carefully fine-tuning the entire network.

Experimental validation of the methods involves first pre-training a binary classifier for the segmentation of T2 lesions based on a large proprietary, multi-scanner, multi-center, longitudinal clinical trial, MRI dataset of 1385 patients with relapsing-remitting Multiple Sclerosis (RRMS), along with expert-labelled T2 lesions. Next, a series of experiments are performed in order to explore the ability of transfer learning to improve the results of an end-to-end multi-class brain tumor segmentation network trained on subsets of the MICCAI 2018 BraTS dataset [16]. Given that both MRI datasets are acquired from patients with neurological diseases that present with focal pathologies (lesions and tumors), the intuition is that the two dataset share common features. As such, the framework should be able to leverage the representation learned by the lesion segmentation network trained on the bigger MS dataset to improve the segmentation results on the smaller brain tumor dataset.

2 Methodology

We use a 3D deep neural network inspired by UNet [6] for the task of focal pathology segmentation. It consists of an encoder followed by a decoder which combines higher resolution features from the contracting path at different levels, in order to learn multi-scale representations. The architecture is depicted in Fig. 1(a), and the implementation details of the model are described in Sect. 3.2.

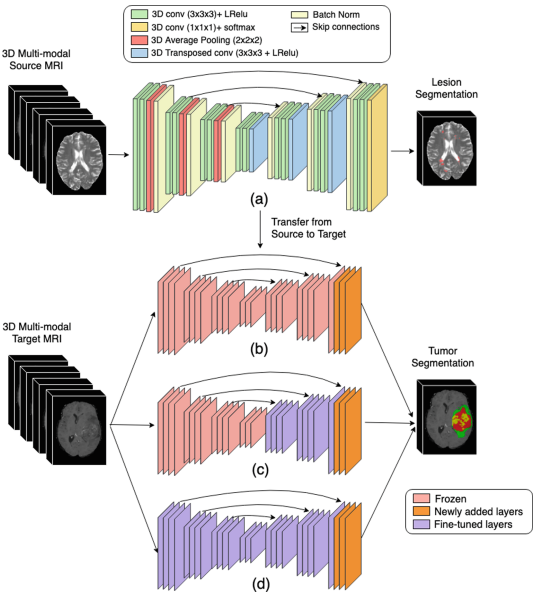

Fig. 1. Transfer learning framework. (a) UNet architecture for pre-trained source network. (b), (c) and (d) depict different methods of adapting the pre-trained source network for the target task. In all three, the last three task-specific layers are replaced with new layers (orange) and the remaining network is fine-tuned such that: (b) only the newly added layers are re-trained (FT-Last Three), (c) only the decoder is fine-tuned (FT-Decoder) and (d) the whole network is fine-tuned (FT-All) with the target data respectively. (Color figure online)

Given a source network trained from scratch on a large source dataset, the objective is to transfer the representation learned by the source network and adapt it to the (smaller) target set in order to improve pathology segmentation performance. A popular strategy for transfer learning consists of fine-tuning the pre-trained source network on the target dataset. In this paper, we explore three different strategies of fine-tuning. The most common way of fine-tuning consists of replacing the last few layers of the source network with new layers, by re-initializing the weights and changing the output dimension of these layers. The remainder of the network is frozen, which prevents the gradient flow. The newly

added layers are trained on the target dataset (See Fig. 1(b)). This strategy has been advocated when the amount of target data available is small and the similarity between the two datasets is high [7], as in the context explored in this paper. The intuition behind this approach is that the initial layers of the network tend to learn low level image features (e.g. edges, orientations) that are generic and therefore useful across different datasets and tasks, while the higher layers of the network tend to capture more complex patterns that are specific to a particular task. When the source and target datasets are similar, and/or more target data is available, more layers can be fine-tuned [5,21]. This leads to the second strategy we explore, which involves freezing the encoder and fine-tuning the entire decoder (See Fig. 1(c)). The third strategy consists of fine-tuning the whole network with target data (See Fig. 1(d)).

3 Experiments and Results

In order to assess the performance of the three different transfer learning approaches in the context of pathology segmentation, we perform experiments using a large source dataset of MS patients, in which the segmentation network is trained to label lesions. The target task is to segment brain tumors and their tissue sub-classes from patient MRI. We compare the performance of the transfer learning approach to training only on the target data, for different dataset sizes. The segmentation performance is assessed using Dice scores.

3.1 Data Description and Preprocessing

Multiple Sclerosis Dataset (Source): The source task involves a binary classification to differentiate T2 hyperintense lesions from healthy tissues in a proprietary, multi-modal MRI dataset acquired from Multiple Sclerosis (MS) patients participating in a multi-site, multi-scanner clinical trial. The dataset consists of 1385 patients, scanned annually for up to a 24-month period, totalling 3630 multi-sequence 3D MRI samples consisting of T1-weighted, T2-weighted, Fluid Attenuated Inverse Recovery (FLAIR), and T1 post-Gadolinium sequences acquired at $1\,\text{mm} \times 1\,\text{mm} \times 3\,\text{mm}$ resolution. They are then interpolated to $1\,\text{mm}^3$ isotropic resolution, which results in MRIs of size $229 \times 193 \times 193$. T2 binary lesion segmentation masks provided with the dataset are obtained through expert manual corrections as a result of a proprietary automatic segmentation method. Preprocessing includes brain extraction [19], N3 bias field in homogeneity correction [18], Nyul image intensity normalization [17], and registration to the MNI-space.

Brain Tumor Dataset (Target): The target datasets are obtained by subsampling datasets of various sizes from the BraTS 2018 MICCAI challenge [2,3,16]. The entire training dataset consists of 210 high-grade glioma (HGG) and 75 low-grade glioma (LGG) patients and the validation set consists of 66 patients. Each sample contains T1-w, T1 post contrast (T1c), T2-w, and FLAIR 3D MR

sequences. Ground truth segmentation labels are provided for the BraTS Training set (used for training the network) but not for the BraTS Validation set[1] (used for testing). Tumors are segmented into 3 classes: edema, necrotic/non-enhancing core, and enhancing tumor. These three classes combined together are referred to as "whole" tumor. The volumes are co-registered, resampled to 1 mm^3 resolution and skull-stripped. Our pre-processing pipeline includes registration of samples to the same space as MS data using ANTs tool [1].

For both MS Dataset and Brain Tumor Dataset, the image intensities are then standardized using mean subtraction, division by standard deviation, and rescaled to range from 0 to 1. The images are standardized to $240 \times 192 \times 192$ using zero-padding and cropping operations.

3.2 Model Implementation Details

The proposed segmentation network takes 3D patient MRI sequences as input and generates a 3D output mask of the same resolution. As is typical of a 3D UNet [6,14], the network consists of an encoder part and a decoder part of 4 resolution steps each. The encoder part consists of two consecutive 3D convolutions of size $3 \times 3 \times 3$ with $k * 2^{(n-1)}$ filters, where n is the resolution step and k is the initial number of filters (4 in our case). Each convolution is followed by a leaky rectified linear unit (L-ReLU). Average pooling of size $2 \times 2 \times 2$ and stride of 2 is performed followed by Batch normalization [11]. In the decoder part, each step consists of 3D transposed convolutions of size $3 \times 3 \times 3$ with $2 \times 2 \times 2$ stride and $k * 2^{(n-1)}$ filters for upsampling, whose output is concatenated with the corresponding output of the encoder part. Batch normalization is applied again following which, two $3 \times 3 \times 3$ convolutions with L-ReLU activation are applied. The last layer consists of $1 \times 1 \times 1$ convolution with F filters, where F denotes the number of classes for the task, followed by a SoftMax non-linearity. The implementation of the model is done in Pytorch.[2]

Segmenting MS lesions is a binary voxel-wise classification task whereas brain tumor sub-type segmentation is a 4-class voxel-wise classification task [16]. For lesion segmentation, the training objective is weighted binary cross entropy loss (to account for class imbalance). For the multi-class brain tumor segmentation task, the training objective is weighted categorical cross entropy loss. The weight of a class c is calculated as the ratio of the total number of voxels divided by the number of voxels belonging to class c in the training set. The class weights are scheduled to decay [12] with a decay rate lower than 1. As the number of epochs increase, the weight for each class converges to 1, ensuring that every class is given equal importance during the later stages of training.

[1] Please note that the predictions made on the BraTS 2018 Validation set must contain all four tumor sub-classes, which are then uploaded onto the BraTS web portal for evaluation.

[2] http://pytorch.org/.

3.3 Experiments

As described in Sect. 2, the baseline experiment consists of training a network from scratch on the brain tumor dataset. The other three experiments use a network trained on the MS dataset from scratch, which is then fine-tuned using the three transfer learning approaches discussed above and denoted as FT-Last Three, FT-Decoder, FT-All in Fig. 1. When pre-training the MS lesion segmentation network, 80% of the MS data (2912 samples) is used for training, and the remaining 20% is left out for validation (718 samples) for 190 epochs. The best validation performance of the pre-trained network is obtained at epoch 186 with an AUC of 0.77.

In order to examine the effect of the size of the target dataset on the transfer learning outcome, the number of patient brain tumor MRI samples extracted from the BraTS 2018 training dataset and used in the target dataset is set to several values: 20, 50, 100, 150. For each case, the fine-tuned networks are compared to the corresponding baseline network. For all experiments, the ratio of high-grade gliomas (HGG) to low-grade gliomas (LGG) is maintained across folds. Four-fold cross validation is performed on the respective training sets to determine the best parameters (see Supplementary Materials[3] document for more information on hyper-parameter tuning). Then, the networks are retrained on the respective complete training sets, using the hyper-parameters that performed best during cross-validation and a local validation set (subset of BraTS 2018 training set) of 50 samples is used to select the operating point. Performance is evaluated on the separate BraTS 2018 Validation set, for which the ground truth is not available.

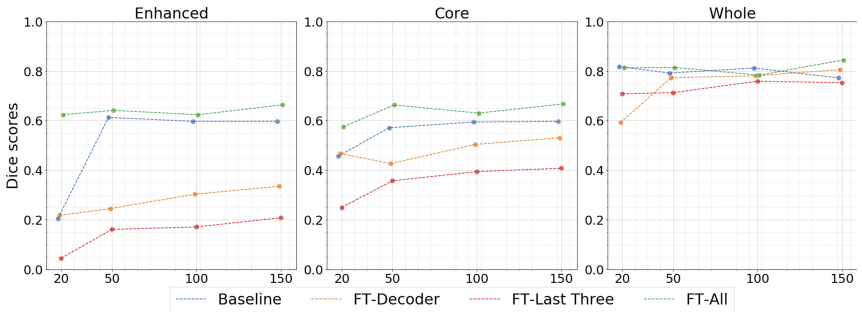

Fig. 2. Comparison of Dice values for baseline method against different fine-tuning methods for enhanced, core and whole tumor segmentation on the Brats 2018 validation set. The x-axis depicts a varying number of brain tumor cases available for training (20, 50, 100, 150).

[3] http://cim.mcgill.ca/~barleenk/MICCAI2019_transfer_appendix.pdf.

3.4 Results

Figure 2 summarizes all the Dice scores obtained on the BraTS 2018 validation set for the baseline and various transfer learning methods, as a function of the number of brain tumor cases available for fine-tuning. The epoch for which the sum of the Dice scores is best on the local validation set, is selected as an operating point. The results indicate that FT-All outperforms the baseline results in almost every case and consistently provides the best Dice scores for core and enhanced tumor, particularly when the number of tumor cases is extremely low, with 25.9% and 204.09% improvement[4] on core and enhanced tumor over baseline respectively when the number of cases is 20. See Supplementary Materials document for more results. Since lesions are smaller in size when compared to tumors, the results indicate that the network is extracting information from the MS pre-trained network that is relevant to segmenting sub-regions of tumor well, even though lesions present quite differently than brain tumors. As the number of brain tumor samples increase, the gain of FT-All over baseline diminishes. FT-Last Three and FT-Decoder don't perform as well as the baseline. This is likely due to low-level representations not getting updated as per the target task, which in turn fuse with high level representations in the UNet to produce an output. Qualitative segmentation results of the different methods on the local

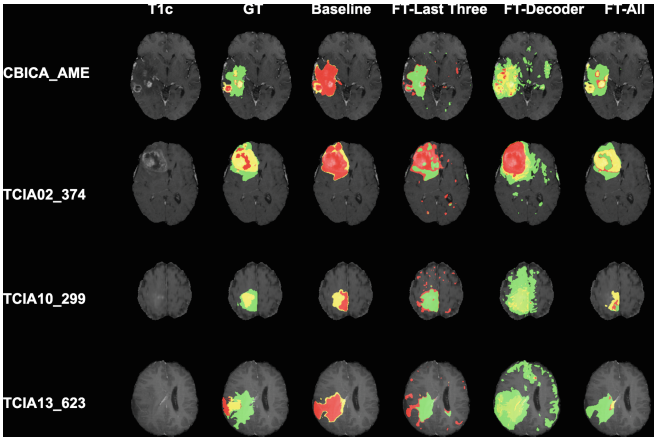

Fig. 3. Examples of visualizations obtained on a local validation set when fine-tuning with 20 BraTS samples for 4 patients (IDs on left). Top two rows and bottom two rows illustrate the segmentation results obtained on HGG and LGG cases respectively. From left to right: T1c MRI (column 1), ground truth segmentation (column 2), results of baseline experiment (column 3), FT-Last Three (column 4), FT-Decoder (column 5) and FT-All (column 6) are shown. Edema, necrotic core and enhancing tumor are shown in green, red and yellow respectively. (Color figure online)

[4] The percentage improvement is calculated as the ratio of difference in the baseline and FT-All Dice scores over the baseline.

validation set for the case of 20 target dataset samples are shown in Fig. 3. Note that with just 20 target dataset samples, FT-All is able to capture different substructures of tumor better than the other methods. Performance is better on the HGG over the LGG cases, as more HGG cases are present in the training dataset. More results are presented in the Supplementary Materials document.

4 Conclusions

In this work, we explore different strategies for transfer learning across diseases for the task of focal pathology segmentation. Fine-tuning the entire network trained on a larger MS dataset improves the multi-class brain tumor segmentation results on target MRI datasets, outperforming the baseline method and the other fine-tuning methods, especially when only very small target datasets are available. This outcome indicates that transfer learning methods can have a significant impact, particularly for diseases where there is little access to large scale, annotated datasets needed for training segmentation networks. The public release of more models that have been pre-trained on large proprietary datasets (e.g. where it is not possible to release the images themselves) will permit the community to leverage them for the large set of applications with small datasets.

Acknowledgments. The MS dataset was provided through an award from the International Progressive MS Alliance (PA-1603-08175). The authors would also like to thank Nicholas J. Tustison for his guidance on using ANTs tool.

References

1. Avants, B.B., et al.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage **54**(3), 2033–2044 (2011)
2. Bakas, S., et al.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Sci. Data **4**, 170117 (2017)
3. Bakas, S., et al.: Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection. TCIA, vol. 286 (2017)
4. Cheplygina, V., et al.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. MIA **54**, 280–296 (2019)
5. Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., Darrell, T.: Best practices for fine-tuning visual classifiers to new domains. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9915, pp. 435–442. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49409-8_34
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 424–432. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_49
7. Ghafoorian, M., et al.: Transfer learning for domain adaptation in MRI: application in brain lesion segmentation. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10435, pp. 516–524. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66179-7_59

8. Hussein, S., Cao, K., Song, Q., Bagci, U.: Risk stratification of lung nodules using 3D CNN-based multi-task learning. In: Niethammer, M., et al. (eds.) IPMI 2017. LNCS, vol. 10265, pp. 249–260. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59050-9_20
9. Huynh, B.Q., et al.: Digital mammographic tumor classification using transfer learning from deep convolutional neural networks. JMI **3**(3), 034501 (2016)
10. Hwang, S., Kim, H.-E.: Self-transfer learning for weakly supervised lesion localization. In: Ourselin, S., Joskowicz, L., Sabuncu, M.R., Unal, G., Wells, W. (eds.) MICCAI 2016. LNCS, vol. 9901, pp. 239–246. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46723-8_28
11. Ioffe, S., et al.: Batch normalization: accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
12. Jesson, A., Arbel, T.: Brain tumor segmentation using a 3D FCN with multi-scale loss. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 392–402. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_34
13. Kamnitsas, K., et al.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. MIA **36**, 61–78 (2017)
14. Mehta, R., Arbel, T.: 3D U-Net for brain tumour segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Keyvan, F., Reyes, M., van Walsum, T. (eds.) BrainLes 2018. LNCS, vol. 11384, pp. 254–266. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-11726-9_23
15. Menegola, A., et al.: Knowledge transfer for melanoma screening with deep learning. ISBI **2017**, 297–300 (2017)
16. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). TMI **34**(10), 1993–2024 (2014)
17. Nyúl, L.G., et al.: New variants of a method of MRI scale standardization. TMI **19**(2), 143–150 (2000)
18. Sled, J.G., et al.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. TMI **17**(1), 87–97 (1998)
19. Smith, S.M.: Fast robust automated brain extraction. HBM **17**(3), 143–155 (2002)
20. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE TMI **35**(5), 1299–1312 (2016)
21. Yosinski, J., et al.: How transferable are features in deep neural networks? In: Proceeding of NIPS, pp. 3320–3328 (2014)
22. Zhang, D., Shen, D., Alzheimer's Disease Neuroimaging Initiative: Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage, **59**(2), 895–907 (2012)