Transfer Learning for Focal Pathology Segmentation across Neuro-degenerative Diseases

Barleen Kaur

Computer Science
McGill University, Montreal

August 4, 2020

Acknowledgements

Firstly, I would like to extend my sincere gratitude towards my supervisors, Prof. Tal Arbel and Prof. Doina Precup, for having faith in me. Their valuable guidance, support and constant encouragement helped me steer the progress in this work in the right direction. Prof. Arbel has remarkably helped me in polishing the design of the methods and framing the bigger picture through timely meetings. Prof. Precup has been an incredible source for machine learning insights and technical expertise. Overall, I'm deeply admired by their leadership qualities and commitment to the research community.

I would also like to express my deepest gratitude to my parents for all their sacrifices. Their constant love and motivation dwarfed the long-distance spanning thousands of kilometres between us and the considerable time gap. I would also like to express my special thanks to Nishanth for his continuous support, care and wisdom. He has been my pillar of strength in all ups and downs of my grad school life.

I also thank my collaborators Paul Lemaitre, Nazanin Mohammadi and Raghav Mehta, with whom I have worked on some of the ideas presented in this thesis. I'm incredibly thankful to Compute Canada, Mila resources and IT support at CIM lab, without whom no experimentation could have been possible to run. I am also grateful to NeuroRx Research and Prof. Douglas L. Arnold for providing a large, clinical dataset that was used in this thesis. The research presented in this thesis was supported by the Canadian Natural Science and Engineering Research Council(NSERC) Collaborative Research and Development grant, and Synaptive Medical. I would also like to thank all members of the Probabilistic Vision group: Brennan, Justin, Joshua, Eric and Amar for creating a positive environment suitable for carrying out research in the lab. I have had interesting conversations with them related to work matters or events happening all across the globe. They gave me moments that I will cherish throughout my life.

Finally, I would like to thank my brother Karan, my relatives and all my friends, especially Anupama and Shashank, for all the laughter and fun during stressful times.

Abstract

This thesis presents an exploration of a variety of transfer learning techniques in the context of deep networks for the segmentation of focal pathological structures across neurodegenerative diseases. Deep learning frameworks have been successful in achieving state-of-the-art performance on a variety of public medical imaging datasets for pathology segmentation. To do so, deep learning models need large amounts of labelled training data, as they contain a huge number of parameters. However, in medical imaging applications, finding a large amount of annotated data is hard. This hinders the performance of these data-hungry deep learning based frameworks in the medical field. Recently, transfer learning has been shown to be effective in dealing with the challenges related to small data regimes in many medical imaging applications. In this thesis, we explore various transfer learning strategies like fine-tuning a pre-trained network, multi-task joint representation learning using a double head network and a novel coupling of the above methods with cascaded networks, specifically for knowledge transfer across neurodegenerative diseases. In addition to this, we also analyze several ways of fine-tuning a pre-trained source network to determine the best way for carrying out knowledge transfer via fine-tuning in the context of medical image segmentation. We evaluate these approaches by leveraging a large, proprietary, multi-scanner, multi-center, clinical trial MRI dataset of 1385 patients with Multiple Sclerosis (MS) in order to improve performance on a multi-class brain tumour sub-tissue segmentation task for which a much smaller, public dataset is available (from the MICCAI BraTS 2018 challenge). We contrast these methods against the baseline method of training a deep network with brain tumour data from scratch. We also present a comprehensive analysis of these methods by varying the number of samples in the target brain tumour set available for training. We obtain $3-61\%$, $3-11\%$, and $0.4-5\%$ relative improvements over the baseline on the target segmentation performance (measured through Dice scores) using transfer learning techniques, when using small, medium and large amounts of training data in the target task respectively.

Résumé

Cette thèse présente une exploration d'une variété de techniques d'apprentissage par transfert appliquées à des maladies neurodégénératives dans le contexte de réseaux profonds pour la segmentation des structures pathologiques focales. Les cadres d'apprentissage profond ont réussi à atteindre des performances de pointe sur une variété d'ensembles de données d'imagerie médicale publiques pour la segmentation de pathologies. Pour ce faire, les modèles d'apprentissage profond ont besoin de grandes quantités de données étiquetées durant la phase d'entraînement car ils contiennent un grand nombre de paramètres. Cependant, dans les applications d'imagerie médicale, de grands ensembles de données annotées sont difficiles à trouver, ce qui entrave les performances de ces cadres basés sur l'apprentissage profond dans le domaine médical. Récemment, l'apprentissage par transfert s'est révélé efficace pour adresser les défis liés à la taille restreinte des ensembles de données dans de nombreuses applications d'imagerie médicale. Dans cette thèse, nous explorons diverses stratégies d'apprentissage par transfert comme le réglage fin de réseau pré-entraîné, l'apprentissage de la représentation conjointe multi-tâches en utilisant un réseau à deux têtes et un ainsi que le couplage des méthodes ci-dessus avec des réseaux en cascade, spécifiquement pour le transfert de connaissances entre plusieurs maladies neurodégénératives. En plus de cela, plusieurs façons de régler avec précision un réseau pré-entraîné sont analysées pour déterminer la meilleure approche pour effectuer un transfert de connaissances via un réglage fin dans le contexte de la segmentation d'images médicales. L'évaluation de ces approches se fait en tirant parti d'un ensemble de données d'IRM d'essais cliniques de grande taille, multi-scanner et multicentrique, de 1385 patients atteints de sclérose en plaques (SEP) dans le but d'améliorer la performance d'une tâche de segmentation de sous-tumeur cérébrale de 4 classes pour laquelle un ensemble de données public beaucoup plus petit est disponible (défi MICCAI BraTS 2018). Ces méthodes sont comparées avec la méthode de base de formation d'un réseau profond avec des données sur les tumeurs cérébrales entraîné sans transfert de connaissances. Nous présentons également une analyse complète de ces méthodes en faisant varier le nombre d'échantillons dans l'ensemble de tumeurs cérébrales cibles disponibles pour la formation. Nous obtenons des améliorations relatives de 3 à 61%, de 3 à 11% et de 0,4 à 5% sur la performance de la tâche de segmentation ciblée (mesurée

par les scores de Dice) en utilisant des techniques d'apprentissage par transfert comparé à la performance de base lors de l'utilisation de petites, moyennes et grandes quantités de données d'entraînement dans l'ensemble cible défini respectivement.

Contribution of Authors

This thesis presents an exploration of a variety of deep transfer learning techniques across neurodegenerative diseases for the segmentation of focal pathological structures. The work presented in chapter 3 is published in the MICCAI 2019 Workshop on Domain Adaptation and Representation Transfer. This is joint work with Paul Lemaitre, Nazanin Mohammadi and Raghav Mehta. Both Paul and Raghav participated in developing the idea. Nazanin and Raghav also contributed to the implementation of parts of the methods initially, while Paul helped with pre-preprocessing of the data.

Contents

List of Figures

List of Tables

1

Introduction

Owing to the success of deep learning in various computer vision applications, the potential impact of these techniques in medical image analysis is enormous. Much research has been done to successfully apply deep learning based approaches to medical image segmentation problems in the brain [42, 94], prostate [144], pancreas [105], and many more applications, with the obtained results outperforming traditional methods. However, the shortage of large, annotated patient imaging datasets available for training hinders the performance of existing deep learning frameworks in medical image segmentation. There are numerous challenges associated with obtaining big datasets. First, acquiring medical imaging data is slow and expensive. Moreover, labelling the data is laborious and dependent on the availability of qualified clinical experts. Datasets for patients with rare diseases are inherently small. Finally, patient consent is not always available, and privacy considerations limit the possibility of making the data public. As a result, the majority of datasets available publicly are small. Large datasets exist, but many are proprietary. Therefore, finding ways to leverage knowledge from large source datasets, if available, might be helpful for other, relatively smaller medical datasets of patients with different diseases.

It has already been shown that transfer learning can be effective in dealing with small data regimes in a wide variety of applications related to natural images [100, 37, 24, 29]. It has become a common practice to use deep Convolutional Neural Network (CNN) models pre-trained on a large ImageNet dataset [27] as a starting point for solving other computer vision tasks based on natural images. In contrast, working with medical images presents more unique challenges. First,

medical images can be 2D, 3D, or even 4D (spatio-temporal) in nature. Second, different anatomical structures or pathologies (e.g. tumours, lesions) can occur non-uniformly, with varying shapes and sizes across patients. Third, pixel intensities can vary across different modalities (MRI, CT) and image acquisition protocols or scanners. This makes it hard to delineate the boundaries of pathological structures accurately. Other factors include motion artifacts, missing edges and low resolution, which increase the complexity of the problem.

Transfer learning has also been explored in many healthcare applications, such as the classification of skin lesions [78] and brain lesion segmentation [36]. Typically, a deep network is trained using a large-scale source dataset. This pre-trained network [19] can then be used in two ways: (1) to extract off-the-shelf features for target datasets [51], or (2) as initialization for further fine-tuning on the target dataset [125]. For example, knowledge could be transferred from medical and/or non-medical datasets like ImageNet [27] to improve classification results in other target medical applications [78, 145]. In the context of medical image segmentation, researchers have explored how fine-tuning can be used to improve performance in segmenting pathologies when data is acquired from multiple scanners [36] or in transferring knowledge from one grade of a disease to another grade [1]. For example, authors in [86] trained a single CNN to perform different tasks, such as tissue segmentation in brain MRI, pectoral muscle segmentation in breast MRI and cardiac CTA segmentation. They found that this single network performed comparably to a CNN model trained for each task individually. More recently, [147] trained a set of models in a self-supervised way and showed how this approach could be leveraged to improve segmentation results across multiple diseases (lung nodule, brain tumour, liver diseases) and multiple modalities (e.g. CT, X-ray, MRI).

In this thesis, we explore transfer learning techniques for the task of segmenting pathological structures across different neurodegenerative diseases, where the imaging modalities are similar (e.g. brain MRI), but the structures of interest differ substantially (e.g. lesions, tumours), and the task can vary from binary to multi-label classification. The intuition is that since the imaging context in both the diseases is similar, i.e. brain images, they share similar features. Hence, segmentation results on the smaller target dataset should improve after leveraging the representation learned by the network pre-trained on the large source dataset, as shown in Figure 1.1. Specifically, we explore

different methods for leveraging a large, proprietary, clinical trial dataset of MRIs acquired from patients with Multiple Sclerosis (MS), intending to improve the results for brain tumour sub-tissue segmentation on a much smaller dataset. Our methods include: fine-tuning the pre-trained network, jointly learning representations for both datasets using multi-task learning, and novel variations of cascaded networks coupled with transfer learning techniques where sub-networks can be trained in sequence or together. We also present a comprehensive study of these methods by varying the number of samples in the target set available for training. Our approach shows great potential when the target dataset is very small. We will now outline the main aspects of our approach, as well as the content of the thesis.

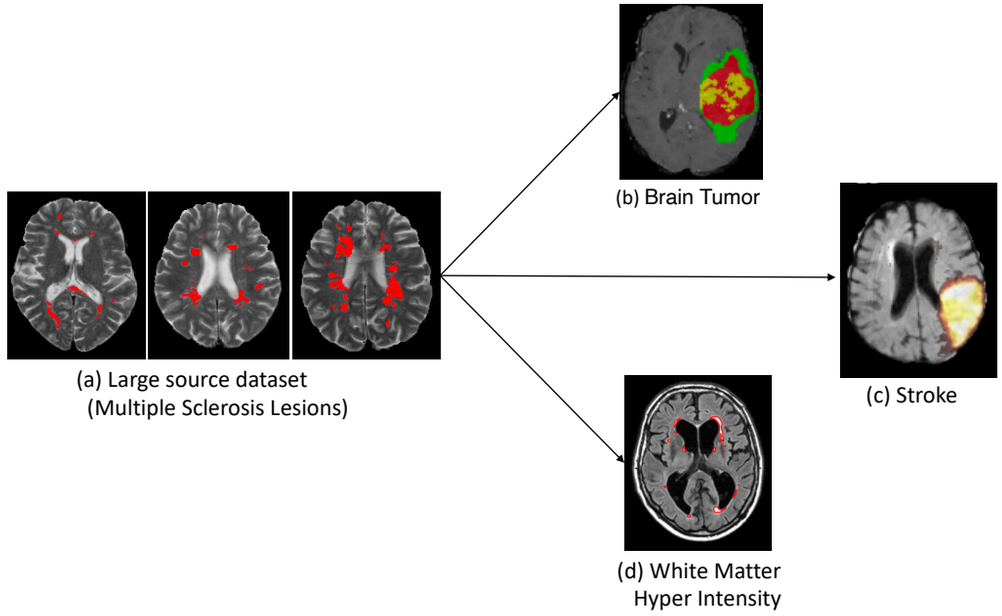


Figure 1.1: Knowledge transfer from models pre-trained on a large annotated dataset of patients having Multiple sclerosis to other relatively smaller datasets of patients having different neurological diseases like brain tumours [9], brain stroke [72], and white matter hyperintensity [61]. Image courtesy of NeuroRx Research for part (a).

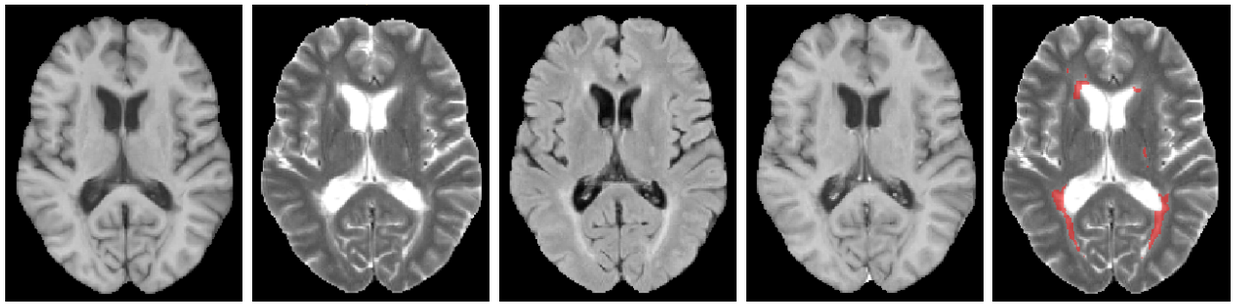
1.1 Multiple Sclerosis

Multiple Sclerosis is a chronic neurological disease characterized by inflammatory demyelination of the Central Nervous system (CNS), which consists of the brain and the spinal cord [41]. It is caused when the body's autoimmune system starts to attack the myelin sheath, a protective layer insulating

the axons in the neuron, which is responsible for the proper transmission of electrical signals. As a consequence, lesions are formed in the affected areas, leading to loss of activity in the central nervous system. These lesions can develop at multiple locations in the CNS, hence the name Multiple Sclerosis (MS). In 2013, 2.3 million people were estimated to be affected by MS globally, with Canada having the highest prevalence rate of 291 per 100,000 people [16]. The risk of developing MS is about 2-3 times higher for women [82] than men. The most common first signs of MS may include weakness in limbs, difficulty balancing the body, vision problems and heat sensitivity. There are primarily four patterns of progression in MS [69]: Clinically Isolated Syndrome (CIS), Relapsing-remitting MS (RRMS), Primary Progressive MS (PPMS) and Secondary Progressive MS (SPMS). CIS is the initial stage, in which a person experiences neurological episodes lasting for at least 24 hours. Around 30-70% of patients who enter the CIS stage develop MS later. More than 80% of patients with confirmed MS after the CIS stage, are diagnosed with RRMS. RRMS is characterized by periods of attacks, called relapses, that can last for months, followed by periods of remission during which symptoms may disappear partially or altogether. At some point, symptoms can become permanent. The remaining 20% of patients with confirmed MS are diagnosed with PPMS. Such patients suffer from the progression of disability right from the onset of the disease, with little to no remission. Around 65% of patients who experience RRMS progress later to the SPMS stage. Although there are treatments available for controlling and slowing down the progression of the disease, no cure exists to date.

Usually, the diagnosis is made by neurologists who check a patient's medical history and conduct a neurological examination, followed by confirmation of the disease after performing specific tests like MRI scanning. Lesions in MRIs are one of the hallmarks of the disease. Lesions appear hypo-intense in T1w MRI sequences in comparison to white matter and grey matter, while they appear hyper-intense in T2w and FLAIR sequences. An example of all the sequences along with expert level T2 lesion ground truth is shown in Figure 1.2.

Experts use MRI scans to manually segment MS lesions, which is an important task for performing proper diagnosis, and evaluating disease progression, activity and treatment response. This process is laborious and time-consuming, because millions of voxels need to be manually examined by



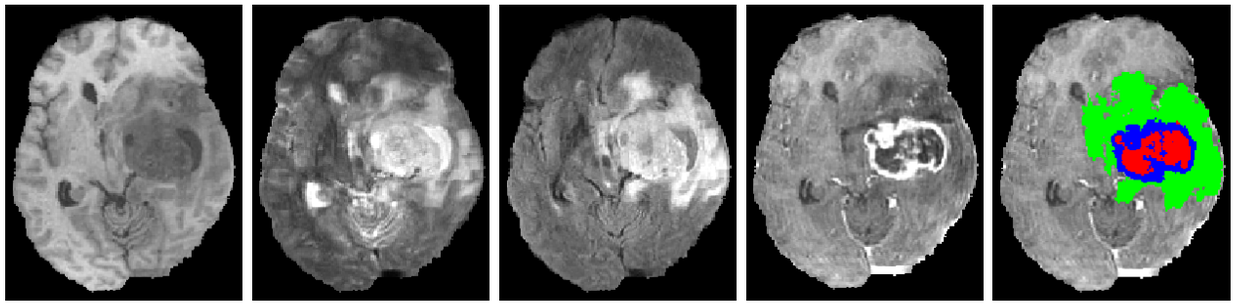
(a) T1w (b) T2w (c) FLAIR (d) T1p (e) Ground Truth

Figure 1.2: An example of different MRI sequences (a)-(d) along with expert labelled lesion labels (e) overlayed on T2 modality in MRI of a MS patient. Image courtesy of NeuroRx Research.

the expert. This manual work is also subject to inter-rater variability, as segmentations can vary across annotators, as well as intra-rater variability, because the same expert can produce different segmentations at different times. This inconsistency motivates the development of automated segmentation tools that can produce consistent segmentations across time. However, the automatic segmentation of MS lesions from a patient's MRI is a challenging task due to many factors. First, MS lesions vary substantially in terms of shape, size, and texture from one patient to another. Multiple lesions may be present at different sites in a patient's brain. Second, lesions are generally present in a small proportion of the brain, causing extreme class imbalances. Other factors like sensor noise, motion artifacts, missing edges and low resolution make this problem even more challenging. In order to deal with these difficulties, researchers have proposed many machine learning-based algorithms for this task, which can be broadly classified into two kinds: traditional machine learning-based approached and deep learning-based approaches. Traditional machine learning methods include, for example, modelling the distribution of healthy and non-healthy tissue separately [32], using graphical models for incorporating spatial context between neighbouring tissue voxels [123] and engineering feature extractors for constructing complex models of the tissue classes [122]. Deep learning based approaches [132, 130, 15] rely on deep neural networks, and often outperform traditional methods because of their ability to combine the feature engineering and classification, thus obtaining features that are automatically tailored for the classification task.

1.2 Brain Tumours

Brain tumours are caused when cells in the brain stop growing normally, leading to the formation of tumours [117]. Brain tumours can be benign or malignant (cancerous) and can be categorized into two types: Primary or Secondary [118]. Primary brain tumours originate in the brain, whereas secondary brain tumours originate in other parts of the body but spread to the brain. Based on factors like the rate of growth of tumours as well as how probable they are to spread to surrounding tissue, tumours can be divided into two grades: Low-grade and High-grade [117]. Low-grade tumours are non-cancerous and unlikely to grow or to infiltrate surrounding tissue. High-grade tumours are cancerous, proliferate and spread to other parts of the CNS. Low-grade tumours can also transform into high-grade tumours. Gliomas [117] are the most common high-grade primary brain tumours, which originate in the glial cells. These can be of three types: astrocytoma, oligodendroglioma and ependymomas. Meningiomas, on the other hand, originate in meninges. Brain tumours affect around 250,000 people worldwide every year [120]. Specific types of tumours are more likely to develop at a certain age. A tumour can manifest through symptoms like headaches, vision problems, seizures, memory loss, difficulty speaking and walking [118]. Once symptoms occur, the diagnosis begins with a neurologist who browses through a patient's medical history and performs various neurological tests and other tests to check muscle strength, eyes, balancing and memory. After that, the neurologist may conduct further tests, typically including imaging of a patient's brain using CT, MRI or PET scans.



(a) T1 (b) T2 (c) FLAIR (d) T1c (e) Ground Truth

Figure 1.3: An example of different MRI sequences (a)-(d) along with ground truth segmentation mask (e) overlayed on T1c modality in BraTS 2018 Training dataset [79].

These scans help radiologists in delineating brain tumours from other healthy parts of the brain. For example, high-grade gliomas are hyper-intense in T1c sequences while edema, the swelling around the tumour, is hyper-intense in T2w and FLAIR sequences. An example of all the MRI sequences along with ground truth segmentation labels for a tumour is shown in Figure 1.3.

The treatment of the tumour depends on various factors such as the type, grade, location, size of the tumour, age of the patient, and medical history. Surgery is usually performed to remove the tumour entirely or partially [118]. For cases in which performing surgery is not feasible or needed, radiation therapy and chemotherapy are used. For surgery based treatments, it is essential to detect and segment the tumour and its sub-structures accurately. Experts examine the patient's MRI to segment different sub-structures of brain tumours manually. This process suffers from similar challenges as in the case of MS lesion segmentation. In order to address those challenges, several automatic brain tumour segmentation methods have been developed by researchers over the years. The overall goal of all these segmentation approaches is to detect the location and delineate the tumour regions in the brain from the healthy tissue. Figure 1.3(e) contains an example delineating the active part of the tumour (in blue), the necrotic core (in red) and the edema (swelling around the tumour, in green). As in MS lesion segmentation, both traditional machine learning and deep learning approaches have been used for tumour segmentation. Some of the traditional machine learning methods focus on distinguishing tumour regions from healthy tissue based on differences in local intensity of the pixels and textual patterns [12] or by aligning MRI scans to a healthy atlas and then classifying tumours as outliers [99], or by building generative, probabilistic models for tissue distribution [80]. Deep Learning methods, on the other hand, learn highly discriminative features from the data without the need for providing any hand-crafted or pre-defined features. Particularly, Convolutional Neural Network-based methods have been leading in this field, with promising results [42, 50, 54, 57]. In order to foster research in this domain, the brain tumour segmentation challenge, BraTS [9], is organized every year since 2012, in conjunction with MICCAI conference, allowing for objective comparison of different brain tumour segmentation methods on standardized datasets.

1.3 Outline of the Proposed Method

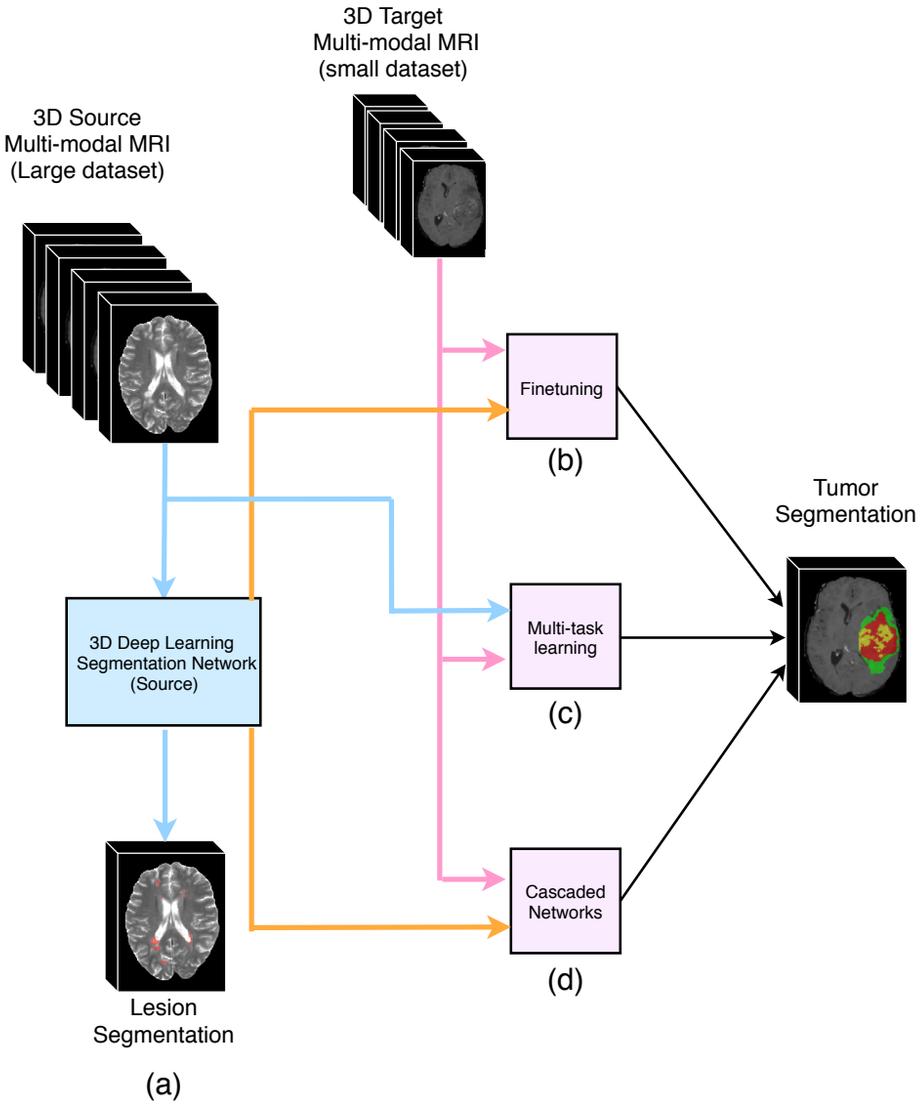


Figure 1.4: Flowchart describing the proposed transfer learning approach from models pre-trained on a large annotated dataset of MS patients (in part (a)) to multi-class brain tumour segmentation on a relatively smaller brain tumour target dataset via fine-tuning, multi-task learning and cascaded networks in part (b), (c) and (d) respectively.

In this thesis, we aim to address the challenges faced by the medical community due to the lack of large annotated datasets in the context of medical image segmentation. In order to tackle this problem, we propose and test several transfer learning methods which can be used on datasets coming from different neurodegenerative diseases. The first objective is to explore different ways of leveraging knowledge obtained from deep networks trained on a large source dataset in order to

improve segmentation performance on a relatively smaller target dataset of patients with a different disease. The second objective is to provide a comprehensive analysis of the impact of the target dataset size on the performance of each transfer learning method. The source dataset used for both objectives is a large, multi-site, multi-scanner, multi-sequence clinical trial MRI image dataset of 1385 patients with Relapsing-Remitting Multiple Sclerosis (RRMS), as well as expert-labelled T2 lesion masks for segmenting binary T2 lesions. The target datasets are subsets of the MICCAI 2018 BraTS dataset [79], which contains a total of 285 annotated brain tumour samples in the training set and 66 unlabelled samples in the validation set.

For the first objective, we approach the problem using three different techniques that have previously proven to be effective for similar tasks. A flow chart giving a high-level view of the methods is shown in Figure 1.4.

1. *FINE-TUNING*: We explore the effectiveness of different fine-tuning techniques for the task of segmenting pathologies across different diseases, as this approach has been shown in prior work to improve segmentation results on data acquired from multiple scanners [36] and knowledge transfer between different grades of a disease [1]. To achieve this aim, a 3D MS lesion segmentation network is first pre-trained using the MS dataset. This network is then fine-tuned for the multi-class brain tumour segmentation task. Furthermore, we evaluate whether fine-tuning just the last few layers works as well for medical images as it does for natural images, by varying the number of fine-tuned layers.

2. *MULTI-TASK LEARNING*: The second approach that we investigate is a multi-task learning method in which a 3D segmentation network with two decoders, one for each dataset, is trained to learn a representation jointly from both datasets. Previous works [86, 14] have shown that sharing features across different tasks leads to better representation learning, resulting in features that are inherently more task-invariant and generic. This has been beneficial in a variety of computer vision domains [28, 84]. Moreover, multi-task learning allows us to train a single network for handling multiple tasks rather than separate networks for individual tasks.

3. *CASCADED NETWORKS*: Cascaded networks have proven useful in various medical image

segmentation applications [134, 42, 20], as they divide the overall task into a series of sub-tasks and build on these sub-tasks. This motivated us to explore cascaded networks in the context of deep transfer learning. We propose novel variations of cascaded networks by augmenting the pre-trained MS network or pre-trained joint head network with another 3D UNet and then training these sub-networks either simultaneously or sequentially. Note that the multi-task learning approach assumes that both the source and the target datasets are available at training time. In contrast, the other approaches work even if only a pre-trained source network is available.

For the second objective, small, medium and large amounts of target brain tumour data are used for training in order to provide a more comprehensive comparison between the performance of different methods. Our quantitative and qualitative results show how different methods perform in comparison to the baseline method of training the network from scratch on the target task, as the number of target samples available for training varies. This comprehensive analysis is crucial in order to understand if our approach can be applied in settings where the quantity of data available for training is limited, in particular, to improve segmentation results on smaller public datasets, real-time clinical data, data belonging to patients having rare diseases, and in analyzing similar data from several hospitals.

1.4 Contributions

The work presented in this thesis contributes to the field of medical image segmentation for applications in which only small amounts of data are available, as follows:

1. **Exploring different transfer learning techniques across neurodegenerative diseases in the context of medical image segmentation.** Transfer learning techniques have been studied extensively in computer vision applications. Yet, their integration into medical image segmentation applications is still an ongoing research. This thesis explores a wide variety of transfer learning techniques across neurodegenerative diseases, including fine-tuning a pre-

trained network, joint learning of multiple tasks and proposes a novel coupling of these two transfer learning methods with cascaded networks. The methods are evaluated using real-world datasets of patients with MS as the source dataset and brain tumours as the target dataset, which presents several challenges: differences in the structures of interest (lesions vs tumours), the nature of task inference (binary to multi-label classification), and the dataset acquisition. The methods we propose and test are general and can be applied to other medical image segmentation tasks as well.

2. **A quantitative and qualitative analysis of different transfer learning techniques as a function of the size of the target dataset.** We evaluate the performance of all the methods using small, medium and large amounts of target brain tumour data. A comprehensive comparison of different methods against the baseline method of training a network from scratch with different amounts of target data is also presented for better insight. Our methods produce the highest performance gain when the amount of target data is small to medium. These results are promising since it suggests that our methods can be applied to improve performance on other small medical imaging datasets.

1.5 Outline of Thesis

This thesis presents an exploration of several transfer learning techniques across neurodegenerative diseases for segmenting pathological structures of interest. It is structured as follows.

Chapter 2 introduces the concept of deep learning and explains the fundamentals of convolutional neural networks along with some tricks for training these networks. The problem of image segmentation is discussed, where the goal is to label every pixel in the image with its corresponding class. We then present a literature review of traditional machine learning and deep learning methods for medical image segmentation. Afterwards, we discuss transfer learning in-depth, covering its definition and its applications in computer vision. Finally, several recent deep transfer learning based methods are reviewed in the context of medical image segmentation.

Chapter 3 discusses different fine-tuning strategies employed for transfer learning across neurological diseases. We present details about the source MS lesion and target brain tumour datasets used for evaluating the proposed methods, along with the pre-processing steps that we took. We present the architecture of the 3D UNet along with its training procedure and the metric used for evaluating the performance of the trained networks. Quantitative and qualitative results illustrate the effectiveness of the proposed methods.

Chapter 4 explores a variety of transfer learning techniques used for leveraging knowledge from the large MS source dataset in order to improve segmentation results on the smaller BraTS dataset. The different methodologies, including fine-tuning, double-head networks for multi-task learning, and novel variations of cascaded networks, are discussed in detail. We then describe our experimental pipeline, as above, and a comparison of the different methods as a function of the amount of data from the target dataset, which is available during training.

Chapter 5 summarizes the key ideas and contributions and discusses ideas for increasing the applicability of this work to other relevant domains, as well as other avenues for future work.

2

Background and Literature Review

This chapter provides a review of relevant literature related to the task of transfer learning in medical image segmentation. First, an overview of important concepts in the field of deep learning is presented. It is followed by an overview of image segmentation in computer vision and in the medical domain. Then, a review of deep learning methods in medical image segmentation, specifically MS lesion segmentation and brain tumour segmentation, is provided. Finally, a review of transfer learning techniques in deep learning for natural images is presented along with its applications in medical image segmentation. Overall, this chapter focuses on providing a foundation for all the work presented in later chapters.

2.1 Introduction to Deep Learning

Deep Learning [62, 39] is a powerful class of machine learning methods which has received immense attention in various practical applications and as a research topic in the last decade. Deep Learning methods rely on simple computational units, linked through parameters that are tuned so as to model existing data. Deep learning methods are nowadays employed in many computer vision applications like image segmentation [68], object detection [101], image generation [40], visual question-answering [3], and image captioning [139]. One of the crucial reasons behind the success of deep learning methods in computer vision applications is that they perform feature engineering automatically from the raw data itself, instead of combining features that are hand-engineered. For example, consider the case of image classification in which the task is to differentiate whether an

object in a given image is a car or a bus. A traditional machine learning framework will first extract *features* from the image, such as whether there are wheels, the size and number of wheels, overall length and width of the object, to name a few, either using feature detectors or manually. These extracted features are then fed into a classifier that learns the correlation of these different features with the output class, i.e. car/bus, in this case. Extracting features is a tricky job and often requires domain knowledge. This decoupling of the two stages of feature engineering and classification limits the potential of traditional machine learning approaches. On the other hand, deep learning methods thrive by learning the feature extraction part and classification part end-to-end.

2.1.1 Convolutional Neural Networks

Convolutional Neural networks (CNNs) [62, 111] are a type of Artificial Neural Network (ANN) that operates on grid-like data, such as an image. This explicit assumption helps in reducing the number of parameters by constraining the architecture of the model. In CNNs, each neuron is connected to only a subset of neurons in its receptive field, and the weights among those subsets of neurons in that layer, also known as kernels, are shared. *Convolutional layers* operate on this set of learnable weights and receptive fields to calculate their dot product, which is used as an input to the next layer during forward propagation. Another key property of CNNs is that they preserve spatial information in the input image, unlike in ANNs, where the input image is flattened to form a vector. CNNs also employ *pooling layers* to reduce the spatial size of the representations between successive convolutional layers. This reduces the number of weights to be learned and hence prevents the network from overfitting. An example of a CNN is presented in Figure 2.1.

Given an input volume [58] of size $W_i \times H_i \times D_i$, one can design a convolutional layer of $\mathbf{K}$ filters, each with a receptive field of size $\mathbf{F}$, a stride $\mathbf{S}$ (with which the filters are slid over the input volume), and an amount of padding $\mathbf{P}$ needed across the border of the input to produce an output of spatial

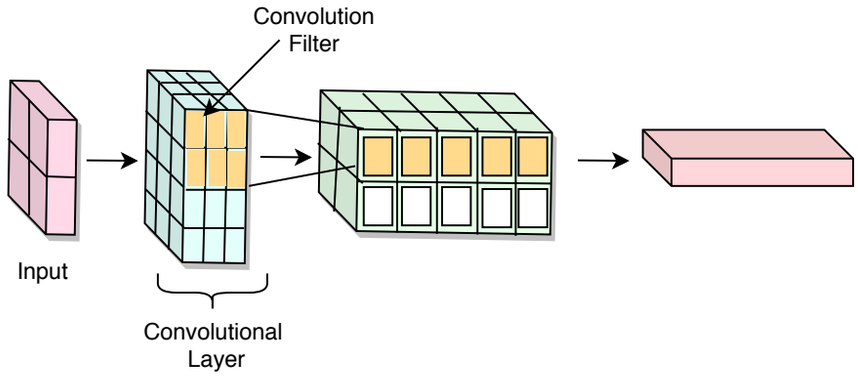


Figure 2.1: An example of a CNN model where K filters of spatial dimension F convolve over an image volume with stride S in order to produce an output volume of depth K.

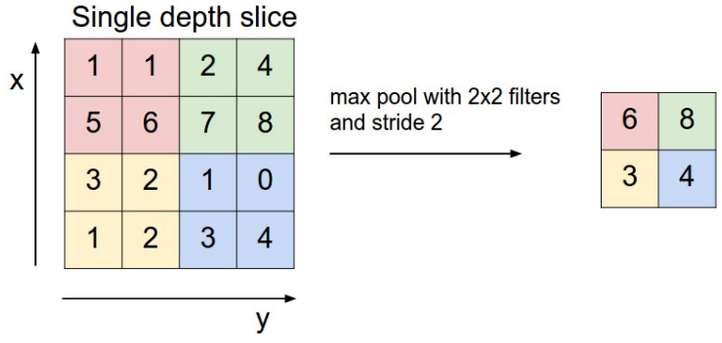


Figure 2.2: An example of a 2x2 max pooling operation with stride of 2. The max is taken out of every square belonging to different color. Image courtesy of [58]

dimension $W_o \times H_o \times D_o$:

$$\begin{aligned} W_o &= \frac{(W_i - F + 2P)}{S+1} \\ H_o &= \frac{(H_i - F + 2P)}{S+1} \\ D_o &= K. \end{aligned} \tag{2.1}$$

The notion of strides also extends to pooling layers with strides equal to the spatial dimension of pooling filters. An example of a 2x2 max-pooling operation is shown in Figure 2.2.

2.1.2 Training Neural Networks

In supervised learning, every input data (e.g. image) is provided with the corresponding ground truth label (e.g. bus or car, in the example we discussed above). After the initial pre-processing of the data, an appropriate neural network architecture is defined, and its weights are initialized. Note

that there are many smart techniques to initialize weights, which lead to faster convergence [38, 44]. The input is fed to the neural network, and the prediction is made at the last layer via forward propagation. This prediction is then compared to the ground truth label, resulting in an error that is captured using the notion of a loss function, **L**. In order to reduce the loss incurred, the weights of the network are updated based on a technique called backpropagation, which is a form of Stochastic Gradient Descent [103]. In practice, the weights are updated based on the loss incurred on a batch of data samples at a time, rather than a single example. There are many other optimizers as well [60]. The networks are trained until the loss function on a validation set saturates. It is quite possible for the network to accurately predict the labels on the training set without generalizing well to the unknown data in the validation/test set. This situation is called *Overfitting* and can be tackled by various regularization methods like reducing network complexity, adding dropout [119] or performing batch normalization [52].

2.2 Image Segmentation

Image segmentation is a problem in the field of computer vision which deals with understanding the content of an image by fragmenting it into sub-regions or segments. These segments are coherent, in that there is usually a high similarity between pixels within a region. Image segmentation mainly has two levels of granularity:

1. **Semantic Segmentation** refers to the labelling of every pixel with the class that best defines it.

2. **Instance Segmentation** deals with identifying all the instances of a class that are present in the input image.

Figure 2.3 illustrates the distinction between instance and semantic segmentation. In the case of semantic segmentation, all the parrots are coloured in the same colour. In contrast, in instance segmentation, every parrot in the image is considered a new instance of the class *parrot* and coloured differently. Image segmentation algorithms have been widely explored in various applications, which

Figure 2.3: An example illustrating the difference between Semantic Segmentation and Instance Segmentation. The input image on the left is courtesy of [93].

include medical imaging and diagnostics [97], autonomous car driving [128], robotics [83], and video surveillance [70].

A variety of approaches have been proposed to deal with the challenges related to the problem of image segmentation. For example, in [95], the authors proposed a new segmentation algorithm based on morphological properties of connected components in high-resolution satellite imagery. The work in [53] presented an unsupervised clustering based texture segmentation algorithm that uses the concept of Gabor filters. In [59], the authors employed a non-parametric information theory based approach on a variety of image segmentation problems. The major disadvantage of all these approaches is that they rely heavily on using image properties like texture, intensity, contrast and entropy to understand the structures in the image, which can be subject to noise and variability that are hard to work with in practice. Researchers also explored discriminative techniques like Support Vector Machines [135, 85] and Random Forest [110] in the context of natural image segmentation. However, all the approaches mentioned above failed to model the local spatial relationships, and some approaches failed to learn from previous training examples. To deal with these issues, researchers also worked on probabilistic graphical models like Conditional Random Fields [11, 48, 98], and Markov Random Fields [33, 26] for modelling contextual and spatial relationships between neighbouring pixels explicitly. In recent years, CNNs and other related models have proven much more successful for image segmentation.

We will now focus on the problem of segmentation in medical images, which is the topic of our work.

2.3 Medical Image Segmentation

Image segmentation, in the context of medical images, is the process of delineating anatomical structures and other regions of interest. These regions could either belong to normal healthy tissue, like hippocampus [65], ventricles [124], liver [2] or to a pathology, such as lesions [122, 34] or tumours [99, 80]. The segmentation of anatomical structures is useful for disease diagnosis, treatment planning, computer-aided surgery and clinical studies. There are various types of imaging modalities such as CT, MRI, X-ray, Microscopy, and Positron emission tomography (PET), which provide insights about the internal anatomy of a patient. Figure 2.4 presents several examples of medical images.

Medical image segmentation comes with its unique challenges, which make it a more complex problem than natural image segmentation. First, medical data can be 2D, 3D or even 4D (spatio-temporal) in nature. Second, there is high variability in medical images (See Figure 2.4). Different anatomical structures can be present non-uniformly with respect to spatial occurrences of pixels or groups of pixels. The shape and size of pathologies can vary across patients. Usually, pathology is a small fraction of the whole image, which leads to large class imbalances. Also, obtaining datasets of patients with rare diseases is a difficult task. There are many other challenges like noise induced by sensors, occlusion, missing boundaries of regions, low-resolution images and motion artifacts, which add to the complexity of the problem [97].

In order to address these specific challenges, a lot of medical image segmentation algorithms have been proposed. The suitability of an algorithm depends on its area of application, the imaging modality used, and the body organ under scrutiny [113]. One of the earliest approaches to medical image segmentation includes several thresholding algorithms [63, 109], which aim to fragment the input image based on different thresholds of image intensity. This kind of approach fails to model the spatial properties of the image and is highly prone to noise and other artifacts. Methods involving the extraction of regions based on manually specified seed points, called region growing methods, have also been developed [73, 140]. Although these methods take into account important image features like edges, in addition to the image intensities, they are still limited in scope due to the

need for manual intervention and sensitivity to noise.

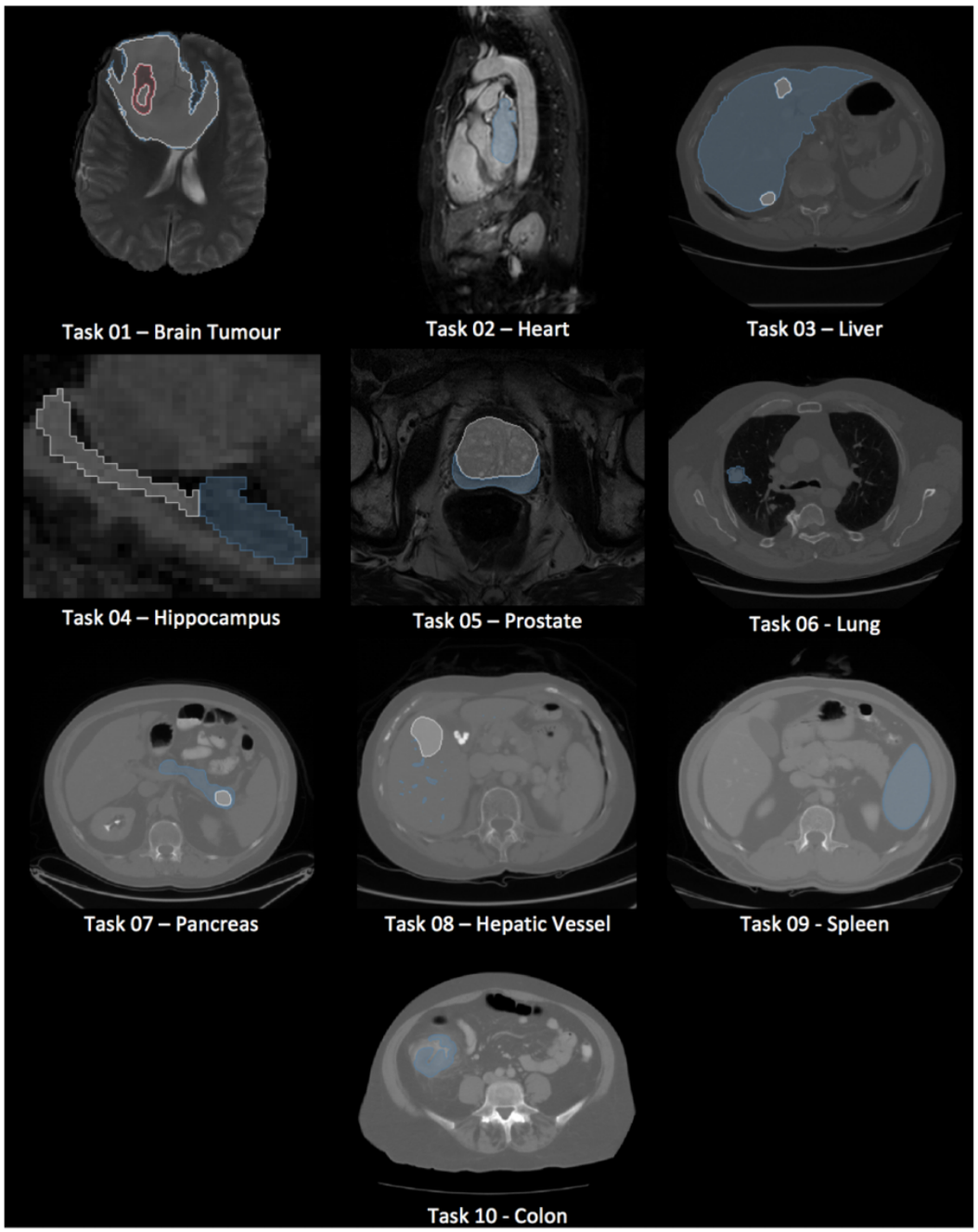


Figure 2.4: Examples of medical images capturing different anatomical structures of interest in different organs. Image courtesy of [114].

Deformable methods [10, 25] use closed curves for defining the continuous boundary of a region, which is obtained by iteratively applying internal and external forces. An advantage of deformable methods is that they take into account the shape and appearance of the region as distinctive features

and are more robust to noise. However, they still require manual intervention [97]. Classifier-based methods like k-nearest neighbour or Bayes classifiers can be used to train a segmentation model from data [76], but they strongly depend on the training set used. Clustering methods [23, 96] are similar to classifier-based methods, but they do not require labelled training data. Both classifiers and clustering methods lack in modelling spatial relationships in the image. Atlas-based approaches [13, 55] work by viewing the segmentation task as a registration problem. A target image is first registered to a pre-annotated atlas image through a process called atlas warping. Afterwards, the target image is segmented by transferring labels from the already segmented atlas image. Atlas-based approaches are known to be prone to anatomical variability [97]. In order to incorporate the spatial context between neighbouring voxels, researchers use statistical models called Markov Random Fields (MRFs) [45, 146, 67]. MRFs assume that pixels which are adjacent to each other should belong to the same class, but other priors can also be used. However, MRFs can be sensitive to the choice of prior. A prior that is too strong can lead to extremely smooth segmentation, with loss of intricate structural information [97].

2.3.1 Deep Learning for Medical Image Segmentation

All the approaches discussed so far have a common disadvantage: they separate feature engineering from the classification phase, as discussed in Section 2.1. With advances in computational power, Deep Learning [62] has gained immense popularity in the field of medical image segmentation, by overcoming the need to extract hand-crafted features from the images. By leveraging advances in parallel computing, it has been possible to train deeper models that have become state-of-the-art in medical image segmentation. The capabilities of deep learning-based approaches have led to their use in a variety of medical image analysis tasks like classification, segmentation, detection, image registration, and disease prognosis. This section is, however, dedicated to the application of deep learning methods specifically for the task of medical image segmentation.

For a long time, researchers have worked on *patch-based methods*, where the goal is to classify the pixels in small 2D or 3D patches extracted from full 3D volumes. The application of patch-based

methods for segmentation in MRI images includes brain tumour segmentation [42, 50], segmentation of different anatomical structures in the brain [133], lesion segmentation in MS patient MRIs [132], segmentation of rectal cancer in the pelvis [127], segmentation of ventricles in cardiac MRI [31], pancreas segmentation in CT images [105]. One limitation of patch-based methods is that they fail to capture a wider or global context, focusing instead on the window size of the patch. Moreover, these methods are computationally demanding, especially when overlapping patches are present.

In order to deal with this problem, researchers started working on incorporating full image volumes as input to deep learning models called fully convolutional neural networks. These networks predict a segmentation map as an output of the same size as the input volume. Some of the most prominent network architectures in this regard are fully Convolutional Networks (FCNs) [68], UNet [104, 22], VNet [81] and SegNet [5].

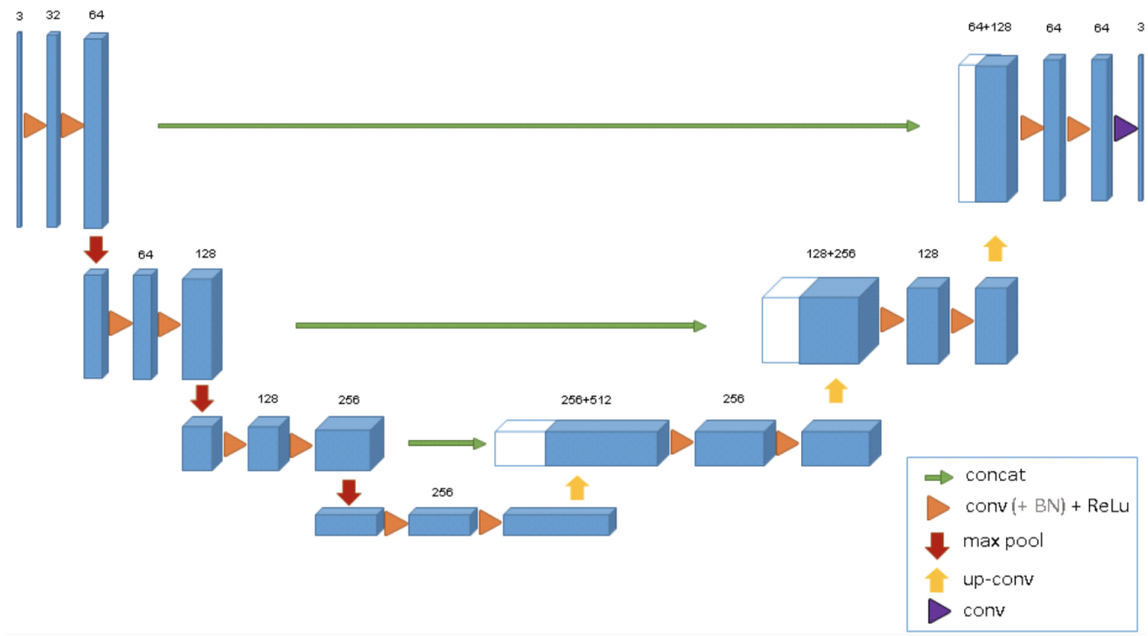


Figure 2.5: Architecture of 3D UNet. Image courtesy of [22].

Figure 2.5 depicts the architecture of the UNet, which will be a starting point in our investigation. The network consists of an encoder, followed by a decoder which outputs a segmentation map. The encoder and decoder are interconnected at different scales via links called skip connections. These skip connections help in passing gradients back through the network more efficiently, thereby

promoting faster training. Although working on full images is computationally more challenging, it has been widely explored in medical image segmentation tasks due to the increasing availability of high-performance computing resources. Applications include the segmentation of MS lesions in the brain [15, 75, 88], brain tumour segmentation [57, 54], prostate gland segmentation [144] and cardiac segmentation [126]; see [74, 66] for a survey.

2.3.2 Deep Learning for MS Lesion Segmentation

One of the early applications of deep learning in MS lesion segmentation is [130], in which the authors present a 3D CNN model trained on 3D MRI patches from the public Longitudinal Multiple Sclerosis Challenge 2015 [17]. The authors of [15] use 3D convolutional encoder networks with shortcut connections to capture features at multiple levels, for better lesion segmentation across different lesion sizes. They also observe that increasing the depth of the network has a positive impact on the segmentation output. Also, authors in [43] develop an automated lesion segmentation network on the Grand Challenge (MSGC) dataset [121] that can deal with missing modalities. Their method learns an embedding for every input modality, where the latent space is acted upon by well-defined arithmetic operations in order to generate the segmentation results. The work in [35] incorporates location-sensitive information or lesion priors into a deep learning model trained on patches of multiple sizes. In [132], the authors propose a cascade of two 3D patch-based CNNs, where the first network predicts prospective lesion voxels, and a secondary network refines the segmentation results by reducing the number of false positives.

2.3.3 Deep Learning for Brain Tumour Segmentation

The segmentation of brain tumours is a crucial task for cancer diagnosis, assessment of tumour growth, treatment recommendation, evaluation of treatment response, and computer-aided surgery.

Recently, deep learning methods have achieved state-of-the-art results on some brain tumour segmentation tasks [87] without the need for hand-crafted features. Some of the early works applying deep learning approaches to brain tumour segmentation include [42, 148, 94]. These are patch-based

methods which extract patches from the image and predict the class label of the central pixel. Initially, researchers mainly opted for 2D patches [42] instead of 3D patches [129], for the sake of managing computational load. In [57, 42], the authors incorporate global and local context using dual pathway CNN models in order to produce accurate segmentation. In [42], the authors introduce cascaded CNN models for brain tumour segmentation, where soft segmentation maps of a base model are concatenated with input image sequences to train a secondary model. The authors of [94] opt for smaller kernels, which allows them to have deeper networks with the same receptive field as shallow networks with bigger kernels. In [71], the authors convert the problem of multi-class brain tumour segmentation into multiple binary sub-tasks for segmenting different sub-structures of the tumour. They also use an ensemble of networks trained on one of the three orthogonal planes in order to incorporate 3D contextual information. Recently, the UNet architecture, which we discussed above [104, 22] has become popular for brain tumour segmentation [30] due to its empirical success.

2.4 Transfer Learning

Humans are capable of transferring knowledge across different tasks. For example, consider a person who knows how to ride a bike. If he or she tries to learn how to ride a motorcycle, the previous biking experience helps to quickly pick up the new skill rather than learning basic principles of riding (like balancing or following traffic rules) from scratch. Unfortunately, artificial agents based on deep learning are designed to learn different tasks separately (see Figure 2.6). They fail to generalize when presented with out-of-distribution data, i.e. data of a type which has not been seen during training. However, in practical applications, it can be difficult or expensive to obtain large annotated datasets for training data-hungry artificial agents from scratch. Transfer learning helps address this challenge.

Before proceeding to the formal definition of transfer learning, we need to introduce some notations. A domain [92] $\mathcal{D} = \{\mathcal{X}, P(X)\}$ has two components: $\mathcal{X}$ is the feature space and $P(X)$ is the marginal probability distribution of $X = \{x_1, x_2, \cdots, x_n\} \in \mathcal{X}$. A difference in either feature space or marginal probability distribution leads to a different domain. Given a specific domain $\mathcal{D}$, a task $\mathcal{T} = \{\mathcal{Y}, f(\cdot)\}$ consists of two parts: $\mathcal{Y}$, the label space, and $f(\cdot)$, a predictive conditional

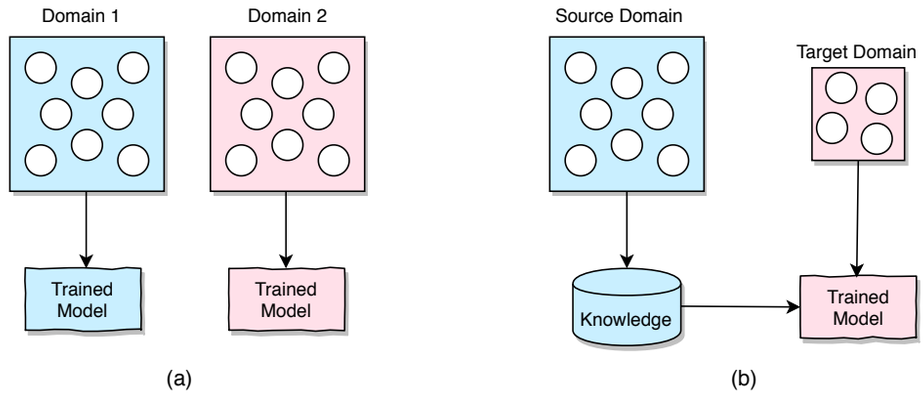


Figure 2.6: Example illustrating the difference between (a) traditional machine learning and (b) transfer learning.

function which learns $P(y \mid x)$ from training data samples $\{x_i, y_i\}$, where $x_i \in X$ and $y_i \in \mathcal{Y}$ [92].

Given a source domain $\mathcal{D}_S$ and source learning task $\mathcal{T}_S$, a target domain $\mathcal{D}_T \neq \mathcal{D}_S$ and target learning task $\mathcal{T}_T \neq \mathcal{T}_S$, transfer learning tries to leverage the knowledge obtained through $\mathcal{D}_S$ and $\mathcal{T}_S$, in order to better learn the target predictive function $f_T(\cdot)$ in $\mathcal{D}_T$ [92]. Transfer learning achieves this by focusing on three main aspects: "*what to transfer*", i.e., which information in the source is of use to the target, "*how to transfer*", i.e., finding the best way to transfer the useful information, and "*when to transfer*", i.e., understanding which information from the past can interfere with the target and how to avoid transferring in such scenarios. The above three aspects vary depending on how similar the source and the target domains are.

The relationship between source and target domains and tasks leads to two primary transfer learning settings: *homogeneous transfer learning* and *heterogeneous transfer learning*. Homogeneous transfer learning deals with cases where $P(\mathcal{X}_T) \neq P(\mathcal{X}_S)$ and/or $P(\mathcal{Y}_T \mid \mathcal{X}_T) \neq P(\mathcal{Y}_S \mid \mathcal{X}_S)$. It tries to narrow the gap of the marginal distribution $P(\mathcal{X})$ and/or conditional distribution $P(\mathcal{Y} \mid \mathcal{X})$ between source and target domains. Heterogeneous transfer learning deals with cases where $\mathcal{X}_T \neq \mathcal{X}_S$ and/or $\mathcal{Y}_T \neq \mathcal{Y}_S$. It focuses on narrowing the gap between the input feature spaces of the source and target domains and further correcting the differences between domain distributions (marginal or conditional) if they exist, using homogeneous transfer learning techniques [136]. The methodologies for homogeneous transfer learning include instance-based, feature-based, parameter-based, hybrid-based and relational based approaches. Instance-based approaches [47] employ a weighing scheme by

which some instances in the source domain are re-weighted and directly used with the target set for training. This re-weighing helps in reducing the difference in marginal distributions between the two domains. Feature-based approaches work in two ways: the first approach [46] finds a transformation of the feature space of the source domain to be closer to the target domain. The other approach [91] transforms both the domains to a common meaningful latent feature representation with better predictive capabilities and therefore reduces the differences in the marginal distribution between the two domains. Parameter-based methods [142, 107] share the parameters of the models learned on the source and target domain. They also create an ensemble of multiple source models combined optimally to improve the performance of the target model. Hybrid-based approaches [138] combine instance and parameter-based approaches for transferring knowledge between the domains, whereas relational based approaches [141] work in scenarios where there exists a well-defined relationship between source and target domains. Heterogeneous transfer learning deals with cases where the feature spaces differ. In that case, feature-based methods help by converting the heterogeneous problem to a homogeneous problem and then using the methodologies mentioned above for homogeneous transfer learning.

In recent years, transfer learning has also been successfully applied in the context of deep learning. Using off-the-shelf pre-trained models [112] as feature extractors has been the most basic way to transfer the knowledge from source to target. The key idea here is to use initial layers of a pre-trained source model as a fixed feature extractor for other target tasks. An advantage of this approach is that it allows the training of shallow classifiers on top of the extracted features.

Another popular strategy for deep transfer learning is to selectively retrain or fine-tune the layers of a pre-trained model [90, 49]. As deep learning models are layered architectures, different layers capture different levels of patterns in the input. The initial layers tend to capture generic features like shape, colour, orientation, whereas deeper layers tend to capture more complex patterns in the input, which are often task-specific in nature [143]. Typically, the initial layers of a source model are frozen while the deeper layers are retrained or fine-tuned on the target task. The selection of layers to freeze or retrain depends on the target task and the amount of target data available. Another slightly different type of deep transfer learning is called multi-task deep learning [108], where a deep

learning model is trained to handle several tasks simultaneously. This leads to an inductive bias in the model, which encourages the model to learn features that can generalize to more than one task.

2.5 Deep Transfer Learning for Medical Image Segmentation

The potential impact of deep learning methods in medical image analysis is immense. Deep learning systems have been successful in early diagnosis, predicting the risk of diseases and thereby aiding in taking preemptive steps to prevent them. Yet, many barriers slow down their progress in the medical field. Since deep learning requires large amounts of data for training, the unavailability of large annotated datasets becomes a hurdle for the advancement of deep learning systems in medical imaging analysis. Acquiring medical imaging data is a challenging task, as the process is slow and expensive. Moreover, labelling the data depends on the availability of qualified clinical experts, which is a laborious and expensive job. Often, the issue of imbalanced data remains in healthcare due to factors like shortage of data for rare diseases or high class imbalances in terms of pathology vs healthy tissues. Moreover, the majority of datasets available publicly are small. Large datasets exist but are mostly proprietary, which further hinders their use freely in the medical community. Recently, transfer learning has proven effective in dealing with small data regimes in healthcare domains for various applications such as the classification of skin lesions [78] and brain lesion segmentation [36]. Typically, a deep network is trained using a large-scale source dataset. This pre-trained network is then used in two ways: either to extract features for the target task [51] or as initialization for further fine-tuning based on the target task [125], as shown in Figure 2.7. For example, in the case of classification tasks, it has been shown that results on smaller target medical datasets [78, 145, 102] could be improved by leveraging models pre-trained on medical and/or non-medical datasets like ImageNet [27].

Transfer learning has also been explored in many medical image segmentation applications. MRI imaging data acquired from different scanners and imaging protocols leads to differences in the

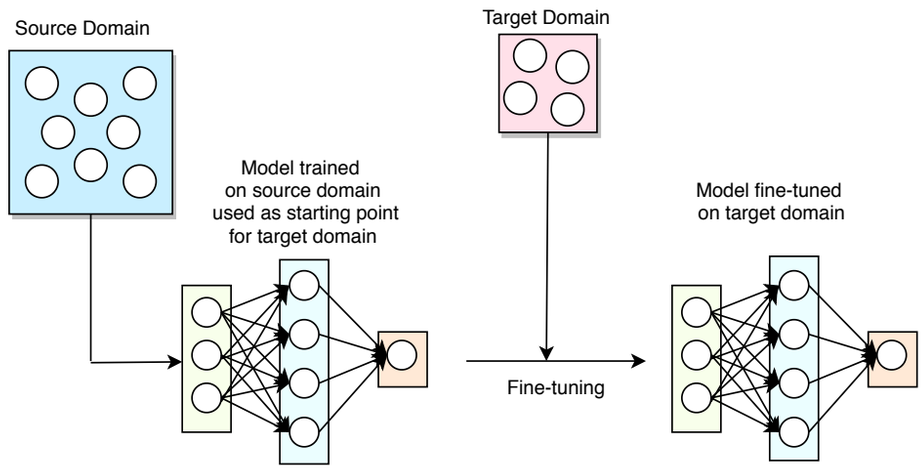


Figure 2.7: An example illustrating how fine-tuning is performed in deep learning networks in order to transfer knowledge from source to target.

nature and quality of the images. Often, models trained with data belonging to one protocol can perform poorly on data from other protocols. In [36], the authors show that fine-tuning can be useful to increase performance in white matter lesion segmentation in the brain, when the data is acquired from multiple scanners and imaging protocols. They also present an extensive study on how much target data is required for successful domain adaptation of the source network and how much of the pre-trained model should be fine-tuned with target data. In [56], the authors address the same problem in the context of brain lesion segmentation, by performing unsupervised domain adaptation, which means that their method does not require ground truth labels for the target set. This involves training a discriminator which classifies the domain of the input image based on activations obtained when the input is passed through a segmentation network, which is trained to learn domain-invariant features. In [1], the authors show how knowledge can be transferred from one grade of a disease to another. Notably, the authors show that transfer learning helps to generalize on unseen test data by using a model pre-trained on high-grade brain tumour cases and fine-tuning it with low-grade glioma cases. In [6], the authors use the prediction of anatomical positions as a pretext task, and train input features in a self-supervised manner for cardiac MR image segmentation. In [86], the authors train a single CNN to perform different tasks, ranging from tissue segmentation in brain MRI, pectoral muscle segmentation in breast MRI and cardiac CTA segmentation. This single network performs comparably to a CNN model trained for each task individually. In [125], the

authors perform extensive experiments on four different medical imaging applications in order to show that fine-tuning CNNs deeply can be effective in many medical image analysis tasks. Recently, the authors of [147] trained a set of models called Model Genesis in a self-supervised way, without requiring any manual annotations, and showed how the knowledge in this model could be leveraged to improve segmentation results across multiple diseases (lung nodule, brain tumour, pulmonary embolism, liver diseases) and multiple modalities (e.g. CT, X-ray, MRI).

3

Transfer Learning via Fine-tuning

In the previous chapter, we saw how the performance of deep learning models suffers in small data regimes in medical imaging applications. We also saw how transfer learning could be beneficial for various medical imaging applications to alleviate such problems. This chapter explores the hypothesis that transfer learning for the task of segmenting pathological structures can be performed across neurodegenerative 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 neurological diseases, (b) the pathological structures are different in the two datasets (lesions vs. tumours), 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 to explore the ability of transfer learning to improve the results of an end-to-end multi-class brain tumour segmentation network trained on subsets of the MICCAI 2018 BraTS dataset [79]. Given that both MRI datasets are acquired from patients with neurological diseases that present with focal pathologies (lesions and

tumours), the intuition is that the two datasets 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 tumour dataset.

3.1 Methodology

This section talks about different fine-tuning strategies that we employ for transferring knowledge across diseases. Mainly, a 3D deep neural network inspired by UNet [22] has been used for the task of focal pathology segmentation. The architecture of the network is depicted in Figure 3.1(a), and the implementation details of the model are described in Section 3.2.2.

Now, 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 to improve pathology segmentation performance. A popular strategy for transfer learning consists of fine-tuning the pre-trained source network on the target dataset. It is also to be noted that fine-tuning is a general approach and can be performed as long as a pre-trained network trained on a source dataset is available or the source dataset is available to train a source network. In this chapter, three different strategies of fine-tuning are explored, considering that we have a source dataset that is used to train a source network. The most common way of fine-tuning consists of replacing the last few layers of the source network with new layers, re-initializing the weights and changing the output dimension of these newly added layers as per the target task. The remainder of the network is frozen, which prevents the gradient flow. In the first strategy, namely *FT-Last Three*, only the newly added layers are trained on the target dataset (See Figure 3.1(b)). This strategy has been advocated when the amount of target data available is small, and the similarity between the two datasets is high, as in the context explored in this paper [36]. 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. In contrast, 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 [143, 21]. This leads to the second strategy we explore, represented as *FT-Decoder*, which involves freezing the encoder and fine-tuning the entire decoder (See Figure 3.1(c)). The third strategy, represented as *FT-All*, consists of fine-tuning the whole network with target data (See Figure 3.1(d)).

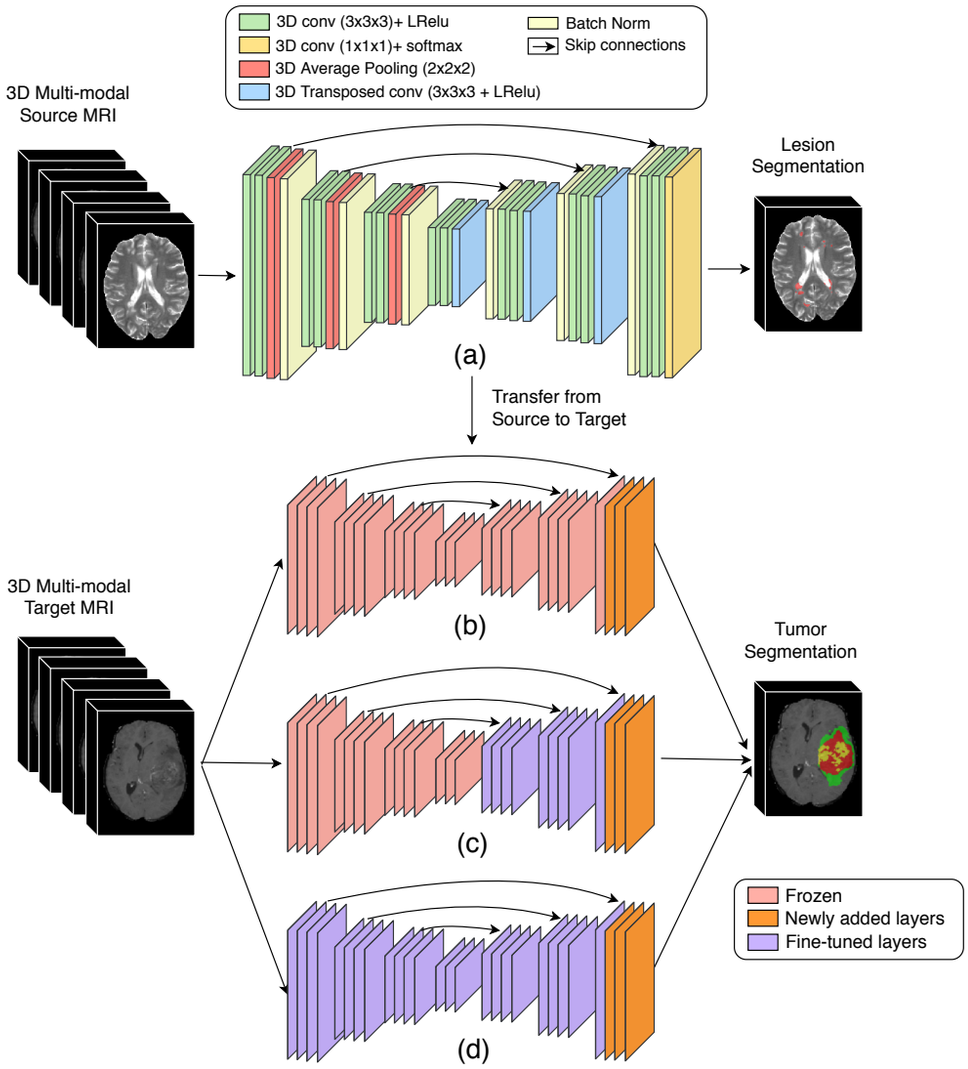


Figure 3.1: Transfer learning framework. (a) UNet architecture for pre-trained source network. (b), (c) and (d) depict different methods of adapting the pre-trained source network for the target task. In all three, the last three task-specific layers are replaced with new layers (orange). The remaining 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.

3.2 Experimentation Pipeline

This section describes the data and its preprocessing. This is followed by a detailed breakdown of the 3D UNet based network architecture along with its training procedure and the experimental setup. To assess the performance of the three different transfer learning approaches, as discussed in Section 3.1, 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 tumours and their tissue sub-classes from patient MRI. We compare the performance of these transfer learning approaches to training only on the target data from scratch, for different dataset sizes.

3.2.1 Data Description and Pre-processing

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 24 months, 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 $1mm \times 1mm \times 3mm$ resolution. They are then interpolated to $1mm^3$ 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. Before using these MRI samples for training the segmentation network, a preprocessing pipeline is followed, which includes brain extraction [116], N3 bias field inhomogeneity correction [115], Nyul image intensity normalization [89], and registration to the MNI-space.

Brain Tumour Dataset (Target): The target datasets are obtained by subsampling datasets of various sizes from the BraTS 2018 MICCAI challenge [79, 7, 8]. 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. Expert 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). Tumours are segmented into three classes: edema, necrotic/non-enhancing core, and enhancing tumour. These three classes combined together are referred to as "whole" tumour. The volumes are co-registered, resampled to $1mm^3$ resolution and skull-stripped. Our pre-processing pipeline includes registration of samples to the same space as MS data using ANTs tool [4].

For both MS dataset and Brain tumour 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.

3.2.2 Architecture of 3D CNN for Segmentation

The proposed segmentation network is a 3D UNet based network that takes 3D patient MRI sequences as input and generates a 3D output mask of the same resolution. As is typical of a 3D UNet [22, 77], the network consists of an encoder part and a decoder part, which helps it to model the entire brain concurrently for the segmentation task preserving all spatial information unlike in patch-based methods. The architecture of the proposed network is shown in Figure 3.1(a).

Like the original 3D UNet, our proposed segmentation network also consists of four downsampling blocks followed by four up-convolution blocks. 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), unlike ReLU in standard 3D UNet. Average pooling of size $2 \times 2 \times 2$ and stride of 2 is performed, followed by batch normalization [52]. 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

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

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/Sigmoid non-linearity. The implementation of the model is done in Pytorch[2].

3.2.3 Training the Network

Segmenting MS lesions is a binary voxel-wise classification task, whereas brain tumour sub-type segmentation is a four-class voxel-wise classification task [79]. For lesion segmentation, the training objective is weighted binary cross-entropy loss. Given an expert labelled image Y^n for a patient volume n consisting of labelled voxels $y_1^n, \ldots, y_i^n$ and the network's predicted output $\hat{Y}^n$ containing voxel predictions $\hat{y}_1^n, \ldots, \hat{y}_i^n$, then, the weighted binary cross-entropy loss is computed as follows:

$$BCE^n = -\sum_i \Big(w_e y_i^n \log(\hat{y}_i^n) + (1 - y_i^n) \log(1 - \hat{y}_i^n) \Big) \tag{3.1}$$

where w_e is the weight of lesion class at epoch e during training as specified in Equation 3.2. The weighted binary cross-entropy loss helps to account for class imbalance, caused due to the presence of fewer lesion voxels in comparison to non-lesion voxels.

$$w_e = max(Pr^e, 1) \tag{3.2}$$

where P, initial weight for the lesion voxel class, is the ratio of the number of non-lesion voxels to the number of lesion voxels in the whole training dataset as defined below:

$$P = \frac{\#\text{non-lesion voxels in whole dataset}}{\#\text{lesion voxels in whole dataset}} \tag{3.3}$$

Initially, due to the presence of a small proportion of lesion voxels, high value of P will encourage the network to put more emphasis on learning voxels belonging to lesion class, thereby over-segmenting the lesion. As the training procedure proceeds, the emphasis of P will decay exponentially with epochs at the rate of r, decay rate until it reaches 1, as specified in Equation 3.5. This weighing scheme

[2] http://pytorch.org/

promotes the network to decrease the number of false positives in prediction, thereby discouraging over-segmentation of lesions as the learning proceeds.

For the multi-class brain tumour segmentation task, the training objective is weighted categorical cross-entropy loss. Given $\hat{y}^n_{i,c}$ be the probability with which the network predicts that voxel i of volume n belongs to class c, then, the weighted categorical cross-entropy loss is calculated as shown in Equation 3.4.

$$CCE^n = -\sum_i \sum_c w^n_{i,c} \log(\hat{y}^n_{i,c}) \tag{3.4}$$

$$w^n_{i,c} = p_{c,e} \times y^n_{i,c}, \quad p_{c,e} = p_c * r^e + 1, \quad \text{where, } p_c = \left(\frac{\sum_{k=0}^{k=C} V_k}{V_c} \right) \tag{3.5}$$

The initial weight of a class c, p_c, is calculated as the ratio of the total number of voxels divided by the number of voxels V_c belonging to class c in the training set. The class weights $p_{c,e}$ at epoch e are scheduled to decay with a decay rate lower than 1 as explained in Equation 3.5. As the number of epoch increases, the weight for each class converges to 1, ensuring that every class is given equal importance during the later stages of training.

3.2.4 Evaluation Metrics

For the task of segmenting MS lesions, a voxel-level analysis of the network's prediction is performed. The network's sigmoid output is compared against a threshold, t, for obtaining a binary segmentation output. It is followed by counting the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) for a given input image. For example, every voxel for which the predicted lesion voxel is correctly classified accounts for True Positive (TP). In contrast, a False Positive (FP) results from the misclassification of a voxel as a lesion voxel. These four metrics are summarised in Table 3.1.

These voxel-wise measures are then used to calculate the True Positive Rate (TPR) and False Detection Rate (FDR), as defined in Equations 3.6 - 3.7. TPR and FDR are calculated at a set of different thresholds t, which capture the range of operating points describing the network's

		Ground Truth	
		1	0
Prediction	1	TP	FP
	0	FN	TN

Table 3.1: Confusion matrix showing how True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) are calculated.

predictions. These resulting vectors of TPR and FDR at different thresholds are then plotted in a Receiver Operating Characteristic (ROC) curve. As the threshold t decreases, the network prediction will classify more voxels as lesion voxels, thereby increasing the number of TP and FP, and therefore, further increasing both TPR and FDR. Ideally, segmentation output with a TPR of 1.0 and FDR of 0.0 is considered optimal.

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP + FN} \tag{3.6}$$

$$\text{False Detection Rate (FDR)} = \frac{FP}{TP + FP} \tag{3.7}$$

Now, for the task of segmenting brain tumours, the segmentation performance is assessed using Dice scores. Every voxel is classified as belonging to the class, which maximizes the network's softmax prediction at that voxel. The Dice score, as explained in Equation 3.8, is calculated for different substructures of the tumour, namely, whole tumour, core tumour and enhancing tumour. The value of the Dice score can range from [0,1]. Higher the value of the Dice score, the better the segmentation performance. A Dice score of 1.0 for segmentation is ideal.

$$\text{Dice} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3.8}$$

3.2.5 Experiments

As described in Section 3.1, the baseline experiment consists of training a network from scratch on the brain tumour 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

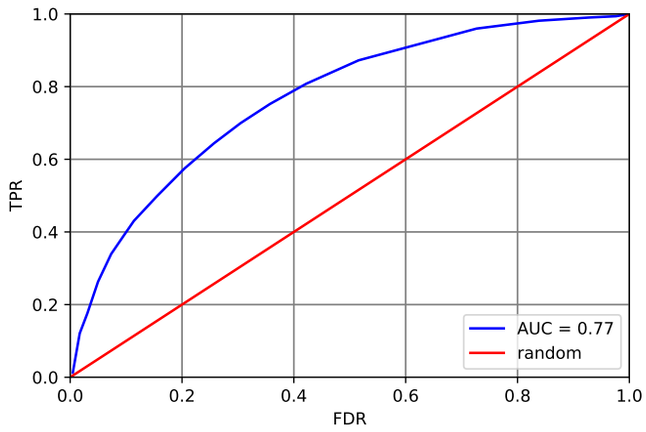


Figure 3.2: Voxel-wise performance ROC curves on the MS validation set of 718 samples. The x-axis depicts False Detection Rate (FDR) and y-axis depicts the True Positive Rate (TPR) captured at a set of different thresholds used to binarize the network's sigmoid predictions.

in Section 3.1 and denoted as *FT-Last Three*, *FT-Decoder*, *FT-All* in Figure 3.1. 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 initial weight of the positive lesion class is found to be 660.68. For all the experiments, Adam is used as an optimizer with $(\beta_1, \beta_2) = (0.9, 0.999)$. Additional information about hyper-parameters used in MS pre-training experiment can be found in Table 3.2. The best validation performance of the pre-trained network is obtained at epoch 186 with an AUC of 0.77, as shown in Figure 3.2.

In order to examine the effect of the size of the target dataset on the transfer learning outcome, the number of patient brain tumour 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 hyperparameters. More information about hyper-parameter tuning is shown in Table 3.2. Then, the networks are re-trained on the respective complete training sets, using the hyper-parameters that performed best during cross-validation for 100 epochs, and a local validation set (a subset of BraTS 2018 training set) of 50 samples is used to select the operating point. The best learning rate obtained after hyper-parameter

tuning is shown in Table 3.3. Finally, the predictions on the separate BraTS 2018 Validation set are registered to the same space as the original BraTS 2018 training set using ANTs tool [4]. The performance on the BraTS 2018 Validation set (for which the ground truth is not available) is then evaluated by uploading the predictions on the BraTS 2018 challenge portal.

Hyper-parameter	MS pre-training	Baseline, FT-Last Three, FT-Decoder	FT-All
Initial Learning rate	$1e-4$	$\{1e-3, 1e-4, 1e-5, 1e-6\}$	$(UNet1, UNet2) \in \{(1e-4, 1e-3), (5e-4, 1e-3), (1e-3, 1e-3), (1e-4, 1e-4), (1e-3, 5e-3)\}$
Batch size	4	2	2
Class weight decay rate	0.95	0.92	0.92
Learning rate decay schedule	0.75 every 50 epochs	0.75 every 25 epochs	0.75 every 25 epochs

Table 3.2: Hyper-parameter tuning for different methods. The intuition behind keeping a lower learning rate for the encoder in case of the FT-All experiment is to alleviate the network from totally forgetting low-level knowledge representations learned about source data, which can also help learn the target task.

Methods	Samples	Best Learning rate
MS pre-training	2912	$1e-4$
Baseline, FT-Last Three, FT-Decoder	20	$1e-3$
	50	$1e-3$
	100	$1e-3$
	150	$1e-3$
FT-All	20	$(UNet1, UNet2) = (1e-3, 1e-3)$
	50	$(UNet1, UNet2) = (1e-4, 1e-3)$
	100	$(UNet1, UNet2) = (1e-4, 1e-3)$
	150	$(UNet1, UNet2) = (1e-4, 1e-4)$

Table 3.3: The best learning rate (LR) obtained after hyper-parameter tuning for different methods as a function of the number of brain tumour cases available for fine-tuning. The learning rate is chosen as the one which provides the best results on the majority of the three Dice scores (Dice enhance, Dice core, and Dice whole).

3.3 Results

Figure 3.3 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 tumour 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.

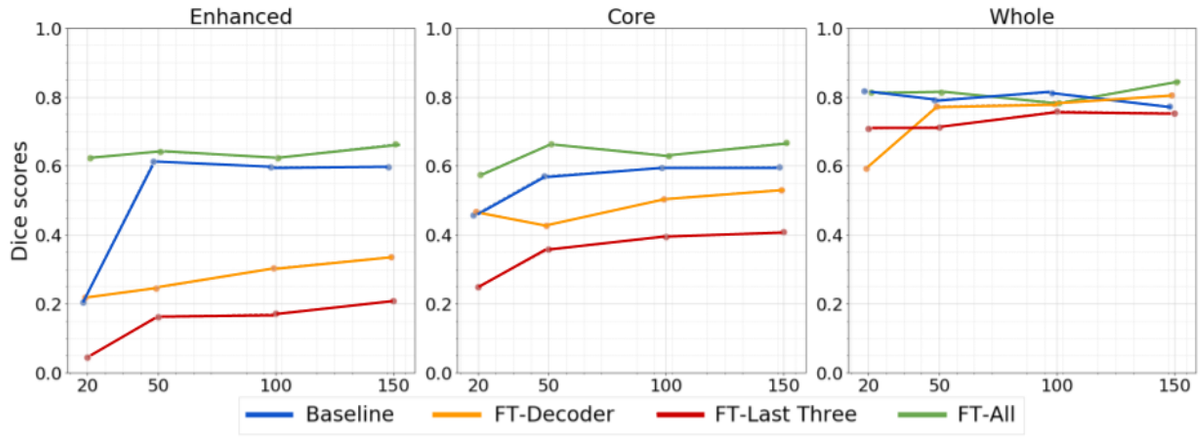


Figure 3.3: Comparison of Dice values for baseline method against different fine-tuning methods for enhanced, core and whole tumour segmentation on the BraTS 2018 Validation set. The x-axis depicts a varying number of brain tumour cases available for training (20, 50, 100, 150).

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 tumour, particularly when the number of tumour cases is extremely low, with 25.9% and 204.09% improvement on core and enhanced tumour over baseline respectively when the number of cases is 20. The percentage improvement is calculated as the ratio of the difference in the baseline and FT-All Dice scores over the baseline. Since lesions are smaller in size when compared to tumours, the results indicate that the network is extracting information from the MS pre-trained network that is relevant to segmenting sub-regions of tumour well, even though lesions present quite differently than brain tumours. As the number of brain tumour samples increases, the gain of FT-All over baseline diminishes. FT-Last Three and FT-Decoder do not 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

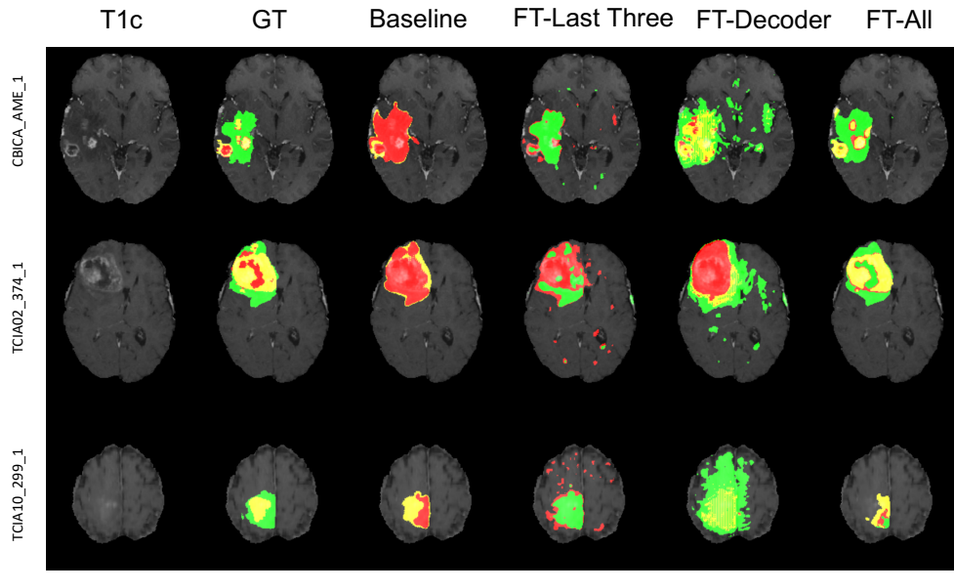
UNet to produce an output.

Qualitative segmentation results of the different methods on the local validation set for the case of 20 and 150 target dataset samples are shown in Figure 3.4(a) and Figure 3.4(b) respectively. Note that with just 20 target dataset samples, FT-All is able to capture different sub-structures of tumour better than the other methods. Performance is better on the HGG over the LGG case, as more HGG cases are present in the training dataset. Also, fine-tuning with 150 brain tumour samples generally produces better results than with just 20 samples in all the experiments. It can also be seen that in the case of 150 brain tumour samples, FT-All produces much more refined results in comparison to all other methods.

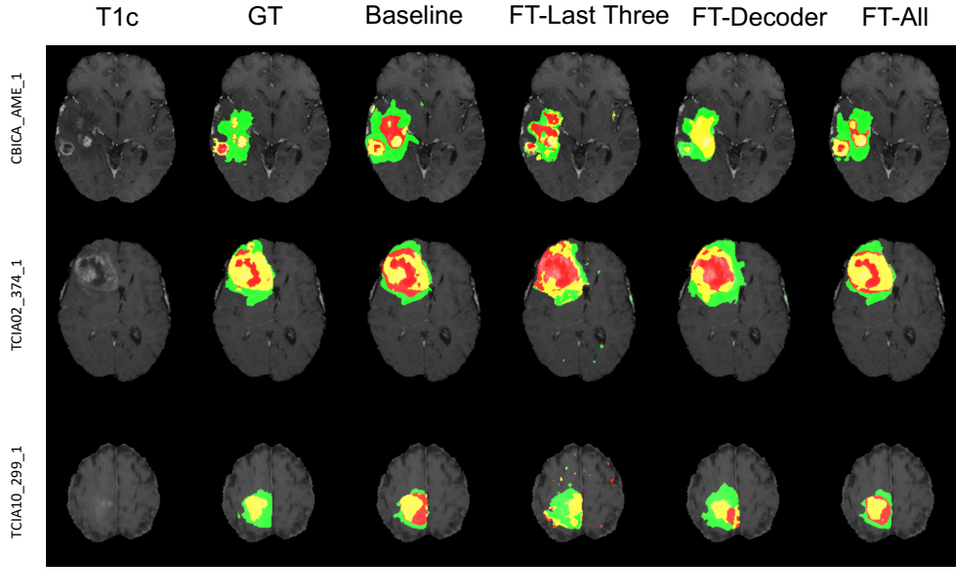
3.4 Discussion

In this chapter, we explore different strategies of fine-tuning for transfer learning across neurodegenerative diseases for the task of focal pathology segmentation. We observe that fine-tuning the entire binary lesion segmentation network trained on a larger MS dataset improves the multi-class brain tumour 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. We also observed that as in the case of natural images, where fine-tuning just the last few layers works, it is not the same case in the medical domain. Moreover, we believe that 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 broad set of applications with small datasets. All the three strategies of fine-tuning are general and can be performed as long as a pre-trained source network is available, or the source dataset is available to train a source network. In this chapter, the methods handle the differences between the nature of inference tasks in the two datasets, i.e. from binary MS lesion segmentation task to multi-class brain tumour segmentation, by replacing the last few layers of the pre-trained network with newly initialized layers having correct output dimensions as per target task. This limits the architecture of the target model to be similar to the pre-trained model, thereby constraining the representation

power of the target network. Nonetheless, there are many sophisticated techniques available in order to tackle the differences in the nature of inference tasks between the two datasets, which are explored in the next chapter.



(a)



(b)

Figure 3.4: Examples of visualizations obtained on a local validation set when fine-tuning with (a) 20 and (b) 150 brain tumour samples, respectively. The top two rows and bottom row illustrate the segmentation results obtained on HGG and LGG cases, respectively. From left to right: T1c MRI (column 1), Expert 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 tumour are shown in green, red and yellow, respectively.

4

Multi-task Learning and Cascaded Networks

In the previous chapter, we saw how knowledge transfer via fine-tuning could be beneficial for improving performance in the task of medical image segmentation across neurodegenerative diseases, particularly in small data regimes. We also discussed how the representation capacity of the target network is constrained to be similar to the source network. Hence, simply fine-tuning this target network with target dataset may not be as effective when the nature of the inference tasks between the two datasets differ significantly. In order to address this issue, this chapter further explores other sophisticated ways of transferring knowledge across diseases where the imaging modalities are similar (e.g. brain MRI), but the structures of interest vary substantially (e.g. lesions, tumours), and the task can vary from binary to multi-label classification. Specifically, we explore different methods for leveraging a large, proprietary, clinical trial dataset of MRIs acquired from patients with Multiple Sclerosis (MS), to improve the results for brain tumour sub-tissue segmentation on a much smaller dataset. We explore several general methods that can be applied for a variety of transfer learning tasks, including fine-tuning the pre-trained network, jointly learning representations for both datasets using multi-task learning, and novel coupling of transfer learning techniques with cascaded networks where sub-networks can be trained in sequence or together. We also present a comprehensive study of these methods by varying the number of samples in the target set available for training.

4.1 Methodology

The goal is to leverage a large source dataset of annotated medical images in order to increase pathology segmentation performance on a much smaller target dataset, where the pathological structures of interest vary (lesions vs tumours), and the inference tasks differ (two-class vs. four-class segmentation). With this aim, we explore several transfer learning techniques, which we detail below.

4.1.1 Multi-task Learning

Multi-task learning [108, 18] is a technique by which shared feature representations are learnt between two or more related tasks, which can ultimately lead to better overall generalization of the network. Multi-task learning imparts an inductive bias in the model to prefer the hypothesis, which can generalize over more than one task. For instance, authors in [87] train a network to learn image reconstruction with semantic segmentation of brain tumours jointly. This is performed by adding a variational autoencoder branch to the network, which helps in regularizing the shared encoder. Their method secured the first position at the BraTS 2018 challenge. Authors in [86] trained a single CNN for joint segmentation of six tissue types in brain MRI, pectoral muscle in breast MRI and coronary arteries from cardiac CTA images with performance comparable to that of a CNN model trained for each task individually. In [131], the authors propose a dual-stream encoder-decoder style architecture to improve segmentation results on multiple organs using multi-modal learning from MRI and CT images.

In this chapter, the effectiveness of multi-task learning is explored in the context of transfer learning across neurodegenerative diseases. Specifically, we use hard parameter sharing methods [108], in which the network consists of an encoder, common to all the tasks at hand, and separate decoders for learning task-specific features. The architecture of the double head network, labelled `DOUBLE`, is shown in Figure 4.1(b). Training a network with more than one task forces it to learn generic features that are common to all the tasks. Moreover, it also allows us to train a single network for performing a variety of tasks instead of training separate networks for every individual task. Although, multi-task learning assumes that data for all the tasks are present at the time of training.

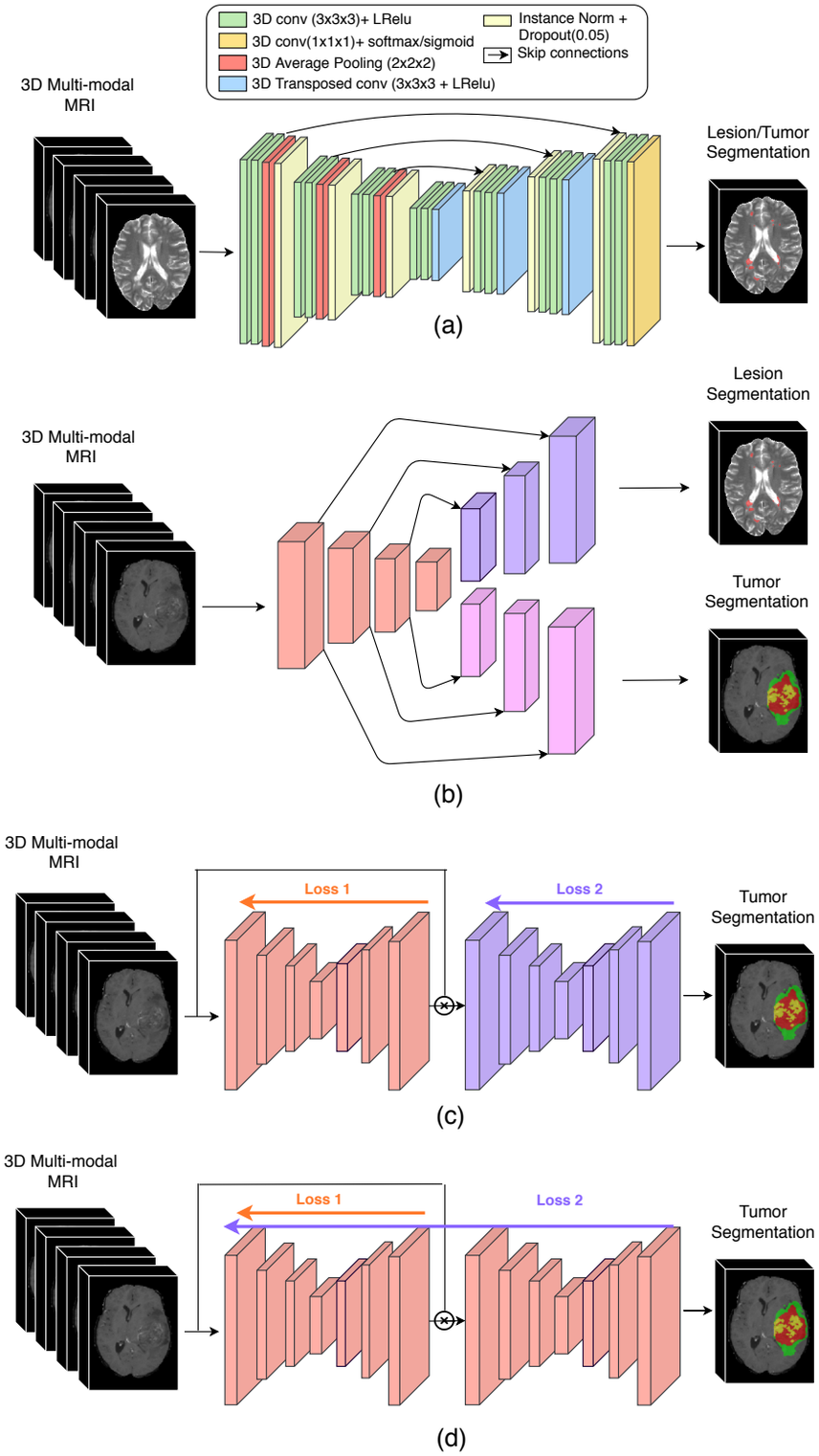


Figure 4.1: (a) `BASE`: UNet architecture for baseline networks. (b) `DOUBLE`: Double head network having a shared encoder and two task-specific decoders (c) `SEQ`: Sequential Cascaded UNet where the UNets are trained one by one in sequence and (d) `SIM`: Simultaneous Cascaded UNet where both the UNets are trained simultaneously.

4.1.2 Cascaded Networks

Cascaded Networks comprise of a series of networks, where every sub-network is trained to handle a specific sub-task. Cascaded networks have shown to be effective in various medical image segmentation problems including brain tumour segmentation [134, 42], liver and lesion segmentation [20], organ and vessel segmentation [106], prenatal fetal head and abdomen ultrasound image segmentation [137] and prostate segmentation [64] to name a few. The main advantage of cascaded networks in medical image segmentation is that they divide the overall segmentation task into a series of sub-tasks and build on the sub-tasks that have already been learned. This enables the cascaded networks to capture different anatomical structures of interest present in medical images at different hierarchical levels and predicting precise segmentation results at the end. In this chapter, we explore different variations of cascaded networks and their novel coupling with transfer learning techniques in order to show their effectiveness in medical image segmentation tasks. More details about this can be found in Section 4.2.5. The first version of cascaded networks that we explore is the *sequential* version in which a first UNet is trained fully before training a second UNet. We use UNets [22] because they have been shown to provide excellent results for medical segmentation tasks in previous works [66]. Once the first UNet is trained, its sigmoid output is multiplied with the input MRI sequences element-wise, and the result is then fed as input to the second UNet for further training. The first UNet is frozen while training the second UNet. This way, the output of the first UNet guides the second UNet, thereby serving a similar role to an attention map when training the second UNet. The second version of cascaded networks, which we call *simultaneous*, involves jointly training both UNets at the same time, with their corresponding objectives. When back-propagating through the second UNet, the gradient is allowed to flow through the first UNet, in this case. Therefore, the output of the first UNet is tailored to facilitate better segmentation on the downstream task of the second UNet. Figure 4.1(c) and (d) present *sequential* and *simultaneous* versions of cascaded networks respectively.

4.2 Experimentation Pipeline

This section talks about the data and its pre-processing steps, followed by a detailed description of the 3D network architecture used and the experimental setup. In order to evaluate the performance of the techniques described in Section 4.1 for the task of pathology segmentation, several experiments are performed using a large source dataset of MS patients and a small public brain tumour dataset. We also compare the performance of different transfer learning approaches to a baseline UNet, on the task of multi-class brain tumour segmentation.

4.2.1 Data Description and Preprocessing

The evaluation of our transfer learning approaches involves the usage of two datasets: Multiple Sclerosis dataset and Brain tumour dataset (BraTS 2018 challenge dataset). The details about these two datasets are already described in section 3.2.1. Pre-processing of data is also done in the same manner as described in section 3.2.1. The only key difference in pre-processing this time is that the images are reduced to $216 \times 176 \times 184$ using zero-padding and cropping operations. This is done to remove the maximum amount of irrelevant background voxels from the MRI images.

4.2.2 Architecture of 3D CNN for Segmentation

The first proposed segmentation network is a variant of 3D UNet, as shown in Figure 4.1(a). An input of 3D MRI sequences is passed to the network, which predicts a 3D output segmentation mask of the same size as the input. The proposed baseline network contains an encoder and a decoder of four resolution steps each. The encoder has two consecutive 3D convolutions at every resolution step (with eight filters initially). Every convolution layer is followed by LReLU activation and average pooling, instance normalization, and a dropout of 0.05 is applied at the end of each step. The decoder uses 3D transposed convolutions for upsampling, whose output is concatenated with the corresponding output of the encoder step. It is followed by instance normalization and a dropout of 0.05, after which two consecutive 3D convolutions with Leaky ReLU activation are applied. The last layer is a $1 \times 1 \times 1$ convolutional layer having C filters, where C is the number of classes. The

output is passed to a softmax/sigmoid layer for the final voxel-wise segmentation prediction. The implementation of all the above models is done in Pytorch[1].

4.2.3 Training the Network

As already explained in sub-section 3.2.3, we use weighted binary cross-entropy loss and weighted cross-entropy loss for two-class and four-class voxel-wise classification, respectively. The initial weight of a class C is determined by the ratio of the total number of voxels to the number of voxels belonging to class C present in the ground truth labels available in the training set. The class weights are scheduled to decay with a decay rate $r \in (0, 1)$ over epochs. This weighing schedule reduces bias and ensures that all classes are given equal importance during the later phases of the training. For the `SEQ` experiment, the first UNet is trained for binary whole tumour segmentation using weighted binary cross-entropy loss function and then kept frozen. The sigmoid output of this frozen UNet is then used for training the second UNet for sub-structure brain tumour segmentation using weighted cross-entropy loss. For the `SIM` experiments, the network follows two levels of training objectives. The first objective only back-propagates the weighted binary cross-entropy loss with respect to whole tumour segmentation in the first UNet. The second objective function, i.e., the weighted cross-entropy loss at the end of the second UNet, for multi-class brain tumour segmentation, gets propagated through both UNets. For `DOUBLE` experiment, binary cross-entropy loss for lesion segmentation output is re-weighted according to the ratio of the number of brain tumour and MS lesion cases to remove biases incurred due to predominance of MS samples.

4.2.4 Evaluation Metrics

As mentioned in sub-section 3.2.4, for the task of segmenting tumours (the binary task of segmenting the whole tumour or segmenting all the different sub-types of tumours), Dice scores are calculated based on the type of tumour of interest.

[1] http://pytorch.org/

4.2.5 Experiments

The baseline experiment(`BASELINE`) involves training a network from scratch for multi-class tumour segmentation on the brain tumour dataset. Besides, another network, denoted `BASE_MS`, is trained from scratch for binary lesion segmentation on the MS dataset. The double head experiment (denoted `DOUBLE`), uses instances from both datasets, which are passed to a common encoder. The resulting encodings are then passed through different decoders, depending on the task (binary segmentation of MS lesions or whole tumour segmentation). The cascaded experiments can be further divided into three types:

1. `SIM_BRATS`: Both UNets are trained simultaneously from scratch with brain tumour data. The first UNet and second UNet are trained in conjunction to perform the whole tumour and multi-class brain tumour segmentation, respectively.

2. `SIM_MS`: This is similar to `SIM_BRATS`, but the first UNet is initialized with the weights of the `BASE_MS` network. Both UNets are then simultaneously fine-tuned/retrained using the brain tumour data, as described in Section 4.1.2.

3. `SEQ_DOUBLE`: Here, the first UNet is initialized with the weights of the double head network, `DOUBLE` (the head corresponding for the whole brain tumour segmentation) and kept frozen. The sigmoid output of this frozen UNet is then used for training the second UNet with brain tumour data for multi-class brain tumour segmentation.

The `BASE_MS` and `DOUBLE` experiments involve the MS dataset. These versions are trained using 80% of the total available MS data (2911 samples), while the remaining 20% is left out for validation (719 samples). For the `DOUBLE` experiment, we combine the MS training set and MS validation set with 228 and 57 samples of brain tumour data, respectively. An AUC of 0.7 is obtained as the top segmentation performance on the validation set for the `BASE_MS` experiment. This pre-trained `BASE_MS` network is then fine-tuned with brain tumour data for the task of segmenting the whole tumour in the `FINETUNE` experiment. For experiments involving the brain tumour dataset (`BASELINE`, `FINETUNE`, `SIM_BRATS`, `SIM_MS`, and `SEQ_DOUBLE`), five-fold cross-validation is performed

for determining the best hyper-parameter setting. The ratio of high-grade glioma patients (HGG) to low-grade glioma patients (LGG) is maintained across all folds. Once the best hyper-parameters are determined, the networks are retrained on the complete BRATS 2018 training set. The details about hyperparameter tuning are presented in section 4.2.5.1. Finally, predictions obtained on the BraTS 2018 validation set are then uploaded on the BraTS 2018 portal for evaluation. We also run experiments on all the above variations using small target datasets for training, of 50 and 20 brain tumour instances respectively, in order to test the effectiveness of all the proposed transfer learning approaches. In this setting, we do not perform five-fold cross-validation, as the size of the training data is very small. Instead, we use a validation set of 50 samples to select the best hyper-parameters (so in effect, the total amount of data used is 100 and 70 examples, considering both training and hyper-parameter tuning).

4.2.5.1 Hyperparameter tuning

For the `BASE_MS` and `DOUBLE` experiments, the networks are trained using an initial learning rate of 1e-3, which is reduced by a factor of 0.75 every 25th epoch. This annealing of the learning rate is performed in order to ensure proper convergence of the model during training. The networks are trained for 100 epochs with a batch size of 1 using SGD with momentum of 0.9 as an optimizer. Due to severe class imbalance in the context of MS lesion segmentation, the ratio of the weight of lesion class and non-lesion class is calculated to be 544:1 in MS training data. The initial class weights are then decayed with rates of 0.95 and 0.92 when training on the MS and brain tumour training sets, respectively. The details about the hyper-parameter tuning of the learning rate are shown in Table 4.1. For 50 and 285 brain tumour training cases, the networks are trained with a batch size of 1 for 100 epochs. For the experiments with 20 brain tumour training samples, the networks are trained for 50 epochs. The learning rate that provides the best performance in terms of Dice scores on the validation set is chosen and fixed.

Methods	Learning rate
base__ms, double	$\{1e-3\}$
baseline, finetune, seq__double	$\{5e-3, 1e-3, 5e-4, 1e-4\}$
sim__brats, sim__ms	$(LR_{1st\ UNet}, LR_{2nd\ UNet}) \in \{(1e-4, 1e-3), (5e-4, 1e-3), (1e-3, 1e-3), (1e-4, 1e-4), (1e-3, 5e-3)\}$

Table 4.1: Hyper-parameter tuning of learning rate (LR) for different methods.

4.3 Results

4.3.1 Quantitative Results

Sample size	20			50			285		
Method	EN	CO	WH	EN	CO	WH	EN	CO	WH
baseline	0.367	0.443	0.789	0.599	0.611	0.83	0.67	0.735	0.88
finetune	n/a	n/a	0.814	n/a	n/a	0.845	n/a	n/a	0.87
double	n/a	n/a	**0.818**	n/a	n/a	**0.856**	n/a	n/a	0.882
sim__brats	0.336	0.48	0.78	0.609	**0.68**	0.825	0.668	0.765	0.872
sim__ms	**0.592**	0.564	0.801	0.594	0.653	0.846	**0.673**	**0.773**	0.883
seq__double	0.531	**0.604**	0.795	**0.63**	0.646	0.853	0.67	**0.773**	**0.886**

Table 4.2: Dice values obtained on the BraTS 2018 Validation for the `BASELINE` compared to all other methods, as a function of the number of brain tumour cases available for training $(20, 50, 285)$. The values marked in bold are the best scores. 'WH', 'CO', and 'EN' stand for whole, core, and enhancing tumours, respectively. N/A means not applicable since it is a two-class whole tumour segmentation problem.

Table 4.2 provides the comparison of all the Dice scores obtained on the BraTS 2018 validation set (used for testing). The results demonstrate that the baseline segmentation performance is surpassed by all the transfer methods in almost every case. Notably, there is a consistent improvement in Dice scores for core tumour in all the proposed methods over baseline. Also, we observe a more significant improvement in all the Dice scores as we decrease the number of samples in the target dataset. Particularly, there is a percentage improvement[2] of 61.3% and 36.3% in the Dice scores for the enhanced and core parts of the tumour, compared to the baseline, when only 20 brain tumour samples are available. We also notice that all the proposed methods outperformed or performed just as well as the baselines most of the time. This suggests that the networks in the proposed methods can leverage aspects that are common among MS lesions and brain tumours. Moreover, in order to

test the effectiveness of the proposed methods against the baseline and reduce the biases induced due to statistical errors, we also calculate the mean Dice scores for the whole, core and enhanced tumour across five-folds based on BraTS 2018 training dataset after performing five-fold cross-validation in case of a total of 285 brain tumour cases. The comparison of results for the same is shown in Table 4.3. Here, every fold has 228 samples for training and 57 samples for validation. The table shows results with a percentage improvement of 4.27% and 3.03% in the mean Dice scores across five folds for the core and whole tumour respectively over the baseline.

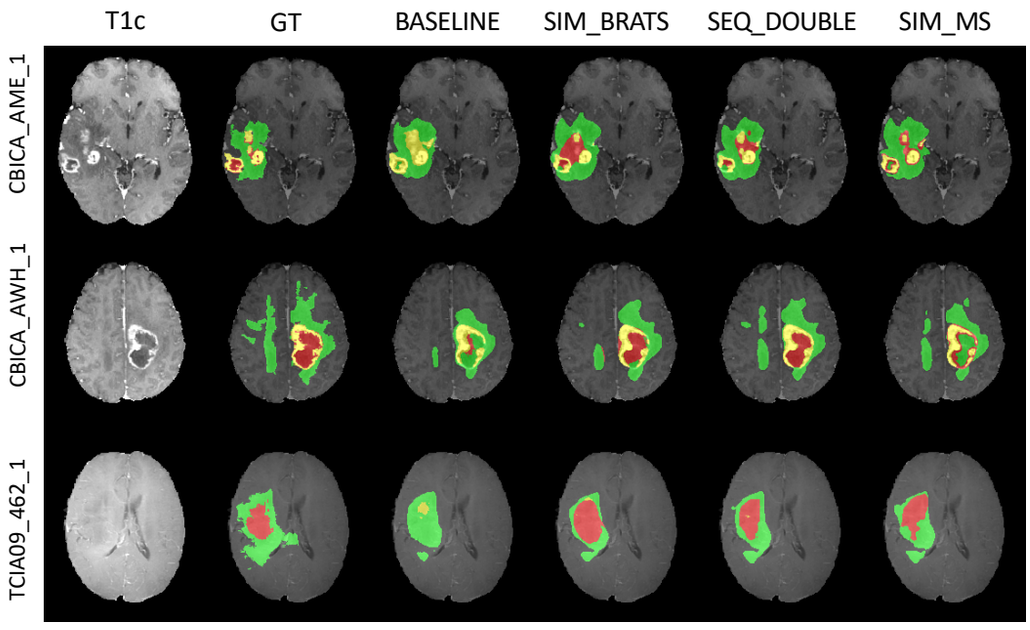
METHOD	EN	CO	WH
BASELINE	**0.618**	0.725	0.856
FINETUNE	N/A	N/A	0.865
DOUBLE	N/A	N/A	0.878
SIM__BRATS	0.592	0.715	0.855
SIM__MS	0.603	0.755	0.864
SEQ__DOUBLE	0.614	**0.756**	**0.882**

Table 4.3: Mean Dice scores obtained across five-folds based on the BraTS 2018 training dataset for different types of tumours after performing five-fold cross-validation in case of a total of 285 brain tumour cases. The values marked in bold are the best scores. 'WH', 'CO', and 'EN' stand for whole, core, and enhancing tumours, respectively. N/A means not applicable since it is a two-class whole tumour segmentation problem.

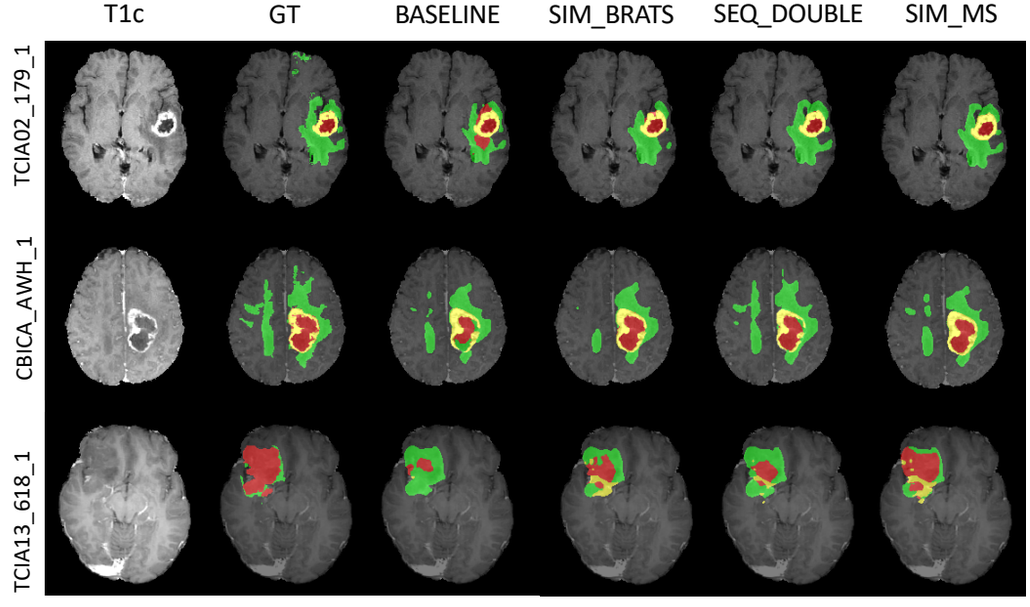
4.3.2 Qualitative Results

Figure 4.2 shows visualizations of the results from the different methods obtained on examples from the BraTS 2018 training dataset when (a) 50 and (b) 285 brain tumour cases are available for training. It is evident that `BASELINE` either over-segments or under-segments the core tumour (in red), thereby performing poorly. At the same time, the proposed methods are better at capturing different sub-structures of the tumour. This also falls in line with the quantitative results discussed before. Overall, all the methods perform better on the HGG cases over the LGG ones. This is due to the abundance of HGG cases in the training dataset.

[2]The percentage improvement is measured as the difference in the baseline and best Dice score obtained across the proposed methods, divided by the baseline result.



(a)



(b)

Figure 4.2: Examples of visualizations obtained on the BraTS 2018 training dataset when 50 (a) and 285 (b) brain tumour cases are available. The top two rows and the bottom row in each figure present the tumour segmentations obtained on HGG and LGG cases, respectively. Green, red and yellow colours represent edema, necrotic core and enhancing tumour respectively. From left to right column: T1c MRI , Ground truth , results of `BASELINE`, `SIM_BRATS`, `SEQ_DOUBLE` and, `SIM_MS` are shown.

4.4 Discussion

In this chapter, we explored a variety of transfer learning techniques for the task of focal pathology segmentation across different neurodegenerative diseases, where the pathological structures of interest vary (e.g. lesions, tumours), and the tasks are different (two-class vs. four-class segmentation). Specifically, we leveraged a deep learning network pre-trained on a large MRI dataset of MS patients in order to improve multi-class brain tumour segmentation on a much smaller public dataset. We showed that cascaded networks, as well as learning joint representations combined with a double head network for each task, substantially improve brain tumour segmentation results, especially when very few instances are available for training. These results show the potential of transfer learning for broadening the application of deep nets in medical imaging.

5

Conclusions and Future Work

In this thesis, we presented an exploration of a variety of transfer learning techniques for segmenting pathological structures of interest across neurodegenerative diseases. The main contribution of this thesis is to show the effectiveness of various transfer learning strategies, such as fine-tuning a pre-trained network and multi-task joint representation learning using double head networks in medical image segmentation based applications, and also to propose a novel augmentation of the above networks in a cascaded approach. We analyzed several ways of fine-tuning to determine the best way of leveraging a source network pre-trained on a large MS dataset, in order to improve segmentation performance on a target brain tumour dataset. The two image datasets are acquired from patients with different neurological diseases, the pathological structures of interest vary in the two datasets (lesions vs. tumours), and the inference tasks differ (binary vs. multi-class segmentation). Extensive quantitative and qualitative analysis as a function of the target dataset size brings several key takeaways, detailed below. First, we observe that fine-tuning the whole network works better than fine-tuning only a portion of the pre-trained network, especially when small target datasets are available. This means that fine-tuning just the last few layers may not be as effective in the medical domain as in the case of natural images. Second, cascaded networks, as well as learning joint representations using multi-tasking, substantially improve brain tumour segmentation results in the case of small target datasets. Our results show the potential of transfer learning and the significant impact it can have, particularly for diseases where there is limited access to large scale, annotated datasets needed for training segmentation networks from scratch. These methods are

general and can be applied to other medical image segmentation tasks such as brain stroke, or white matter hyperintensity segmentation. Furthermore, their application can be extended to transfer knowledge between different organs or different grades of the same disease, or for domain adaptation between medical images obtained from different scanners/hospitals, thereby further broadening the application of deep nets in medical imaging.

One possible direction for extending the work done in this thesis is to explore different kinds of loss functions or their combinations, instead of simply using the cross-entropy loss. For example, using the Dice loss or a combination of the Dice loss and cross-entropy loss could help in boosting segmentation performance, by producing crisper boundaries. Another immediate experiment that could be performed is multi-class brain tumour segmentation in the `DOUBLE` dataset, instead of binary whole tumour segmentation. For the experiments involving the coupling of cascaded networks with transfer learning, `SIM_MS`, `SEQ_DOUBLE`, only the first UNet is initialized with the weights of a pre-trained network. This initialization could also be applied to the second UNet. In this way, both the UNets would be initialized with a better starting point, which could lead to faster convergence as well.

One drawback of the techniques discussed in this thesis, especially multi-task learning and cascaded networks, is an increase in the number of parameters needed to train in the target segmentation task. Another drawback is that the network architectures for the source and target networks are taken to be similar. One possible direction to address these problems is to answer concretely the questions of what to transfer from the source network and where to transfer it in the target model. The first question deals with identifying the particular layers in the pre-trained network that can be useful for the target task. The second part deals with finding the relative positioning of those selected layers in the target network. This would enable network architectures to be different in the source and the target task, which could improve the segmentation performance. It would also be interesting to see how the approaches explored in this work perform when transferring across different imaging modalities (MRI to CT) or different organs (pancreas to liver, for example).

Bibliography

[1] Varghese Alex, Kiran Vaidhya, Subramaniam Thirunavukkarasu, Chandrasekharan Kesavadas, and Ganapathy Krishnamurthi. "Semisupervised learning using denoising autoencoders for brain lesion detection and segmentation". In: *Journal of Medical Imaging* 4.4 (2017), p. 041311.

[2] Raja S Alomari, Suryaprakash Kompalli, and Vipin Chaudhary. "Segmentation of the liver from abdominal CT using Markov random field model and GVF snakes". In: *2008 International Conference on Complex, Intelligent and Software Intensive Systems*. IEEE. 2008, pp. 293–298.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2425–2433.

[4] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. "A reproducible evaluation of ANTs similarity metric performance in brain image registration". In: *Neuroimage* 54.3 (2011), pp. 2033–2044.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation". In: *IEEE transactions on pattern analysis and machine intelligence* 39.12 (2017), pp. 2481–2495.

[6] Wenjia Bai, Chen Chen, Giacomo Tarroni, Jinming Duan, Florian Guitton, Steffen E Petersen, Yike Guo, Paul M Matthews, and Daniel Rueckert. "Self-Supervised learning for cardiac MR image segmentation by anatomical position prediction". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 541–549.

[7] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. "Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features". In: *Scientific data* 4 (2017), p. 170117.

[8] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. "Segmentation labels and radiomic features for the pre-operative scans of the TCGA-LGG collection". In: *The cancer imaging archive* 286 (2017).

[9] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. "Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge". In: *arXiv preprint arXiv:1811.02629* (2018).

[10] Eric Bardinet, Laurent D Cohen, and Nicholas Ayache. "A parametric deformable model to fit unstructured 3D data". In: *Computer vision and image understanding* 71.1 (1998), pp. 39–54.

[11] Dhruv Batra, Rahul Sukthankar, and Tsuhan Chen. "Learning class-specific affinities for image labelling". In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. 2008, pp. 1–8.

[12] Stefan Bauer, Thomas Fejes, Johannes Slotboom, Roland Wiest, Lutz-P Nolte, and Mauricio Reyes. "Segmentation of brain tumor images based on integrated hierarchical classification and regularization". In: *MICCAI BraTS Workshop. Nice: Miccai Society*. 2012, p. 11.

[13] Pierre-Louis Bazin and Dzung L Pham. "Statistical and topological atlas based brain image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2007, pp. 94–101.

[14] Hakan Bilen and Andrea Vedaldi. "Integrated perception with recurrent multi-task neural networks". In: *Advances in neural information processing systems*. 2016, pp. 235–243.

[15] Tom Brosch, Lisa YW Tang, Youngjin Yoo, David KB Li, Anthony Traboulsee, and Roger Tam. "Deep 3D convolutional encoder networks with shortcuts for multiscale feature integration applied to multiple sclerosis lesion segmentation". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1229–1239.

[16] Paul Browne, Dhia Chandraratna, Ceri Angood, Helen Tremlett, Chris Baker, Bruce V Taylor, and Alan J Thompson. "Atlas of multiple sclerosis 2013: a growing global problem with widespread inequity". In: *Neurology* 83.11 (2014), pp. 1022–1024.

[17] Aaron Carass, Snehashis Roy, Amod Jog, Jennifer L Cuzzocreo, Elizabeth Magrath, Adrian Gherman, Julia Button, James Nguyen, Ferran Prados, Carole H Sudre, et al. "Longitudinal multiple sclerosis lesion segmentation: resource and challenge". In: *NeuroImage* 148 (2017), pp. 77–102.

[18] Rich Caruana. "Multitask learning". In: *Machine learning* 28.1 (1997), pp. 41–75.

[19] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. "Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis". In: *Medical image analysis* 54 (2019), pp. 280–296.

[20] Patrick Ferdinand Christ, Florian Ettlinger, Felix Grün, Mohamed Ezzeldin A Elshaera, Jana Lipkova, Sebastian Schlecht, Freba Ahmaddy, Sunil Tatavarty, Marc Bickel, Patrick Bilic, et al. "Automatic liver and tumor segmentation of CT and MRI volumes using cascaded fully convolutional neural networks". In: *arXiv preprint arXiv:1702.05970* (2017).

[21] Brian Chu, Vashisht Madhavan, Oscar Beijbom, Judy Hoffman, and Trevor Darrell. "Best practices for fine-tuning visual classifiers to new domains". In: *European conference on computer vision*. Springer. 2016, pp. 435–442.

[22] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. "3D U-Net: learning dense volumetric segmentation from sparse annotation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2016, pp. 424–432.

[23] Guy Barrett Coleman and Harry C Andrews. “Image segmentation by clustering”. In: *Proceedings of the IEEE* 67.5 (1979), pp. 773–785.

[24] Jifeng Dai, Kaiming He, and Jian Sun. “Instance-aware semantic segmentation via multi-task network cascades”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3150–3158.

[25] Christos Davatzikos and N Bryan. “Using a deformable surface model to obtain a shape representation of the cortex”. In: *IEEE transactions on medical imaging* 15.6 (1996), pp. 785–795.

[26] Huawu Deng and David A Clausi. “Unsupervised image segmentation using a simple MRF model with a new implementation scheme”. In: *Pattern recognition* 37.12 (2004), pp. 2323–2335.

[27] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “Imagenet: A large-scale hierarchical image database”. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee. 2009, pp. 248–255.

[28] Carl Doersch and Andrew Zisserman. “Multi-task self-supervised visual learning”. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017, pp. 2051–2060.

[29] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. “Long-term recurrent convolutional networks for visual recognition and description”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 2625–2634.

[30] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. “Automatic brain tumor detection and segmentation using u-net based fully convolutional networks”. In: *annual conference on medical image understanding and analysis*. Springer. 2017, pp. 506–517.

[31] Omar Emad, Inas A Yassine, and Ahmed S Fahmy. “Automatic localization of the left ventricle in cardiac MRI images using deep learning”. In: *2015 37th Annual International*

Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE. 2015, pp. 683–686.

[32] Daniel Garcia-Lorenzo, Simon Francis, Sridar Narayanan, Douglas L Arnold, and D Louis Collins. "Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging". In: *Medical image analysis* 17.1 (2013), pp. 1–18.

[33] Stuart Geman and Donald Geman. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In: *IEEE Transactions on pattern analysis and machine intelligence* 6 (1984), pp. 721–741.

[34] Ezequiel Geremia, Olivier Clatz, Bjoern H Menze, Ender Konukoglu, Antonio Criminisi, and Nicholas Ayache. "Spatial decision forests for MS lesion segmentation in multi-channel magnetic resonance images". In: *NeuroImage* 57.2 (2011), pp. 378–390.

[35] Mohsen Ghafoorian, Nico Karssemeijer, Tom Heskes, Inge WM van Uden, Clara I Sanchez, Geert Litjens, Frank-Erik de Leeuw, Bram van Ginneken, Elena Marchiori, and Bram Platel. "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities". In: *Scientific Reports* 7.1 (2017), pp. 1–12.

[36] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. "Transfer learning for domain adaptation in mri: Application in brain lesion segmentation". In: *International conference on medical image computing and computer-assisted intervention*. Springer. 2017, pp. 516–524.

[37] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. "Rich feature hierarchies for accurate object detection and semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 580–587.

[38] Xavier Glorot and Yoshua Bengio. "Understanding the difficulty of training deep feedforward neural networks". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 2010, pp. 249–256.

[39] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[40] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets". In: *Advances in neural information processing systems*. 2014, pp. 2672–2680.

[41] Stephen L Hauser and Jorge R Oksenberg. "The neurobiology of multiple sclerosis: genes, inflammation, and neurodegeneration". In: *Neuron* 52.1 (2006), pp. 61–76.

[42] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. "Brain tumor segmentation with deep neural networks". In: *Medical image analysis* 35 (2017), pp. 18–31.

[43] Mohammad Havaei, Nicolas Guizard, Nicolas Chapados, and Yoshua Bengio. "Hemis: Hetero-modal image segmentation". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 469–477.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 1026–1034.

[45] Karsten Held, E Rota Kops, Bernd J Krause, William M Wells, Ron Kikinis, and H-W Muller-Gartner. "Markov random field segmentation of brain MR images". In: *IEEE transactions on medical imaging* 16.6 (1997), pp. 878–886.

[46] Judy Hoffman, Erik Rodner, Jeff Donahue, Brian Kulis, and Kate Saenko. "Asymmetric and category invariant feature transformations for domain adaptation". In: *International journal of computer vision* 109.1-2 (2014), pp. 28–41.

[47] Jiayuan Huang, Arthur Gretton, Karsten Borgwardt, Bernhard Schölkopf, and Alex J Smola. "Correcting sample selection bias by unlabeled data". In: *Advances in neural information processing systems*. 2007, pp. 601–608.

[48] Qixing Huang, Mei Han, Bo Wu, and Sergey Ioffe. "A hierarchical conditional random field model for labeling and segmenting images of street scenes". In: *CVPR 2011*. IEEE. 2011, pp. 1953–1960.

[49] Zhongling Huang, Zongxu Pan, and Bin Lei. "Transfer learning with deep convolutional neural network for SAR target classification with limited labeled data". In: *Remote Sensing* 9.9 (2017), p. 907.

[50] Saddam Hussain, Syed Muhammad Anwar, and Muhammad Majid. "Brain tumor segmentation using cascaded deep convolutional neural network". In: *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE. 2017, pp. 1998–2001.

[51] Benjamin Q Huynh, Hui Li, and Maryellen L Giger. "Digital mammographic tumor classification using transfer learning from deep convolutional neural networks". In: *Journal of Medical Imaging* 3.3 (2016), p. 034501.

[52] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[53] Anil K Jain and Farshid Farrokhnia. "Unsupervised texture segmentation using Gabor filters". In: *1990 IEEE international conference on systems, man, and cybernetics conference proceedings*. IEEE. 1990, pp. 14–19.

[54] Andrew Jesson and Tal Arbel. "Brain tumor segmentation using a 3D FCN with multi-scale loss". In: *International MICCAI Brainlesion Workshop*. Springer. 2017, pp. 392–402.

[55] Micheline Kamber, Rajjan Shinghal, D Louis Collins, Gordon S Francis, and Alan C Evans. "Model-based 3-D segmentation of multiple sclerosis lesions in magnetic resonance brain images". In: *IEEE Transactions on Medical Imaging* 14.3 (1995), pp. 442–453.

[56] Konstantinos Kamnitsas, Christian Baumgartner, Christian Ledig, Virginia Newcombe, Joanna Simpson, Andrew Kane, David Menon, Aditya Nori, Antonio Criminisi, Daniel Rueckert, et al. "Unsupervised domain adaptation in brain lesion segmentation with adversarial networks".

In: *International conference on information processing in medical imaging*. Springer. 2017, pp. 597–609.

[57] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical image analysis* 36 (2017), pp. 61–78.

[58] Andrej Karpathy. "Convolutional neural networks (cnns/convnets)". In: *CS231n Convolutional Neural Networks for Visual Recognition* (2016).

[59] Junmo Kim, John W Fisher, Anthony Yezzi, Müjdat Çetin, and Alan S Willsky. "A non-parametric statistical method for image segmentation using information theory and curve evolution". In: *IEEE Transactions on Image processing* 14.10 (2005), pp. 1486–1502.

[60] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[61] Hugo J Kuijf, J Matthijs Biesbroek, Jeroen De Bresser, Rutger Heinen, Simon Andermatt, Mariana Bento, Matt Berseth, Mikhail Belyaev, M Jorge Cardoso, Adria Casamitjana, et al. "Standardized Assessment of Automatic Segmentation of White Matter Hyperintensities and Results of the WMH Segmentation Challenge". In: *IEEE transactions on medical imaging* 38.11 (2019), pp. 2556–2568.

[62] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[63] Chulhee Lee, Shin Huh, Terence A Ketter, and Michael Unser. "Unsupervised connectivity-based thresholding segmentation of midsagittal brain MR images". In: *Computers in biology and medicine* 28.3 (1998), pp. 309–338.

[64] Suiyi Li, Yuxuan Chen, Su Yang, and Wuyang Luo. "Cascade dense-Unet for prostate segmentation in MR images". In: *International Conference on Intelligent Computing*. Springer. 2019, pp. 481–490.

[65] Fedde van der Lijn, Tom Den Heijer, Monique MB Breteler, and Wiro J Niessen. "Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts". In: *Neuroimage* 43.4 (2008), pp. 708–720.

[66] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. "A survey on deep learning in medical image analysis". In: *Medical image analysis* 42 (2017), pp. 60–88.

[67] Xin Liu, Deanna L Langer, Masoom A Haider, Yongyi Yang, Miles N Wernick, and Imam Samil Yetik. "Prostate cancer segmentation with simultaneous estimation of Markov random field parameters and class". In: *IEEE Transactions on Medical Imaging* 28.6 (2009), pp. 906–915.

[68] Jonathan Long, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 3431–3440.

[69] Fred D Lublin, Stephen C Reingold, Jeffrey A Cohen, Gary R Cutter, Per Soelberg Sørensen, Alan J Thompson, Jerry S Wolinsky, Laura J Balcer, Brenda Banwell, Frederik Barkhof, et al. "Defining the clinical course of multiple sclerosis: the 2013 revisions". In: *Neurology* 83.3 (2014), pp. 278–286.

[70] Rafael Marcos Luque, Enrique Dominguez, Esteban J Palomo, and Jose Munoz. "A neural network approach for video object segmentation in traffic surveillance". In: *International Conference Image Analysis and Recognition*. Springer. 2008, pp. 151–158.

[71] Mark Lyksborg, Oula Puonti, Mikael Agn, and Rasmus Larsen. "An ensemble of 2D convolutional neural networks for tumor segmentation". In: *Scandinavian Conference on Image Analysis*. Springer. 2015, pp. 201–211.

[72] Oskar Maier, Bjoern H Menze, Janina von der Gablentz, Levin Häni, Mattias P Heinrich, Matthias Liebrand, Stefan Winzeck, Abdul Basit, Paul Bentley, Liang Chen, et al. "ISLES 2015-A public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI". In: *Medical image analysis* 35 (2017), pp. 250–269.

[73] IN Manousakas, PE Undrill, GG Cameron, and TW Redpath. "Split-and-merge segmentation of magnetic resonance medical images: performance evaluation and extension to three dimensions". In: *Computers and Biomedical Research* 31.6 (1998), pp. 393–412.

[74] Maciej A Mazurowski, Mateusz Buda, Ashirbani Saha, and Mustafa R Bashir. "Deep learning in radiology: An overview of the concepts and a survey of the state of the art with focus on MRI". In: *Journal of magnetic resonance imaging* 49.4 (2019), pp. 939–954.

[75] Richard McKinley, Rik Wepfer, Tom Gundersen, Franca Wagner, Andrew Chan, Roland Wiest, and Mauricio Reyes. "Nabla-net: A deep dag-like convolutional architecture for biomedical image segmentation". In: *International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer. 2016, pp. 119–128.

[76] Michael F McNitt-Gray, HK Huang, and James W Sayre. "Feature selection in the pattern classification problem of digital chest radiograph segmentation". In: *IEEE Transactions on Medical Imaging* 14.3 (1995), pp. 537–547.

[77] Raghav Mehta and Tal Arbel. "3D U-Net for Brain Tumour Segmentation". In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 254–266.

[78] Afonso Menegola, Michel Fornaciali, Ramon Pires, Flávia Vasques Bittencourt, Sandra Avila, and Eduardo Valle. "Knowledge transfer for melanoma screening with deep learning". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 297–300.

[79] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. "The multimodal brain tumor image segmentation benchmark (BRATS)". In: *IEEE transactions on medical imaging* 34.10 (2014), pp. 1993–2024.

[80] Bjoern H Menze, Koen Van Leemput, Danial Lashkari, Marc-André Weber, Nicholas Ayache, and Polina Golland. "A generative model for brain tumor segmentation in multi-modal images". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2010, pp. 151–159.

[81] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. "V-net: Fully convolutional neural networks for volumetric medical image segmentation". In: *2016 Fourth International Conference on 3D Vision (3DV)*. IEEE. 2016, pp. 565–571.

[82] Ron Milo and Esther Kahana. "Multiple sclerosis: geoepidemiology, genetics and the environment". In: *Autoimmunity reviews* 9.5 (2010), A387–A394.

[83] Ajay K Mishra and Yiannis Aloimonos. "Visual segmentation of "Simple" objects for robots". In: *Robotics: Science and Systems VII* 217 (2012).

[84] Ishan Misra, Abhinav Shrivastava, Abhinav Gupta, and Martial Hebert. "Cross-stitch networks for multi-task learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 3994–4003.

[85] Pabitra Mitra, B Uma Shankar, and Sankar K Pal. "Segmentation of multispectral remote sensing images using active support vector machines". In: *Pattern recognition letters* 25.9 (2004), pp. 1067–1074.

[86] Pim Moeskops, Jelmer M Wolterink, Bas HM van der Velden, Kenneth GA Gilhuijs, Tim Leiner, Max A Viergever, and Ivana Išgum. "Deep learning for multi-task medical image segmentation in multiple modalities". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2016, pp. 478–486.

[87] Andriy Myronenko. "3D MRI brain tumor segmentation using autoencoder regularization". In: *International MICCAI Brainlesion Workshop*. Springer. 2018, pp. 311–320.

[88] Tanya Nair, Doina Precup, Douglas L Arnold, and Tal Arbel. "Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation". In: *Medical image analysis* 59 (2020), p. 101557.

[89] László G Nyúl, Jayaram K Udupa, and Xuan Zhang. "New variants of a method of MRI scale standardization". In: *IEEE transactions on medical imaging* 19.2 (2000), pp. 143–150.

[90] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. "Learning and transferring mid-level image representations using convolutional neural networks". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2014, pp. 1717–1724.

[91] Sinno Jialin Pan, Ivor W Tsang, James T Kwok, and Qiang Yang. "Domain adaptation via transfer component analysis". In: *IEEE Transactions on Neural Networks* 22.2 (2010), pp. 199–210.

[92] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[93] Imelda Parrenas. *Parakeets*. URL: https://www.pinterest.ca/pin/102879172708195511/ (visited on 03/26/2020).

[94] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. "Brain tumor segmentation using convolutional neural networks in MRI images". In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1240–1251.

[95] Martino Pesaresi and Jon Atli Benediktsson. "A new approach for the morphological segmentation of high-resolution satellite imagery". In: *IEEE transactions on Geoscience and Remote Sensing* 39.2 (2001), pp. 309–320.

[96] Dzung L Pham and Jerry L Prince. "An adaptive fuzzy C-means algorithm for image segmentation in the presence of intensity inhomogeneities". In: *Pattern recognition letters* 20.1 (1999), pp. 57–68.

[97] DL Phan, Chenyang Xu, and J Price. "A survey of current methods in medical image segmentation". In: *Annual review of biomedical engineering* (1998).

[98] Nils Plath, Marc Toussaint, and Shinichi Nakajima. "Multi-class image segmentation using conditional random fields and global classification". In: *Proceedings of the 26th Annual International Conference on Machine Learning*. 2009, pp. 817–824.

[99] Marcel Prastawa, Elizabeth Bullitt, Sean Ho, and Guido Gerig. "A brain tumor segmentation framework based on outlier detection". In: *Medical image analysis* 8.3 (2004), pp. 275–283.

[100] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. "CNN features off-the-shelf: an astounding baseline for recognition". In: *2014 IEEE conference on computer vision and pattern recognition workshops*. IEEE. 2014, pp. 512–519.

[101] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. "Faster r-cnn: Towards real-time object detection with region proposal networks". In: *Advances in neural information processing systems*. 2015, pp. 91–99.

[102] Eduardo Ribeiro, Michael Häfner, Georg Wimmer, Toru Tamaki, Jens JW Tischendorf, Shigeto Yoshida, Shinji Tanaka, and Andreas Uhl. "Exploring texture transfer learning for colonic polyp classification via convolutional neural networks". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 1044–1048.

[103] Herbert Robbins and Sutton Monro. "A stochastic approximation method". In: *The annals of mathematical statistics* (1951), pp. 400–407.

[104] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation". In: *International Conference on Medical image computing and computer-assisted intervention*. Springer. 2015, pp. 234–241.

[105] Holger R Roth, Amal Farag, Le Lu, Evrim B Turkbey, and Ronald M Summers. "Deep convolutional networks for pancreas segmentation in CT imaging". In: *Medical Imaging 2015: Image Processing*. Vol. 9413. International Society for Optics and Photonics. 2015, 94131G.

[106] Holger R Roth, Hirohisa Oda, Xiangrong Zhou, Natsuki Shimizu, Ying Yang, Yuichiro Hayashi, Masahiro Oda, Michitaka Fujiwara, Kazunari Misawa, and Kensaku Mori. "An application of cascaded 3D fully convolutional networks for medical image segmentation". In: *Computerized Medical Imaging and Graphics* 66 (2018), pp. 90–99.

[107] Artem Rozantsev, Mathieu Salzmann, and Pascal Fua. "Beyond sharing weights for deep domain adaptation". In: *IEEE transactions on pattern analysis and machine intelligence* 41.4 (2018), pp. 801–814.

[108] Sebastian Ruder. "An overview of multi-task learning in deep neural networks". In: *arXiv preprint arXiv:1706.05098* (2017).

[109] Prasanna K Sahoo, SAKC Soltani, and Andrew KC Wong. "A survey of thresholding techniques". In: *Computer vision, graphics, and image processing* 41.2 (1988), pp. 233–260.

[110] Jakob Santner, Markus Unger, Thomas Pock, Christian Leistner, Amir Saffari, and Horst Bischof. "Interactive Texture Segmentation using Random Forests and Total Variation." In: *BMVC*. Citeseer. 2009, pp. 1–12.

[111] Jürgen Schmidhuber. "Deep learning in neural networks: An overview". In: *Neural networks* 61 (2015), pp. 85–117.

[112] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. "CNN features off-the-shelf: an astounding baseline for recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 2014, pp. 806–813.

[113] Neeraj Sharma and Lalit M Aggarwal. "Automated medical image segmentation techniques". In: *Journal of medical physics/Association of Medical Physicists of India* 35.1 (2010), p. 3.

[114] Amber L Simpson, Michela Antonelli, Spyridon Bakas, Michel Bilello, Keyvan Farahani, Bram Van Ginneken, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, et al. "A large annotated medical image dataset for the development and evaluation of segmentation algorithms". In: *arXiv preprint arXiv:1902.09063* (2019).

[115] John G Sled, Alex P Zijdenbos, and Alan C Evans. "A nonparametric method for automatic correction of intensity nonuniformity in MRI data". In: *IEEE transactions on medical imaging* 17.1 (1998), pp. 87–97.

[116] Stephen M Smith. "Fast robust automated brain extraction". In: *Human brain mapping* 17.3 (2002), pp. 143–155.

[117] Canadian Cancer Society. *Brain and spinal tumours*. 2020. URL: https://www.cancer.ca/en/cancer-information/cancer-type/brain-spinal/brain-and-spinal-tumours/?region=qc (visited on 03/26/2020).

[118] National Brain Tumor Society. *Understanding brain tumors*. 2020. URL: https://braintumor.org/brain-tumor-information/understanding-brain-tumors/ (visited on 03/26/2020).

[119] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[120] BWKP Stewart, Christopher P Wild, et al. "World cancer report 2014". In: (2014).

[121] Martin Styner, Joohwi Lee, Brian Chin, M Chin, Olivier Commowick, H Tran, S Markovic-Plese, V Jewells, and S Warfield. "3D segmentation in the clinic: A grand challenge II: MS lesion segmentation". In: *Midas Journal* 2008 (2008), pp. 1–6.

[122] Nagesh Subbanna, Doina Precup, Douglas Arnold, and Tal Arbel. "IMaGe: iterative multilevel probabilistic graphical model for detection and segmentation of multiple sclerosis lesions in brain MRI". In: *International Conference on Information Processing in Medical Imaging*. Springer. 2015, pp. 514–526.

[123] Nagesh Subbanna, M Shah, SJ Francis, Sridar Narayanan, DL Collins, DL Arnold, and Tal Arbel. "MS lesion segmentation using Markov Random Fields". In: *Proceedings of International Conference on Medical Image Computing and Computer Assisted Intervention, London, UK*. 2009.

[124] Jasjit S Suri. "Computer vision, pattern recognition and image processing in left ventricle segmentation: The last 50 years". In: *Pattern Analysis & Applications* 3.3 (2000), pp. 209–242.

[125] Nima Tajbakhsh, Jae Y Shin, Suryakanth 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?" In: *IEEE transactions on medical imaging* 35.5 (2016), pp. 1299–1312.

[126] Phi Vu Tran. "A fully convolutional neural network for cardiac segmentation in short-axis MRI". In: *arXiv preprint arXiv:1604.00494* (2016).

[127] Stefano Trebeschi, Joost JM van Griethuysen, Doenja MJ Lambregts, Max J Lahaye, Chintan Parmar, Frans CH Bakers, Nicky HGM Peters, Regina GH Beets-Tan, and Hugo JWL Aerts. "Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR". In: *Scientific reports* 7.1 (2017), pp. 1–9.

[128] Michael Treml, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, et al. "Speeding up semantic segmentation for autonomous driving". In: *MLITS, NIPS Workshop*. Vol. 2. 2016, p. 7.

[129] Gregor Urban, M Bendszus, F Hamprecht, and J Kleesiek. "Multi-modal brain tumor segmentation using deep convolutional neural networks". In: *MICCAI BraTS (brain tumor segmentation) challenge. Proceedings, winning contribution* (2014), pp. 31–35.

[130] Suthirth Vaidya, Abhijith Chunduru, Ramanathan Muthuganapathy, and Ganapathy Krishnamurthi. "Longitudinal multiple sclerosis lesion segmentation using 3D convolutional neural networks". In: *Proceedings of the 2015 Longitudinal Multiple Sclerosis Lesion Segmentation Challenge* (2015), pp. 1–2.

[131] Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. "Multi-modal learning from unpaired images: Application to multi-organ segmentation in CT and MRI". In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. 2018, pp. 547–556.

[132] Sergi Valverde, Mariano Cabezas, Eloy Roura, Sandra Gonzalez-Villa, Deborah Pareto, Joan C Vilanova, Lluis Ramio-Torrenta, Alex Rovira, Arnau Oliver, and Xavier Llado. "Improving automated multiple sclerosis lesion segmentation with a cascaded 3D convolutional neural network approach". In: *NeuroImage* 155 (2017), pp. 159–168.

[133] Christian Wachinger, Martin Reuter, and Tassilo Klein. "DeepNAT: Deep convolutional neural network for segmenting neuroanatomy". In: *NeuroImage* 170 (2018), pp. 434–445.

[134] Guotai Wang, Wenqi Li, Sébastien Ourselin, and Tom Vercauteren. "Automatic brain tumor segmentation using cascaded anisotropic convolutional neural networks". In: *International MICCAI brainlesion workshop*. Springer. 2017, pp. 178–190.

[135] Xiang-Yang Wang, Ting Wang, and Juan Bu. "Color image segmentation using pixel wise support vector machine classification". In: *Pattern Recognition* 44.4 (2011), pp. 777–787.

[136] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. "A survey of transfer learning". In: *Journal of Big data* 3.1 (2016), p. 9.

[137] Lingyun Wu, Yang Xin, Shengli Li, Tianfu Wang, Pheng-Ann Heng, and Dong Ni. "Cascaded fully convolutional networks for automatic prenatal ultrasound image segmentation". In: *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*. IEEE. 2017, pp. 663–666.

[138] Rui Xia, Chengqing Zong, Xuelei Hu, and Erik Cambria. "Feature ensemble plus sample selection: domain adaptation for sentiment classification". In: *IEEE Intelligent Systems* 28.3 (2013), pp. 10–18.

[139] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention". In: *International conference on machine learning*. 2015, pp. 2048–2057.

[140] Jianhua Xuan, Tülay Adali, and Yue Wang. "Segmentation of magnetic resonance brain image: integrating region growing and edge detection". In: *Proceedings., International Conference on Image Processing*. Vol. 3. IEEE. 1995, pp. 544–547.

[141] Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W Cohen, Ruslan Salakhutdinov, and Yann LeCun. "Glomo: Unsupervisedly learned relational graphs as transferable representations". In: *arXiv preprint arXiv:1806.05662* (2018).

[142] Yi Yao and Gianfranco Doretto. "Boosting for transfer learning with multiple sources". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE. 2010, pp. 1855–1862.

[143] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. "How transferable are features in deep neural networks?" In: *Advances in neural information processing systems*. 2014, pp. 3320–3328.

[144] Lequan Yu, Xin Yang, Hao Chen, Jing Qin, and Pheng Ann Heng. "Volumetric ConvNets with mixed residual connections for automated prostate segmentation from 3D MR images". In: *Thirty-first AAAI conference on artificial intelligence*. 2017.

[145] Ruikai Zhang, Yali Zheng, Tony Wing Chung Mak, Ruoxi Yu, Sunny H Wong, James YW Lau, and Carmen CY Poon. "Automatic detection and classification of colorectal polyps by transferring low-level CNN features from nonmedical domain". In: *IEEE journal of biomedical and health informatics* 21.1 (2016), pp. 41–47.

[146] Yongyue Zhang, Michael Brady, and Stephen Smith. "Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm". In: *IEEE transactions on medical imaging* 20.1 (2001), pp. 45–57.

[147] Zongwei Zhou, Vatsal Sodha, Md Mahfuzur Rahman Siddiquee, Ruibin Feng, Nima Tajbakhsh, Michael B Gotway, and Jianming Liang. "Models genesis: Generic autodidactic models for 3d medical image analysis". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 384–393.

[148] Darko Zikic, Yani Ioannou, Matthew Brown, and Antonio Criminisi. "Segmentation of brain tumor tissues with convolutional neural networks". In: *Proceedings MICCAI-BRATS* (2014), pp. 36–39.