Transfer Learning for Brain Segmentation: Pre-task Selection and Data Limitations

Jack Weatheritt[1(✉)], Daniel Rueckert[2], and Robin Wolz[1,2]

[1] IXICO plc, London, UK
jweatheritt@ixico.com
[2] Imperial College, London, UK

Abstract. Manual segmentations of anatomical regions in the brain are time consuming and costly to acquire. In a clinical trial setting, this is prohibitive and automated methods are needed for routine application. We propose a deep-learning architecture that automatically delineates sub-cortical regions in the brain (example biomarkers for monitoring the development of Huntington's disease). Neural networks, despite typically reaching state-of-the-art performance, are sensitive to differing scanner protocols and pre-processing methods. To address this challenge, one can pre-train a model on an existing data set and then fine-tune this model using a small amount of labelled data from the target domain. This work investigates the impact of the pre-training task and the amount of data required via a systematic study. We show that use of just a few samples from the same task (but a different domain) can achieve state-of-the-art performance. Further, this pre-training task utilises automated labels, meaning the pipeline requires very few manually segmented data points. On the other hand, using a different task for pre-training is shown to be less successful. We then conclude, by showing that, whilst fine-tuning is very powerful for a specific data distribution, models developed in this fashion are considerably more fragile when used on completely unseen data.

Keywords: Brain segmentation · Deep learning · Transfer learning

1 Introduction

In clinical trials for neurodegenerative diseases, the progress of the disease (and efficacy of preventative treatment) is often monitored by calculating the volume of regions of interest (ROIs) in the brain [12]. For example, changes in volume of the caudates are known biomarkers for Huntington's disease (HD) [9]. The gold-standard procedure is for an expert clinician to manually delineate the region(s) of interest (ROI) and calculate the enclosing volume. However, often due to the number of patients and the amount of required follow-up scans, manual delineation is simply too costly and time consuming. Further, with multiple

B. W. Papież et al. (Eds.): MIUA 2020, CCIS 1248, pp. 118–130, 2020.
https://doi.org/10.1007/978-3-030-52791-4_10

clinicians and longitudinal data, inter- and intra-rater variability can become an issue.

A wealth of automated approaches [3] have been proposed that attempt to alleviate this issue. Notably, atlas-based algorithms [17,26], that segment anatomical regions by registering target scans to some ground-truth atlases, are widely used and perform well for many biomarkers. That said, one faces many challenges when employing such automated methodologies for clinical trials [11]. For example, some pipelines require labour-intensive manual steps (e.g. boundary shift integral analysis [8]).

More recently, deep-learning algorithms are being turned to the problem [18]. These are computationally much more efficient and reach state-of-the-art performance for many ROIs [21]. However, they are known to be sensitive to differing input data distributions [25,27], which vary due to scanner protocol, pre-processing techniques and image quality. This means that results can deteriorate on unseen data. Models can also deteriorate between cohorts that vary in pathology [5], however this is not addressed in this study.

Transfer learning, an attempt to bridge this gap, is an active area of research and currently the most popular knowledge transfer technique for MRI deep learning [6]. Transfer learning takes information from one task and, using a small amount of labelled data, makes the model generalise to a new problem. The new problem can differ by definition (i.e. the structure to be segmented), be on a new data set (differing by scanner protocol, pre-processing and subject population), or both.

Transfer learning is a promising deployment strategy on new clinical trials, for which little or no labelled data exists. This work is an investigation into such an application, in particular the choice of pre-task and the amount of labelled clinical data required to transfer information. We only know of two systematic studies for brain MRI segmentation which investigates the amount of data required and both are for 2D lesion segmentation [2,10]. Other medical imaging studies take models pre-trained on non-medical data [22,24], which is effective but inherently limited to 2D inputs. This study is for 3D anatomical delineation, where we vary the pre-task ROI and the pre-processing techniques employed. Altering the pre-processing addresses a common discussion between practitioners of medical imaging and AI. The medical imaging community tend to remove variance between images by registering to a common template [16]. This provides a strong spatial prior to the model. However, in the deep-learning community, there is a tendency to increase variability in the training process, via data augmentation, in order to generalise better to unseen data [4]. This has the advantage of not requiring registration at inference time. Further, matching pre-processing techniques between data is not always possible in a clinical setting. This often happens when building models from legacy data sets, for which the raw data may not be available. Therefore, it is informative to investigate this influence.

We take the real-world case study [14] of caudate segmentation for an HD population, as the target problem, and assess a range of pre-tasks to transfer

knowledge from, in order to boost final segmentation accuracy. We use three separate tasks: automatically generated labels from an Alzheimer's disease (AD) population [13], manually-segmented lateral ventricles and manually-segmented tissue masks. The latter two use the same HD subjects as the target caudate problem. Further, despite the promise of transfer learning for such tasks, little systematic work has been carried out investigating the amount of data required for fully 3D anatomical segmentation from MRI images.

We contribute a methodical study into the points raised above. In particular:

1. How much data is enough (end-to-end and fine-tuning) for 3D dimensional anatomical segmentation.
2. How the choice of pre-training, which varies in terms of pre-processing, augmentation and task affects results.
3. That models fine-tuned on little data are fragile when applied to new sets.
4. That utilising automated labels yields close to state-of-the-art performance with just a few manual labels.

2 Data Sets and Learning Tasks

For an overview of all the data used in this work, consult Table 1. These data sources have been labelled HDT-C, MAL-C, HDT-V, HDT-WB and OAS-C. This labelling reflects the source of the data and the ROI delineated. Therefore there are 3 unique data sources (HDT, MAL and OAS) and 3 different structures (C: caudate, V: lateral ventricles and WB: whole brain). We show the range of data sets in Fig. 1.

The overall goal of each network is to accurately predict caudate labels from the HDT data. Therefore, success is measured by calculating segmentation performance for predictions on HDT-C. OAS-C is used as a validation data set, to assess how the models generalise to unseen distributions—which raises several interesting discussion points. The remainder of the data sets are used to pre-train networks, in order to simulate a scenario where only a handful of labelled data points are available in the target domain. Consult Fig. 2 for a schematic on how all models are trained in this paper. Neural networks are named after their initial pre-task data set.

2.1 Target Task (HDT-C)

The target problem is the segmentation of the caudate from a multi-centre HD clinical trial data set. As part of the clinical trial process, the caudates were manually delineated on each subject's screening 3D T1-weighted image in MNI305 space. The data consists of 306 training subjects and 76 test subjects. This data set will be referred to as HDT-C.

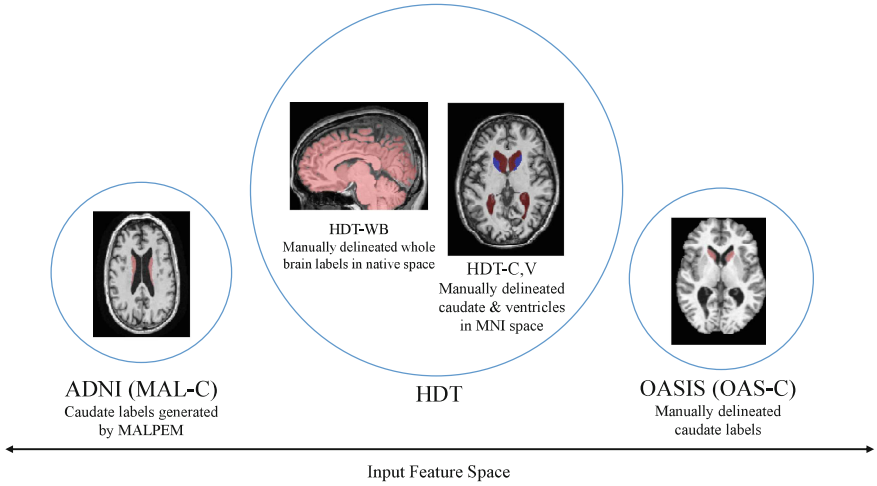

Fig. 1. The five data sets used in this study. HDT-C and HDT-V are delineated on the same T1 scans. We have schematically included a depiction of the input space, with circles grouping data sources.

Table 1. Summary of all data and pre-processing applied. *Same data, with same pre-processing and train/test split. **Same subjects (and split) with differing pre-processing.

	HDT-C*	MAL-C	HDT-V*	HDT-WB**	OAS-C
Segmented ROI	Caudate	Caudate	Ventricles	Whole brain	Caudate
Source	HD trial	ADNI (MALPEM labels)	HD trial	HD trial	OASIS
Space	MNI305	Native	MNI305	Native	MNI305
Resampled dimensions	$172 \times 220 \times 156$	$192 \times 192 \times 160$	$172 \times 220 \times 156$	$256 \times 256 \times 192$	$172 \times 220 \times 156$
Voxel spacing	$1 \times 1 \times 1$	Various	$1 \times 1 \times 1$	$1 \times 1 \times 1.2$	$1 \times 1 \times 1$
Patch size	$72 \times 112 \times 112$	$80 \times 112 \times 112$	$96 \times 136 \times 112$	$64 \times 64 \times 64$	$72 \times 112 \times 112$
Skull strip	No	No	No	No	Yes
Augmentation	Flip L/R	Flip L/R	Flip L/R	Random patch, random 90 degree rotations and flips	–

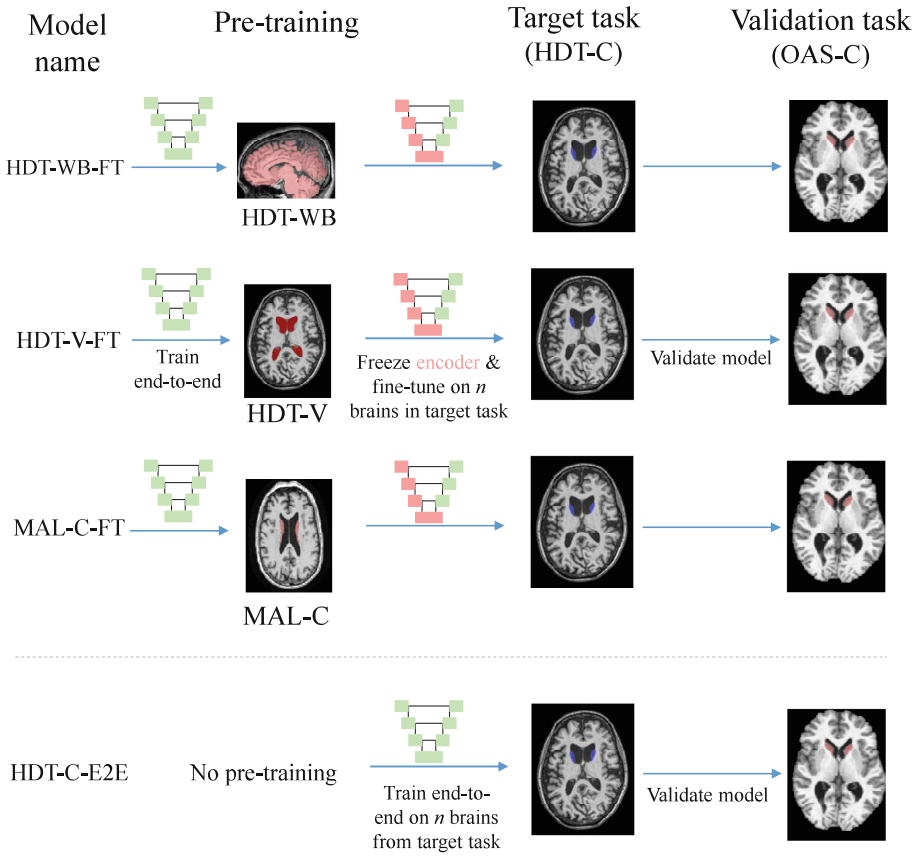

Fig. 2. Schematic of how each model is trained. The first three models are all pre-trained on a different task and then fine-tuned on a subset of the target task data set. The fourth model is not pre-trained and is trained end-to-end on a subset of the target task data set. All models are then validated on a hold out set from the target domain and finally a further validation set from an unseen data distribution. See Fig. 1, for the relationship between these data sets. In the network schematics, green implies that weights at that particular level are trainable, where as red implies those layers are frozen (see Fig. 3 for a detailed description of the network). Please refer to online version for colours. (Color figure online)

2.2 Pre-training Tasks

Caudate (MAL-C). To learn a segmentation of the caudate on a different data set represents an out-of-distribution same-task pre-training case.

We obtained data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [13] (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. For up-to-date information, see www.adni-info.org.

The caudate labels were automatically generated, using the MALPEM multi-atlas approach [17] which are available at https://github.com/ledigchr/MALPEM. We chose a random sample of 544 training subjects and 118 for testing.

Lateral Ventricles (HDT-V). For the HDT data, we utilise manual labels of the lateral ventricles (in MNI space). This is using the same input as the target task. The training/test split is kept the same, to ensure there is no cross-over of learned feature maps when fine-tuning. This data is denoted HDT-V.

Whole Brain (HDT-WB). Using the same HD subjects as HDT-C and HDT-V, we use manual delineations of the whole-brain tissue from native space. Note, the train/test split is kept the same.

2.3 Generalisation Task (OAS-C)

In order to test the generalisability of the models presented, we use a validation data set. The data consists of manual labels created by Neuromorphometrics, Inc (www.neuromorphometrics.com) for 39 images from the OASIS (www.oasis-brains.org) project.

3 Methodology

3.1 Neural Network

Architecture. We use the same network architecture for all tasks presented in this paper, a 3D UNET [7]—which is a generalisation of the now ubiquitous 2D UNET [20]. The 3D version consists of 3D convolutions, allowing for full representation of volumes and the utilisation of information in all directions. The trade-off comes with network size, as volumes require more memory. Figure 3 is a schematic of the graph structure, with feature encoder on the left and decoder on the right. The encoder consists of successive blocks of convolution, batch normalisation and ReLU activation, followed by max pooling. For the decoder, we use similar blocks, followed by a max-unpooling layer [28]. The final layer is a $1 \times 1 \times 1$ convolution, with two channels for the classification. We do not use any fully connected layers, to allow the network to be input size invariant (the only constraint being each dimension must be divisible by 8). The total number of trainable parameters 220,000.

To fine-tune this network, the encoder is frozen and only the decoder weights are allowed to change—see Fig. 3.

Optimisation. All training is run using the ADAM optimiser [15] with an initial learning rate of 0.001. The smoothing parameter (typically denoted ε) is set to 0.1, in order to avoid a vanishing denominator. The neural network was implemented using Tensorflow 1.15 [1]. All training was done on an NVIDIA RTX 2080 TI graphics card and the batch size was 2.

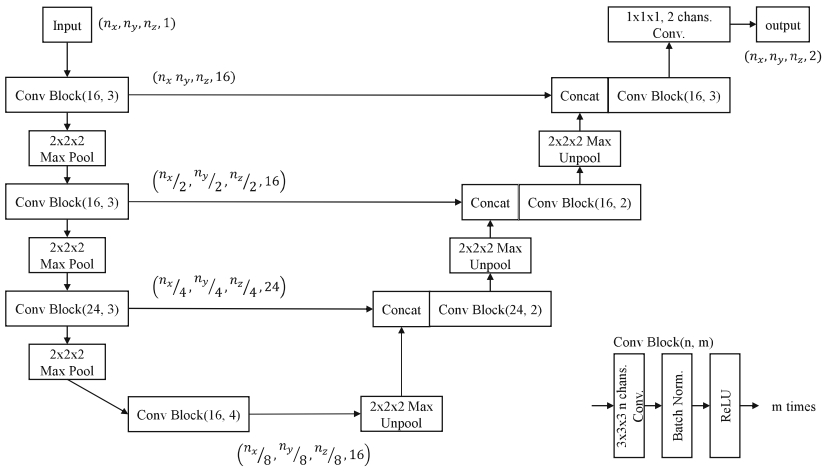

Fig. 3. Neural network architecture used throughout.

Pre-processing. All cases considered in this paper are pre-processed in different ways, to simulate differing input distributions. Table 1 highlights the pre-processing steps taken for each data set. The order of processing is: registration (if performed), bias field correction [23], resampling, z-score normalisation, patch extraction and then augmentation. For OAS-C, skull stripping happens before normalisation. The key differentiators are MNI/native space—with the target task being in MNI—and the generalisation data, OAS-C, has been skull stripped. Large portions of skull and CSF are present in the target patch; their absence in OAS-C causes high activation in these regions. This provides unseen conditions for all the neural networks considered thus far.

A network applied to the whole brain is too large to fit into GPU memory, therefore we extract a generous bounding box around the centre of mass of the training segmentations. This was sufficient to encompass all the ROIs in the respective test sets. For the whole brain task, random $(64 \times 64 \times 64)$ patches are extracted every epoch.

4 Results and Discussion

4.1 Pre-training Task Results

For each task, a model has been trained end-to-end and then tested on a hold out set. In Table 2, performance statistics for the network are presented. The mean intersection over union (IOU), also called the Jaccard index, and its standard deviation are shown. The network is able to learn the differing tasks well and with low variance between subjects. Note, the whole brain task has a relatively low IOU (when compared to state of the art), as each brain is evaluated without any overlap of patches when predicting (which is often done to boost performance [19]). The outlier is the model trained on the automated labels on the

AD data set. This can be accounted for by considering assumed inaccuracies in the automated 'ground truth' which MALPEM would make according to some random variable. This introduced noise in the labels is not inferable at test time by the network.

Table 2. Mean and standard deviation IOU on the pre-training tasks test sets.

Task	Mean (± std.)
Ventricles (HDT-V)	0.983 (±0.0149)
Whole Brain (HDT-WB)	0.940 (±0.0010)
Caudate (MAL-C)	0.837 (±0.0942)

4.2 Fine-Tuning on Caudate Task

For each set of pre-trained weights, the network is fine-tuned on a subset of the target HD data set. This allows the model to learn some in-distribution features and thus boost performance. As a comparison, an end-to-end model is also trained (from scratch) on the same amount of data—so the reader may infer the benefits of fine-tuning. Figure 4 shows plots of the same slice from a subject from the test set, highlighting the errors made by each network. The left hand column is the ground truth and the amount of training data increases as the row number does. Each network is labelled by its source data set and whether it has been trained end-to-end (E2E) or has had additional fine-tuning (FT) separately. For instance HDT-WB-FT is initially trained on the HDT-WB data for whole brain segmentation and then fine-tuned on HDT-C for caudate segmentation (see Fig. 2). It is apparent that all methods qualitatively perform worse as n decreases and that the MAL-C-FT methodology appears to make comparatively fewer mistakes. In contrast, the HDT-WB-FT method with $n = 2$, has completely missed the caudate and end-to-end training (HDT-C-E2E), which had no pre-training, similarly makes large errors. When all available data is used from the target domain, $n = 306$, all methods converge to a very similar solution.

Figure 5 plots test IOU as a function of data used from the HD set. Despite having the worst performance during pre-training, the model trained initially on MALPEM labels can reach almost state-of-the-art performance with minimal fine-tuning. This should be emphasised; with just two manually labelled brains, a neural network can be built that reliably segments the caudate. In contrast, the end-to-end model is the worst approach until 16 scans are used, at which point fine-tuning becomes redundant. For $n > 16$, all models perform similarly, as they asymptote to a mean IOU of approximately 0.93.

In terms of task choice, it is clear that pre-training by learning the caudate is preferential. With limited data, the model pre-trained using the ventricle delineations gives reasonable results. This is because initial and target tasks are in

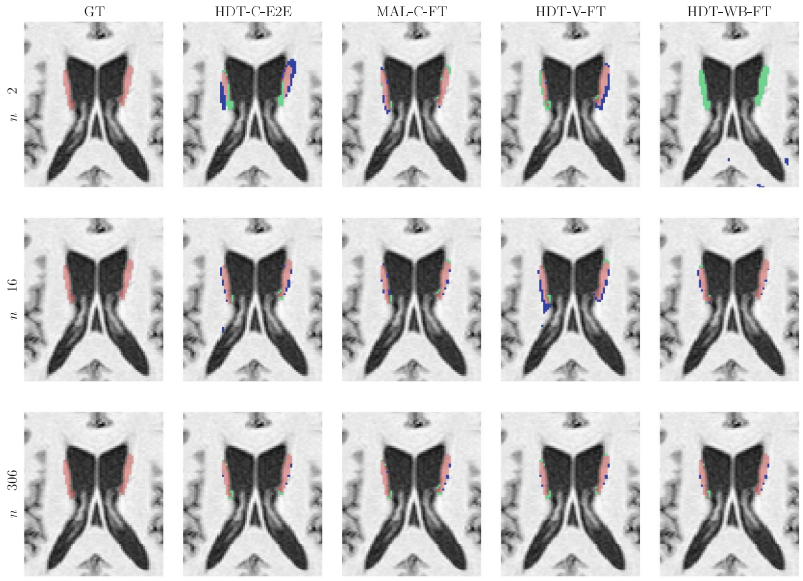

Fig. 4. Axial contours of caudate predictions (compared to ground truth, GT), as a function of n—the number of subjects in the training data. Red: true positive. Green: false negative. Blue: false positive. All plots are the same slice of the same subject. Please refer to online version for colours. (Color figure online)

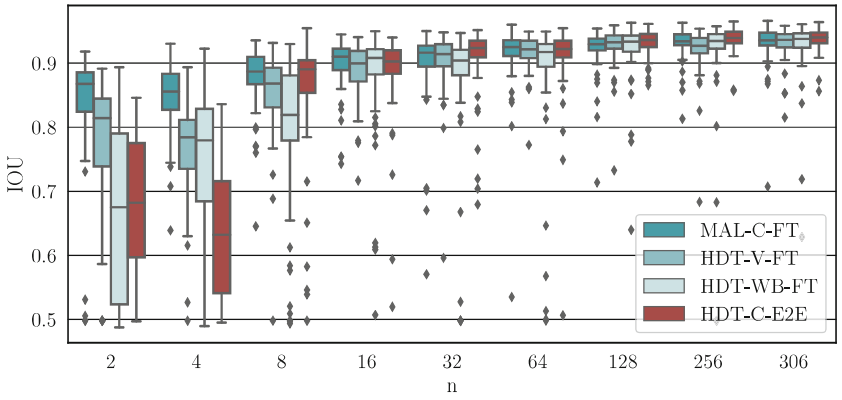

Fig. 5. Distribution of IOU scores by the amount of fine-tuning data used. Each box plot is constructed by testing on the 76 test subjects. For comparison, a model is trained from scratch using n subjects. Boxes are ordered in groups of MAL-C-FT, HDT-V-FT, HDT-WB-FT and HDT-C-E2E—please see online version for colours. (Color figure online)

the same space. On the contrary, the whole-brain model, subject to a greater level of augmentation, which should in theory be sensitive to a richer feature set performs no better than the model without any pre-training.

These three pre-training tasks highlight that choosing the same task on out-of-distribution data outweighs either a related task using the target data, or a task that maximises the variance in previously acquired knowledge (through augmentation and required labelling of the whole brain). This means that, surprisingly, even though the ventricle task contains the strong spatial prior provided by MNI registration and the ventricles are neighbouring structures, it is less important than learning the same structure from a different disease population, in non-registered native space.

4.3 Generalisation

The above conclusions are relevant to the problem of maximising performance on a given data set, but the question remains regarding generalisation to new data distributions. Using the OAS-C data (Sect. 2), we can ascertain the generalisability of all models considered thus far. OAS-C has been previously brain masked, which provides a significantly different input distribution. Not only because z-normalisation acts on the whole domain, but large portions of skull and CSF were present in the target patch. This provides unseen conditions for all the neural networks considered thus far.

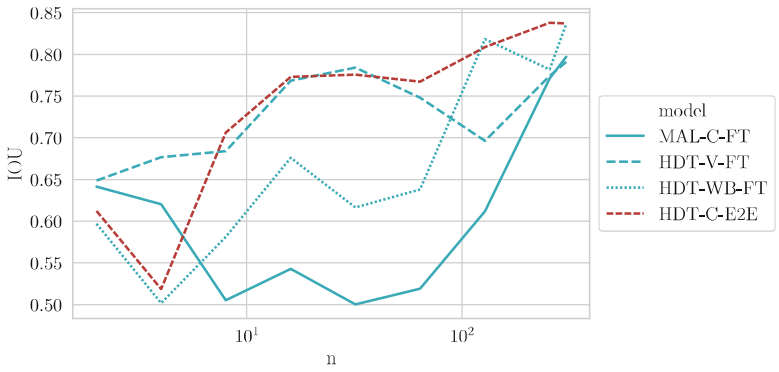

Fig. 6. Mean IOU on the OAS-C data set, as a function of the number of fine-tuning subjects (n) used. In the case of HDT-C-E2E, n is the amount of data used in totality.

In Fig. 6, the mean IOU is plotted for each method applied to OAS-C, as a function of the number of subjects in the fine-tuning set. It becomes highly apparent that the models fine-tuned on a few scans do not generalise well, despite performing well on the target task (Fig. 5). This implies that the connection between the frozen part of the network and the fine-tuned layers is very fragile. This is corroborated by the HDT-V-FT method performing reasonably well for

small n, as the early layers are 'seeing' data more like their original training case (HDT-V, which is in MNI space too).

Therefore, whilst task selection was most important for tuning to a specific data set, for complete generalisation, the choice of pre-processing appears more influential. The cases that have been initially trained in MNI space, like OAS-C, perform better. This is especially contrasted against the MAL-C-FT, which cannot generalise until $n > 200$—at which point the encoder is probably learning an independent representation without utilising pre-trained features in the decoder. The other fine-tuned models begin to generalise as n increases, implying exposure to more examples inherently improves the connection between the two components of the network. It is worth emphasising that performance on the HDT-C data (Fig. 5) plateaus before performance on the OAS-C data (Fig. 6), implying that the increase in HDT-C samples has a hidden benefit on future predictions on unseen data.

5 Conclusions and Extensions

This work has shown that, when few in-distribution labelled scans are available, it is possible to construct a neural network that gives state-of-the-art performance. This is particularly encouraging for deploying neural networks in clinical trials. In such a scenario, it appears sensible to first generate automated labels on a different data set and then fine-tune using a few manually labelled scans (say screening patients). Such a pipeline can be rapidly deployed across trials.

Fine-tuning a model trained on the same task was far more beneficial than matching the input distribution of the data. That said, on the flip side, fine-tuning appears to do little for generalisation to a new distribution. This is at least the case for the problem considered here, whereby using a co-registered pre-task data set, yielded the strongest generalisation performance. Given the generalisation results, it appears that matching the input distribution, at all stages in the pipeline, is advisable and it is not sufficient to fine-tune a network which expects a differing distribution.

Several avenues should now be explored. An investigation into whether a fine-tuned model initially trained in MNI space would yield improvements over one from native space (e.g. MAL-C-FT) should be carried out. Secondly, a demonstration of this model generalising to new data would be promising. Finally, whilst we have elucidated impacts of task selection and pre-processing methods, a full parameter sweep could be devised that fully separates task, disease population and pre-processing technique.

Acknowledgements. Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. Data collection and sharing for this project was

funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012).

References

1. Abadi, M., et al.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). software available from tensorflow.org https://www.tensorflow.org/
2. Alex, V., Vaidhya, K., Thirunavukkarasu, S., Kesavadas, C., Krishnamurthi, G.: Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation. J. Med. Imaging **4**(4), 041311 (2017)
3. Balafar, M.A., Ramli, A.R., Saripan, M.I., Mashohor, S.: Review of brain MRI image segmentation methods. Artif. Intell. Rev. **33**(3), 261–274 (2010). https://doi.org/10.1007/s10462-010-9155-0
4. Bowles, C., et al.: GAN augmentation: augmenting training data using generative adversarial networks. arXiv preprint arXiv:1810.10863 (2018)
5. Brusini, I., Lindberg, O., Muehlboeck, J.S., Smedby, Ö., Westman, E., Wang, C.: Shape information improves the cross-cohort performance of deep learning-based segmentation of the hippocampus. Front. Neurosci. **14**, 15 (2020)
6. Cheplygina, V., de Bruijne, M., Pluim, J.P.: Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. Med. Image Anal. **54**, 280–296 (2019)
7. 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
8. Freeborough, P.A., Fox, N.C.: The boundary shift integral: an accurate and robust measure of cerebral volume changes from registered repeat MRI. IEEE Trans. Med. Imaging **16**(5), 623–629 (1997)
9. Georgiou-Karistianis, N., Hannan, A.J., Egan, G.F.: Magnetic resonance imaging as an approach towards identifying neuropathological biomarkers for Huntington's disease. Brain Res. Rev. **58**(1), 209–225 (2008)
10. 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
11. Giorgio, A., De Stefano, N.: Clinical use of brain volumetry. J. Magn. Reson. Imaging **37**(1), 1–14 (2013)
12. Henley, S.M., Bates, G.P., Tabrizi, S.J.: Biomarkers for neurodegenerative diseases. Curr. Opin. Neurol. **18**(6), 698–705 (2005)
13. Jack Jr., C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging Off. J. Int. Soc. Magn. Reson. Med. **27**(4), 685–691 (2008)
14. Johnson, E.B., et al.: Recommendations for the use of automated gray matter segmentation tools: evidence from Huntington's disease. Front. Neurol. **8**, 519 (2017)
15. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

16. Krivov, E., Pisov, M., Belyaev, M.: MRI augmentation via elastic registration for brain lesions segmentation. In: Crimi, A., Bakas, S., Kuijf, H., Menze, B., Reyes, M. (eds.) BrainLes 2017. LNCS, vol. 10670, pp. 369–380. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75238-9_32
17. Ledig, C., Schuh, A., Guerrero, R., Heckemann, R.A., Rueckert, D.: Structural brain imaging in Alzheimer's disease and mild cognitive impairment: biomarker analysis and shared morphometry database. Sci. Rep. **8**(1), 1–16 (2018)
18. Litjens, G., et al.: A survey on deep learning in medical image analysis. Med. Image Anal. **42**, 60–88 (2017)
19. Milletari, F., et al.: Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound. Comput. Vis. Image Underst. **164**, 92–102 (2017)
20. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
21. Roy, A.G., Conjeti, S., Navab, N., Wachinger, C., Initiative, A.D.N., et al.: QuickNAT: a fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage **186**, 713–727 (2019)
22. Shin, H.C., et al.: Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
23. Sled, J.G., Zijdenbos, A.P., Evans, A.C.: A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging **17**(1), 87–97 (1998)
24. Tajbakhsh, N., et al.: Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging **35**(5), 1299–1312 (2016)
25. Weese, J., Lorenz, C.: Four challenges in medical image analysis from an industrial perspective. Med. Image Anal. **33**, 44–49 (2016)
26. Wolz, R., Aljabar, P., Hajnal, J.V., Hammers, A., Rueckert, D., Initiative, A.D.N., et al.: LEAP: learning embeddings for atlas propagation. NeuroImage **49**(2), 1316–1325 (2010)
27. Zavala-Romero, O., et al.: Segmentation of prostate and prostate zones using deep learning. Strahlentherapie und Onkologie (2020). https://doi.org/10.1007/s00066-020-01607-x
28. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 818–833. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_53