Efficient Multi-Scale 3D CNN with fully connected CRF for Accurate Brain Lesion Segmentation

Konstantinos Kamnitsas[a], Christian Ledig[a], Virginia F.J. Newcombe[b,c],
Joanna P. Simpson[b], Andrew D. Kane[b], David K. Menon[b,c],
Daniel Rueckert[a], Ben Glocker[a]

[a] *Biomedical Image Analysis Group, Imperial College London, UK*
[b] *University Division of Anaesthesia, Department of Medicine, Cambridge University, UK*
[c] *Wolfson Brain Imaging Centre, Cambridge University, UK*

Abstract

We propose a dual pathway, 11-layers deep, three-dimensional Convolutional Neural Network for the challenging task of brain lesion segmentation. The devised architecture is the result of an in-depth analysis of the limitations of current networks proposed for similar applications. To overcome the computational burden of processing 3D medical scans, we have devised an efficient and effective dense training scheme which joins the processing of adjacent image patches into one pass through the network while automatically adapting to the inherent class imbalance present in the data. Further, we analyze the development of deeper, thus more discriminative 3D CNNs. In order to incorporate both local and larger contextual information, we employ a dual pathway architecture that processes the input images at multiple scales simultaneously. For post-processing of the network's soft segmentation, we use a 3D fully connected Conditional Random Field which effectively removes false positives. Our pipeline is extensively evaluated on three challenging tasks of lesion segmentation in multi-channel MRI patient data with traumatic brain injuries, brain tumors, and ischemic stroke. We improve on the state-of-the-art for all three applications, with top ranking performance on the public benchmarks BRATS 2015 and ISLES 2015. Our method is computationally efficient, which allows its adoption in a variety of research and clinical settings. The source code of our implementation is made publicly available.

Keywords: 3D Convolutional Neural Network, Fully Connected CRF, Segmentation, Brain Lesions, Deep Learning

1. Introduction

Segmentation and the subsequent quantitative assessment of lesions in medical images provide valuable information for the analysis of neuropathologies and are important for planning of treatment strategies, monitoring of disease progression and prediction of patient outcome. For a better understanding of the pathophysiology of diseases, quantitative imaging can reveal clues about the disease characteristics and effects on particular anatomical structures. For example, the associations of different lesion types, their spatial distribution and extent with acute and chronic sequelae after traumatic brain injury (TBI) are still poorly understood (Maas et al. (2015)). However, there is growing evidence that quantification of lesion burden may add insight into the functional outcome of patients (Ding et al. (2008); Moen et al. (2012)). Additionally, exact locations of injuries relate to particular deficits depending on the brain structure that is affected (Lehtonen et al. (2005); Warner et al. (2010); Sharp et al. (2011)). This is in line with estimates that functional deficits caused by stroke are associated with the extent of damage to particular parts of the brain (Carey et al. (2013)). Lesion burden is commonly quantified by means of volume and number of lesions, biomarkers that have been shown to be related to cognitive deficits. For example, volume of white matter lesions (WML) correlates with cognitive decline and increased risk of dementia (Ikram et al. (2010)). In clinical research on multiple sclerosis (MS), lesion count and volume are used to analyse disease progression and effectiveness of pharmaceutical treatment (Rovira and León (2008); Kappos et al. (2007)). Finally, accurate delineation of the pathology is important in the case of brain tumors, where estimation of the relative volume of a tumor's sub-components is required for planning radiotherapy and treatment follow-up (Wen et al. (2010)).

The quantitative analysis of lesions requires accurate lesion segmentation in multi-modal, three-dimensional images which is a challenging task for a number of reasons. The heterogeneous appearance of lesions including the large variability in location, size, shape and frequency make it difficult to devise effective segmentation rules. It is thus highly non-trivial to delineate contusions, edema and haemorrhages in TBI (Irimia et al. (2012)), or sub-components of brain tumors such as proliferating cells and necrotic core (Menze et al. (2015)). The arguably most accurate segmentation results can be obtained through manual delineation by a human expert which is tedious, expensive, time-consuming, impractical in larger studies, and intro-

duces inter-observer variability. Additionally, for deciding whether a particular region is part of a lesion multiple image sequences with varying contrasts need to be considered, and the level of expert knowledge and experience are important factors that impact segmentation accuracy. Hence, in clinical routine often only qualitative, visual inspection, or at best crude measures like approximate lesion volume and number of lesions are used (Yuh et al. (2012); Wen et al. (2010)). In order to capture and better understand the complexity of brain pathologies it is important to conduct large studies with many subjects to gain the statistical power for drawing conclusions across a whole patient population. The development of accurate, automatic segmentation algorithms has therefore become a major research focus in medical image computing with the potential to offer objective, reproducible, and scalable approaches to quantitative assessment of brain lesions.

Figure 1 illustrates some of the challenges that arise when devising a computational approach for the task of automatic lesion segmentation. The figure summarizes statistics and shows examples of brain lesions in the case of TBI, but is representative of other pathologies such as brain tumors and ischemic stroke. Lesions can occur at multiple sites, with varying shapes and sizes, and their image intensity profiles largely overlap with non-affected, healthy parts of the brain or lesions which are not in the focus of interest. For example, stroke and MS lesions have a similar hyper-intense appearance in FLAIR sequences as other WMLs (Mitra et al. (2014); Schmidt et al. (2012)). It is generally difficult to derive statistical prior information about lesion shape and appearance. On the other hand, in some applications there is an expectation on the spatial configuration of segmentation labels, for example there is a hierarchical layout of sub-components in brain tumors. Ideally, a computational approach is able to adjust itself to application specific characteristics by learning from a set of a few example images.

1.1. Related Work

A multitude of automatic lesion segmentation methods have been proposed over the last decade, and several main categories of approaches can be identified. One group of methods poses the lesion segmentation task as an abnormality detection problem, for example by employing image registration. The early work of Prastawa et al. (2004) and more recent ones by Schmidt et al. (2012) and Doyle et al. (2013) align the pathological scan to a healthy atlas and lesions are detected based on deviations in tissue appearance between the patient and the atlas image. Lesions, however, may cause
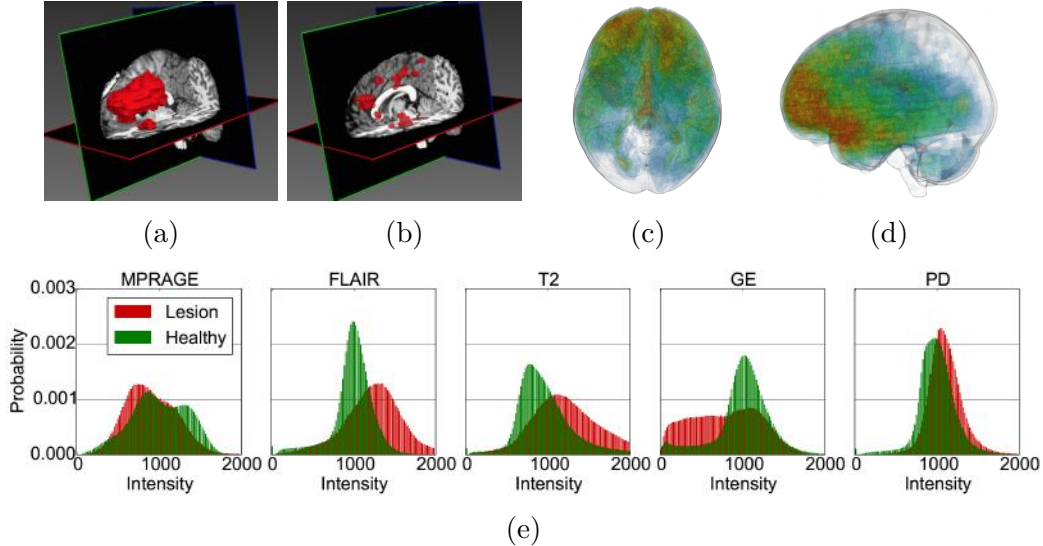
Figure 1: Heterogeneous appearance of TBI lesions poses challenges in devising discriminative models. Lesion size varies significantly with both large, focal and small, diffused lesions (a,b). Alignment of manual lesion segmentations reveals the wide spatial distribution of lesions in (c,d) with some areas being more likely than others. (e) shows the average of the normalized intensity histograms of different MR channels over all the TBI cases in our database, for healthy (green) and injured (red) tissue. One can observe a large overlap between the distributions of healthy and non-healthy tissue.

large structural deformations that may lead to incorrect segmentation due to incorrect registration. Gooya et al. (2011); Parisot et al. (2012) alleviate this problem by jointly solving the segmentation and registration tasks. Liu et al. (2014) showed that registration together with a low-rank decomposition gives as a by-product the abnormal structures in the sparse components, although, this may not be precise enough for detection of small lesions. Abnormality detection has also been proposed within image synthesis works. Representative approaches are those of Weiss et al. (2013) using dictionary learning and Ye et al. (2013) using a patch-based approach. The idea is to synthesize pseudo-healthy images that when compared to the patient scan allow to highlight abnormal regions. In this context, Cardoso et al. (2015) present a generative model for image synthesis that yields a probabilistic seg-

mentation of abnormalities. Another unsupervised technique is proposed by Erihov et al. (2015), a saliency-based method that exploits brain asymmetry in pathological cases. A common advantage of the above methods is that they do not require a training dataset with corresponding manual annotations. In general, these approaches are more suitable for detecting lesions rather than accurately segmenting them.

Some of the most successful, supervised segmentation methods for brain lesions are based on voxel-wise classifiers, such as Random Forests. Representative work is that of Geremia et al. (2010) on MS lesions, employing intensity features to capture the appearance of the region around each voxel. Zikic et al. (2012) combine this with a generative Gaussian Mixture Model (GMM) to obtain tissue-specific probabilistic priors (Van Leemput et al. (1999)). This framework was adopted in multiple works, with representative pipelines for brain tumors by Tustison et al. (2013) and TBI by Rao et al. (2014). Both works incorporate morphological and contextual features to better capture the heterogeneity of lesions. Rao et al. (2014) also incorporate brain structure segmentation results obtained from a multi-atlas label propagation approach (Ledig et al. (2015)) to provide strong tissue-class priors to the Random Forests. Tustison et al. (2013) additionally use a Markov Random Field (MRF) to incorporate spatial regularization. MRFs are commonly used to encourage spatial continuity of the segmentation (Schmidt et al. (2012); Mitra et al. (2014)). Although those methods have been very successful, it appears that their modeling capabilities still have significant limitations. This is confirmed by the results of the most recent challenges [1], and also by our own experience and experimentation with such approaches.

At the same time, deep learning techniques have emerged as a powerful alternative for supervised learning with great model capacity and the ability to learn highly discriminative features for the task at hand. These features often outperform hand-crafted and pre-defined feature sets. In particular, Convolutional Neural Networks (CNNs) (LeCun et al. (1998); Krizhevsky et al. (2012)) have been applied with promising results on a variety of biomedical imaging problems. Ciresan et al. (2012) presented the first GPU implementation of a two-dimensional CNN for the segmentation of neural membranes. From the CNN based work that followed, related to our approach are the methods of Zikic et al. (2014); Havaei et al. (2015); Pereira et al. (2015),

---

[1]links: http://braintumorsegmentation.org/, www.isles-challenge.org

with the latter being the best performing automatic approach in the BRATS 2015 challenge (Menze et al. (2015)). These methods are based on 2D CNNs, which have been used extensively in computer vision applications on natural images. Here, the segmentation of a 3D brain scan is achieved by processing each 2D slice independently, which is arguably a non-optimal use of the volumetric medical image data. Despite the simplicity in the architecture, the promising results obtained by these methods indicate the potential of CNNs.

Fully 3D CNNs come with an increased number of parameters and significant memory and computational requirements. Previous work discusses problems and apparent limitations when employing a 3D CNN on medical imaging data (Prasoon et al. (2013); Li et al. (2014); Roth et al. (2014)). To incorporate 3D contextual information, multiple works used 2D CNNs on three orthogonal 2D patches (Prasoon et al. (2013); Roth et al. (2014); Lyksborg et al. (2015)). In their work for structural brain segmentation, Brebisson and Montana (2015) extracted large 2D patches from multiple scales of the image and combined them with small single-scale 3D patches, in order to avoid the memory requirements of fully 3D networks.

One of the reasons that discouraged the use of 3D CNNs is the slow inference due to the computationally expensive 3D convolutions. In contrast to the 2D/3D hybrid variants (Roth et al. (2014); Brebisson and Montana (2015)), 3D CNNs can fully exploit *dense-inference* (LeCun et al. (1998); Sermanet et al. (2014)), a technique that greatly decreases inference times and which we will further discuss in section 2.1. By employing dense-inference with 3D CNNs, Brosch et al. (2015) and Urban et al. (2014) reported computation times of a few seconds and approximately a minute respectively for the processing of a single brain scan. Even though the size of their developed networks was limited, a factor that is directly related to a network's representational power, their results on MS and brain tumor segmentation respectively were very promising.

Performance of CNNs is significantly influenced by the strategy for extracting training samples. A commonly adopted approach is training on image patches that are equally sampled from each class. This, however, biases the classifier towards rare classes and may result in over-segmentation. To counter this, Cireşan et al. (2013) proposes to train a second CNN on samples with a class distribution close to the real one, but oversample pixels that were incorrectly classified in the first stage. A secondary training stage was also suggested by Havaei et al. (2015), who retrain the classification layer on patches extracted uniformly from the image. In practice, two stage train-

6

ing schemes can be prone to overfitting and sensitive to the state of the first classifier. Alternatively, *dense training* (Long et al. (2015)) has been used to train a network on multiple or all voxels of a single image per optimisation step (Urban et al. (2014); Brosch et al. (2015); Ronneberger et al. (2015)). This can introduce severe class imbalance, similarly to uniform sampling. Weighted cost functions have been proposed in the two latter works to alleviate this problem. Brosch et al. (2015) manually adjusted the sensitivity of the network, but the method can become difficult to calibrate for multi-class problems. Ronneberger et al. (2015) first balance the cost from each class, which has an effect similar to equal sampling, and further adjust it for the specific task by estimating the difficulty of segmenting each pixel.

1.2. Contributions

We present a fully automatic approach for lesion segmentation in multi-modal brain MRI based on an 11-layers deep, multi-scale, 3D CNN with the following main contributions:

1. We propose an efficient hybrid training scheme, utilizing *dense training* (Long et al. (2015)) on sampled image segments, and analyze its behaviour in adapting to class imbalance of the segmentation problem at hand.
2. We analyze in depth the development of deeper, thus more discriminative, yet computationally efficient 3D CNNs. We exploit the utilization of small kernels, a design approach previously found beneficial in 2D networks (Simonyan and Zisserman (2014)) that impacts 3D CNNs even more, and present adopted solutions that enable training deeper networks.
3. We employ parallel convolutional pathways for multi-scale processing, a solution to efficiently incorporate both local and contextual information which greatly improves segmentation results.
4. We demonstrate the generalization capabilities of our system, which without significant modifications outperforms the state-of-the-art on a variety of challenging segmentation tasks, with top ranking results in two MICCAI challenges, ISLES and BRATS.

Furthermore, a detailed analysis of the network reveals valuable insights into the powerful black box of deep learning with CNNs. For example, we have found that our network is capable of learning very complex, high level

features that separate gray matter (GM), cerebrospinal fluid (CSF) and other anatomical structures to identify the image regions corresponding to lesions.

Additionally, we have extended the fully-connected Conditional Random Field (CRF) model by Krähenbühl and Koltun (2011) to 3D which we use for final post-processing of the CNN's soft segmentation maps. This CRF overcomes limitations of previous models as it can handle arbitrarily large neighborhoods while preserving fast inference times. To the best of our knowledge, this is the first use of a fully connected CRF on medical data.

To facilitate further research and encourage other researchers to build upon our results, the source code of our lesion segmentation method including the CNN and the 3D fully connected CRF is made publicly available on https://biomedia.doc.ic.ac.uk/software/deepmedic/.

2. Method

Our proposed lesion segmentation method consists of two main components, a 3D CNN that produces highly accurate, soft segmentation maps, and a fully connected 3D CRF that imposes regularization constraints on the CNN output and produces the final hard segmentation labels. The main contributions of our work are within the CNN component which we describe first in the following.

2.1. 3D CNNs for Dense Segmentation – Setting the Baseline

CNNs produce estimates for the voxel-wise segmentation labels by classifying each voxel in an image independently taking the neighborhood, i.e. local and contextual image information, into account. This is achieved by sequential convolutions of the input with multiple filters at the cascaded layers of the network. Each layer $l \in [1, L]$ consists of C_l *feature maps* (FMs), also referred to as *channels*. Every FM is a group of neurons that detects a particular pattern, i.e. a feature, in the channels of the previous layer. The pattern is defined by the kernel weights associated with the FM. If the neurons of the m-th FM in the l-th layer are arranged in a 3D grid, their activations constitute the image $\mathbf{y}_l^m = f(\sum_{n=1}^{C_{l-1}} \mathbf{k}_l^{m,n} \star \mathbf{y}_{l-1}^n + b_l^m)$. This is the result of convolving each of the previous layer's channels with a 3-dimensional *kernel* $\mathbf{k}_l^{m,n}$, adding a learned bias b_l^m and applying a non-linearity f. Each kernel is a matrix of learned hidden weights $\mathbf{W}_l^{m,n}$. The images $\mathbf{y}_0^n$, input to the first layer, correspond to the channels of the original input image, for instance a multi-sequence 3D MRI scan of the brain. The concatenation of

the kernels $\mathbf{k}_l = (\mathbf{k}_l^{m,1}, ..., \mathbf{k}_l^{m,C_{l-1}})$ can be viewed as a 4-dimensional kernel convolving the concatenated channels $\mathbf{y}_{l-1} = (\mathbf{y}_{l-1}^1, ..., \mathbf{y}_{l-1}^{C_{l-1}})$, which then intuitively expresses that the neurons of higher layers combine the patterns extracted in previous layers, which results in the detection of increasingly more complex patterns. The activations of the neurons in the last layer L correspond to particular segmentation class labels, hence this layer is also referred to as the classification layer. The neurons are thus grouped in C_L FMs, one for each of the segmentation classes. Their activations are fed into a position-wise softmax function that produces the predicted posterior $p_c(\mathbf{x}) = \exp(\mathbf{y}_L^c(\mathbf{x}))/\sum_{c=1}^{C_L} \exp(\mathbf{y}_L^c(\mathbf{x}))$ for each class c, which form soft segmentation maps with (pseudo-)probabilities. $\mathbf{y}_L^c(\mathbf{x})$ is the activation of the c-th classification FM at position $\mathbf{x} \in \mathbb{N}^3$. This baseline network is depicted in Fig. 2.
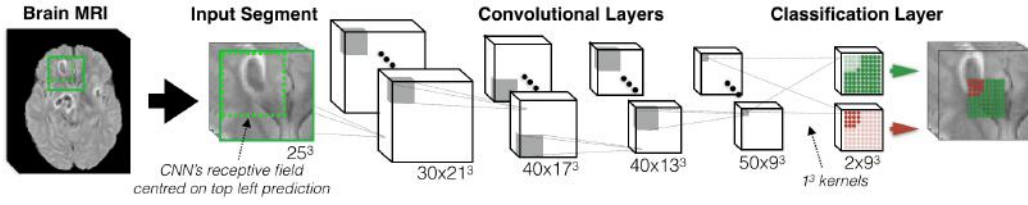

Figure 2: Our baseline CNN consists of four layers with 5^3 kernels for feature extraction, leading to a receptive field of size 17^3. The classification layer is implemented as convolutional with 1^3 kernels, which enables efficient *dense-inference*. When the network segments an input it predicts multiple voxels simultaneously, one for each shift of its receptive field over the input. Number of FMs and their size depicted as (*Number* × *Size*).

The neighborhood of voxels in the input that influence the activation of a neuron is its *receptive field*. Its size, $\boldsymbol{\varphi}_l$, increases at each subsequent layer l and is given by the 3-dimensional vector:

$$\boldsymbol{\varphi}_l^{\{x,y,z\}} = \boldsymbol{\varphi}_{l-1}^{\{x,y,z\}} + (\boldsymbol{\kappa}_l^{\{x,y,z\}} - 1)\boldsymbol{\tau}_l^{\{x,y,z\}} \ , \tag{1}$$

where $\boldsymbol{\kappa}_l, \boldsymbol{\tau}_l \in \mathbb{N}^3$ are vectors expressing the size of the kernels and stride of the receptive field at layer l. $\boldsymbol{\tau}_l$ is given by the product of the strides of kernels in layers preceding l. In this work only unary strides are used, as larger strides downsample the FMs (Springenberg et al. (2014)), which is unwanted behaviour for accurate segmentation. Thus in our system $\boldsymbol{\tau}_l = (1, 1, 1)$. The receptive field of a neuron in the classification layer corresponds to the image

patch that influences the prediction for its central voxel. This is called the *CNN's receptive field*, with $\boldsymbol{\varphi}_{CNN} = \boldsymbol{\varphi}_L$.

If input of size $\boldsymbol{\delta}_{in}$ is provided, the dimensions of the FMs in layer l are given by:

$$\boldsymbol{\delta}_l^{\{x,y,z\}} = \lfloor(\boldsymbol{\delta}_{in}^{\{x,y,z\}} - \boldsymbol{\varphi}_l^{\{x,y,z\}})/\boldsymbol{\tau}_l^{\{x,y,z\}} + 1\rfloor \tag{2}$$

In the common patch-wise classification setting, an input patch of size $\boldsymbol{\delta}_{in} = \boldsymbol{\varphi}_{CNN}$ is provided and the network outputs a single prediction for its central voxel. In this case the classification layer consists of FMs with size 1^3. Networks that are implemented as fully-convolutionals are capable of *dense-inference*, which is performed when input of size greater than $\boldsymbol{\varphi}_{CNN}$ is provided (Sermanet et al. (2014)). In this case, the dimensions of FMs increase according to Eq. (2). This includes the classification FMs which then output multiple predictions simultaneously, one for each stride of the CNN's receptive field on the input (Fig. 2). All predictions are equally trustworthy, as long as the receptive field is fully contained within the input and captures only original content, i.e. no padding is used. This strategy significantly reduces the computational costs and memory loads since the otherwise repeated computations of convolutions on the same voxels in overlapping patches are avoided. Optimal performance is achieved if the whole image is scanned in one forward pass. If GPU memory constraints do not allow it, such as in the case of large 3D networks where a large number of FMs need to be cached, the volume is tiled in multiple *image-segments*, which are larger than individual patches, but small enough to fit into memory.

Before analyzing how we exploit the above dense-inference technique for training, which is the first main contribution of our work, we present the commonly used setting in which CNNs are trained patch-by-patch. Random patches of size $\boldsymbol{\varphi}_{CNN}$ are extracted from the training images. A *batch* is formed out of B of these samples, which is then processed by the network for one training iteration of Stochastic Gradient Descent (SGD). This step aims to alter the network's parameters $\boldsymbol{\Theta}$, such as weights and biases, in order to maximize the log likelihood of the data or, equally, minimize the Cross Entropy via the cost function:

$$J(\boldsymbol{\Theta}; \mathbf{I}^i, c^i) = -\frac{1}{B}\sum_{i=1}^{B} \log\left(P(Y = c^i | \mathbf{I}^i, \boldsymbol{\Theta})\right) = -\frac{1}{B}\sum_{i=1}^{B} \log(p_{c^i}) , \tag{3}$$

where the pair $(\mathbf{I}^i, c^i), \forall i \in [1, B]$ is the i-th patch in the batch and the true

label of its central voxel, while the scalar value p_{c^i} is the predicted posterior for class c^i. Regularization terms were omitted for simplicity. Multiple sequential optimization steps over different batches gradually lead to convergence.

2.2. Dense Training on Image Segments and Class Balance

Larger training batch sizes B are preferred as they approximate the overall data more accurately and lead to better estimation of the true gradient by SGD. However, the memory requirement and computation time increase with the batch size. This limitation is especially relevant for 3D CNNs, where only a few dozens of patches can be processed within reasonable time on modern GPUs.

To overcome this problem, we devise a training strategy that exploits the dense inference technique on image segments. Following from Eq. (2), if an image segment of size greater than $\boldsymbol{\varphi}_{CNN}$ is given as input to our network, the output is a posterior probability for multiple voxels $V = \prod_{i=\{x,y,z\}} \boldsymbol{\delta}_L^{(i)}$. If the training batches are formed of B segments extracted from the training images, the cost function (3) in the case of *dense-training* becomes:

$$ J_D(\boldsymbol{\Theta}; \mathbf{I}_s, \mathbf{c}_s) = -\frac{1}{B \cdot V} \sum_{s=1}^{B} \sum_{v=1}^{V} \log(p_{c_s^v}(\mathbf{x}^v)) \ , \tag{4} $$

where $\mathbf{I}_s$ and $\mathbf{c}_s$ are the s-th segment of the batch and the true labels of its V predicted voxels respectively. c_s^v is the true label of the v-th voxel, $\mathbf{x}^v$ the corresponding position in the classification FMs and $p_{c_s^v}$ the output of the softmax function. The effective batch size is increased by a factor of V without a corresponding increase in computational and memory requirements, as earlier discussed in Sec. 2.1. Notice that this is a hybrid scheme between the commonly used training on individual patches and the dense training scheme on a whole image (Long et al. (2015)), with the latter being problematic to apply for training large 3D CNNs on volumes of high resolution due to memory limitations.

An appealing consequence of this scheme is that the sampling of input segments provides a flexible and automatic way to balance the distribution of training samples from different segmentation classes which is an important issue that directly impacts the segmentation accuracy. Specifically, we build the training batches by extracting segments from the training images with 50% probability being centred on a foreground or background voxel,
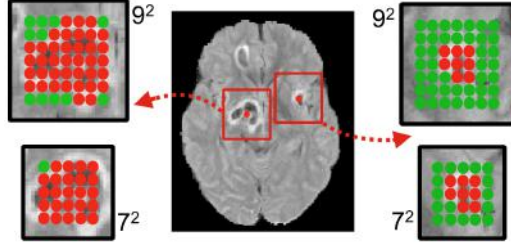
Figure 3: Consider a network with a 2D receptive field of 3^2 (for illustration) densely-applied on the depicted lesion-centred image segments of size 7^2 or 9^2. Relatively more background (green) is captured by larger segments and around smaller lesions.

alleviating class-imbalance. Note that the predicted voxels V in a segment do not have to be of the same class, something that occurs when a segment is sampled from a region near class boundaries (Fig. 3). Hence, the sampling rate of the proposed hybrid method adjusts to the true distribution of the segmentation task's classes. Specifically, the smaller a labelled object, the more background voxels will be captured within segments centred on the foreground voxel. Implicitly, this yields a balance between sensitivity and specificity in the case of binary segmentation tasks. In multi-class problems, the rate at which different classes are captured within a segment centred on foreground reflects the real relative distribution of the foreground classes, while adjusting their frequency relatively to the background.

2.3. Building Deeper Networks

Deeper networks have greater discriminative power due to the additional non-linearities and better quality of local optima (Choromanska et al. (2015)). However, convolutions with 3D kernels are computationally expensive in comparison to the 2D variants, which hampers the addition of more layers. Additionally, 3D architectures have a larger number of trainable parameters, with each layer adding $C_l C_{l-1} \prod_{i=\{x,y,z\}} \kappa_l^{(i)}$ weights to the model. C_l is the number of FMs in layer l and $\kappa_l^{\{x,y,z\}}$ the size of its kernel in the respective spatial dimension. Overall this makes the network increasingly prone to over-fitting.

In order to build a deeper 3D architecture, we adopt the sole use of small 3^3 kernels that are faster to convolve with and contain less weights. This design approach was previously found beneficial for classification of natural

images (Simonyan and Zisserman (2014)) but its effect is even more drastic on 3D networks. When compared to common kernel choices of 5^3 (Zikic et al. (2014); Urban et al. (2014); Prasoon et al. (2013)) and in our baseline CNN, the smaller 3^3 kernels reduce the element-wise multiplications by a factor of approximately $5^3/3^3 \approx 4.6$ while reducing the number of trainable parameters by the same factor. Thus deeper network variants that are implicitly regularised and more efficient can be designed by simply replacing each layer of common architectures with more layers that use smaller kernels (Fig. 4).
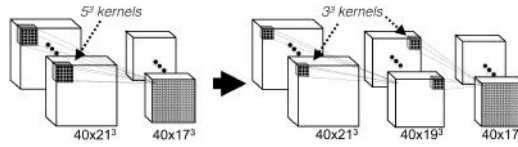

Figure 4: The replacement of the depicted layer with 5^5 kernels (left) with two successive layers using 3^3 kernels (right) introduces an additional non-linearity without altering the CNN's receptive field. Additionally, the number of weights is reduced from 200k to 86.4k and the required convolutions are cheaper (see text). Number of FMs and their size depicted as (*Number* × *Size*).

However, deeper networks are more difficult to train. It has been shown that the forward (neuron activations) and backwards (gradients) propagated signal may explode or vanish if care is not given to retain its variance (Glorot and Bengio (2010)). This occurs because at every successive layer l, the variance of the signal is multiplied by $n_l^{in} \cdot var(\mathbf{W}_l)$, where $n_l^{in} = C_{l-1} \prod_{i=\{x,y,z\}} \boldsymbol{\kappa}_l^{(i)}$ is the number of weights through which a neuron of layer l is connected to its input and $var(\mathbf{W}_l)$ is the variance of the layer's weights. To better preserve the signal in the initial training stage we adopt a scheme recently derived for ReLu-based networks by He et al. (2015) and initialize the kernel weights of our system by sampling from the normal distribution $\mathcal{N}(0, \sqrt{2/n_l^{in}})$.

A phenomenon of similar nature that hinders the network's performance is the "internal covariate shift" (Ioffe and Szegedy (2015)). It occurs throughout training, because the weight updates to deeper layers result in a continuously changing distribution of signal at higher layers, which hinders the convergence of their weights. Specifically, at training iteration t the weight updates may cause deviation $\epsilon_{l,t}$ to the variance of the weights. At the next iteration the signal will be amplified by $n_l^{in} \cdot var(\mathbf{W}_{l,t+1}) = n_l^{in} \cdot (var(\mathbf{W}_{l,t}) + \epsilon_{l,t})$. Thus

before influencing the signal, any deviation $\epsilon_{l,t}$ is amplified by n_l^{in} which is exponential in the number of dimensions. For this reason the problem affects training of 3D CNNs more severely than conventional 2D systems. For countering it, we adopt the recently proposed Batch Normalisation (BN) technique to all hidden layers (Ioffe and Szegedy (2015)), which allows normalization of the FM activations at every optimization step in order to better preserve the signal.

2.4. Multi-Scale Processing via Parallel Convolutional Pathways

The segmentation of each voxel is performed by taking into account the contextual information that is captured by the receptive field of the CNN when it is centred on the voxel. The spatial context is providing important information for being able to discriminate voxels that otherwise appear very similar when considering only local appearance. From Eq. (1) follows that an increase of the CNN's receptive field requires bigger kernels or more convolutional layers, which increases computation and memory requirements. An alternative would be the use of pooling (LeCun et al. (1998)), which however leads to loss of the exact position of the segmented voxel and thus can negatively impact accuracy.

In order to incorporate both local and larger contextual information into our 3D CNN, we add a second pathway that operates on down-sampled images. Thus, our dual pathway 3D CNN simultaneously processes the input image at multiple scales (Fig. 5). Higher level features such as the location within the brain are learned in the second pathway, while the detailed local appearance of structures is captured in the first. As the two pathways are decoupled in this architecture, arbitrarily large context can be processed by the second pathway by simply adjusting the down-sampling factor F_D. The size of the pathways can be independently adjusted according to the computational capacity and the task at hand, which may require relatively more or less filters focused on the down-sampled context.

To preserve the capability of dense inference, spatial correspondence of the activations in the FMs of the last convolutional layers of the two pathways, $L1$ and $L2$, should be ensured. In networks where only unary kernel strides are used, such as the proposed architecture, this requires that for every F_D shifts of the receptive field $\boldsymbol{\varphi}_{L1}$ over the normal resolution input, only one shift is performed by $\boldsymbol{\varphi}_{L2}$ over the down-sampled input. Hence it is required that the dimensions of the FMs in $L2$ are $\boldsymbol{\delta}_{L2}^{\{x,y,z\}} = \lceil \boldsymbol{\delta}_{L1}^{\{x,y,z\}} / F_D \rceil$. From Eq. (2), the size of the input to the second pathway is $\boldsymbol{\delta}_{in2}^{\{x,y,z\}} = \boldsymbol{\varphi}_{L2}^{\{x,y,z\}} +$
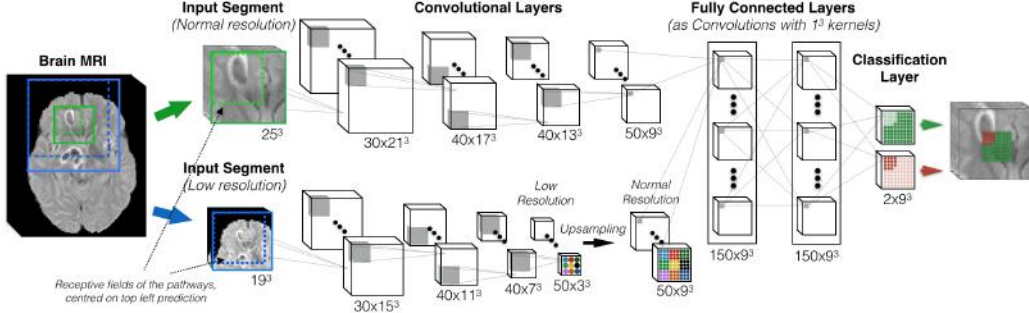
14

Figure 5: Multi-scale 3D CNN with two convolutional pathways. The kernels of the two pathways are here of size 5^3 (for illustration only to reduce the number of layers in the figure). The neurons of the last layers of the two pathways thus have receptive fields of size 17^3 voxels. The inputs of the two pathways are centered at the same image location, but the second segment is extracted from a down-sampled version of the image by a factor of 3. The second pathway processes context in an actual area of size 51^3 voxels. *DeepMedic*, our proposed 11-layers architecture, results by replacing each layer of the depicted pathways with two that use 3^3 kernels (see Sec. 2.3). Number of FMs and their size depicted as (*Number* $\times$ *Size*).

$\boldsymbol{\delta}_{L2}^{\{x,y,z\}} - 1$ and similar is the relation between $\boldsymbol{\delta}_{in1}$ and $\boldsymbol{\delta}_{L1}$. These establish the relation between the required dimensions of the input segments from the two resolutions, which can then be extracted centered on the same image location. The FMs of $L2$ are up-sampled to match the dimensions of $L1$'s FMs and are then concatenated together. We add two more hidden layers for combining the multi-scale features before the final classification, as shown in Fig. 5. Integration of the multi-scale parallel pathways in architectures with non-unary strides is discussed in Appendix A.

Combining multi-scale features has been found beneficial in other recent works (Long et al. (2015); Ronneberger et al. (2015)), in which whole 2D images are processed in the network by applying a few number of convolutions and then down-sampling the FMs for further processing at various scales. Our decoupled pathways allow arbitrarily large context to be provided while avoiding the need to load large parts of the 3D volume into memory. Additionally, our architecture extracts features completely independently from the multiple resolutions. This way, the features learned by the first pathway retain finest details, as they are not involved in processing low resolution

context.

2.5. 3D Fully Connected CRF for Structured Prediction

Because neighboring voxels share substantial spatial context, the soft segmentation maps produced by the CNN tend to be smooth, even though neighborhood dependencies are not modeled directly. However, local minima in training and noise in the input images can still result in some spurious outputs, with small isolated regions or holes in the predictions. We employ a fully connected CRF (Krähenbühl and Koltun (2011)) as a post-processing step to achieve more structured predictions. As we describe below, this CRF is capable of modeling arbitrarily large voxel-neighborhoods but is also computationally efficient, making it ideal for processing 3D multi-modal medical scans.

For an input image $\mathbf{I}$ and the label configuration (segmentation) $\mathbf{z}$, the Gibbs energy in a CRF model is given by

$$E(\mathbf{z}) = \sum_i \psi_u(z_i) + \sum_{ij, i \neq j} \psi_p(z_i, z_j) \ . \tag{5}$$

The unary potential is the negative log-likelihood $\psi_u(z_i) = -logP(z_i|\mathbf{I})$, where in our case $P(z_i|\mathbf{I})$ is the CNN's output for voxel i. In a fully connected CRF, the pairwise potential is of form $\psi_p(z_i, z_j) = \mu(z_i, z_j)k(\mathbf{f_i}, \mathbf{f_j})$ between any pair of voxels, regardless of their spatial distance. The Pott's Model is commonly used as the label compatibility function, giving $\mu(z_i, z_j) = [z_i \neq z_j]$. The corresponding energy penalty is given by the function k, which is defined over an arbitrary feature space, with $\mathbf{f_i}, \mathbf{f_j}$ being the feature vectors of the pair of voxels. Krähenbühl and Koltun (2011) observed that if the penalty function is defined as a linear combination of Gaussian kernels, $k(\mathbf{f_i}, \mathbf{f_j}) = \sum_{m=1}^{M} w^{(m)} k^{(m)}(\mathbf{f_i}, \mathbf{f_j})$, the model lends itself for very efficient inference with mean field approximation, after expressing message passing as convolutions with the Gaussian kernels in the space of the feature vectors $\mathbf{f_i}, \mathbf{f_j}$.

We extended the work of the original authors and implemented a 3D version of the CRF for processing multi-modal scans. We make use of two Gaussian kernels, which operate in the feature space defined by the voxel coordinates $p_{i,d}$ and the intensities of the c-th modality-channel $I_{i,c}$ for voxel i. The smoothness kernel, $k^{(1)}(\mathbf{f_i}, \mathbf{f_j}) = exp\left(-\sum_{d=\{x,y,z\}} \frac{|p_{i,d}-p_{j,d}|^2}{2\sigma_{\alpha,d}^2}\right)$, is defined by a diagonal covariance matrix with elements the configurable parameters $\sigma_{\alpha,d}$, one for each axis. These parameters express the size and shape of

16

neighborhoods that homogeneous labels are encouraged. The appearance kernel $k^{(2)}(\mathbf{f_i}, \mathbf{f_j}) = exp\Big(- \sum_{d=\{x,y,z\}} \frac{|p_{i,d}-p_{j,d}|^2}{2\sigma_{\beta,d}^2} - \sum_{c=1}^{C} \frac{|I_{i,c}-I_{j,c}|^2}{2\sigma_{\gamma,c}^2} \Big)$ is defined similarly. The additional parameters $\sigma_{\gamma,c}$ can be interpreted as how strongly to enforce homogeneous appearance in the C input channels, when voxels in an area spatially defined by $\sigma_{\beta,d}$ are identically labelled. Finally, the configurable weights $w^{(1)}, w^{(2)}$ define the relative strength of the two factors.

3. Analysis of Network Architecture

In this section we present a series of experiments in order to analyze the impact of each of the main contributions and to justify the choices made in the design of the proposed 11-layers, multi-scale 3D CNN architecture, referred to as the *DeepMedic*. Starting from the CNN baseline as discussed in Sec. 2.1, we first explore the benefit of our proposed dense training scheme (cf. Sec. 2.2), then investigate the use of deeper models (cf. Sec. 2.3) and then evaluate the influence of the multi-scale dual pathway (cf. Sec. 2.4). Finally, we compare our method with corresponding 2D variants to assess the benefit of processing 3D context.

3.1. Experimental Setting

The following experiments are conducted using the TBI dataset with 61 multi-channel MRIs which is described in more detail later in Sec. 4.1. Here, the images are randomly split into a validation and training set, with 15 and 46 images each. The same sets are used in all analyses. To monitor the progress of segmentation accuracy during training, we extract 10k random patches at regular intervals, with equal numbers extracted from each of the validation images. The patches are uniformly sampled from the brain region in order to approximate the true distribution of lesions and healthy tissue. Full segmentation of the validation datasets is performed every five epochs and the mean Dice similarity coefficient (DSC) is determined. Details on the configuration of the networks are provided in Appendix B.

3.2. Effect of Dense Training on Image Segments

We compare our proposed dense training method with two other commonly used training schemes on the 5-layers baseline CNN (see Fig. 2). The first common scheme trains on 17^3 patches extracted uniformly from the brain region, and the second scheme samples patches equally from the lesion
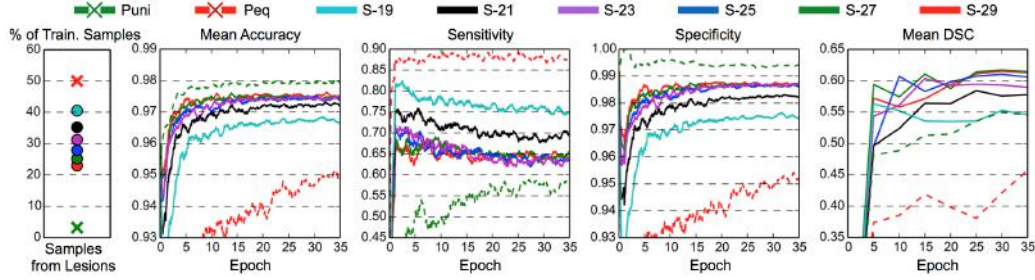
17

Figure 6: Comparison of the commonly used methods for training on patches uniformly sampled from the brain region (P_{uni}) and equally sampled from lesion and background (P_{eq}) against our proposed scheme (S-d) on cubic segments of side length d, also equally sampled from lesion and background. We varied d to observe its effect. From left to right: percentage of training samples extracted from the lesion class, mean accuracy, sensitivity, specificity calculated on uniformly sampled validation patches and, finally, the mean DSC of the segmentation of the validation datasets. The progress throughout training is plotted. Because lesions are small, P_{uni} achieves very high voxel-wise accuracy by being very specific but not sensitive, with the opposite being the case for P_{eq}. Our method achieves an effective balance between the two, resulting in better segmentation as reflected by higher DSC.

and background class. We refer to these schemes as P_{uni} and P_{eq}. The results shown in Fig. 6 show a correlation of sensitivity and specificity with the percentage of training samples that come from the lesion class. P_{eq} performs poorly because of over-segmentation (high sensitivity, low specificity). P_{uni} has better classification on the background class (high specificity), which leads to high mean voxel-wise accuracy since the majority corresponds to background, but not particularly high DSC scores due to under-segmentation (low sensitivity).

To evaluate our dense training scheme, we train multiple models with varying sized image segments, equally sampled from lesions and background. The tested sizes of the segments go from 19^3 upwards to 29^3. The models are referred to as "S-d", where d is the side length of the cubic segments. For fair comparison, the batch sizes in all the experiments are adjusted to have a similar memory footprint and lead to similar training times as compared

18

to training on P_{uni} and P_{eq}[2]. We observe a great performance increase for model S-19 over P_{eq}. We account this partly to the efficient increase of the effective batch size ($B \cdot V$ in Eq. (4)), but also to the altered distribution of training samples. As we increase the size of the training segments further, we quickly reach a balance between the sensitivity of P_{eq} and the specificity of P_{uni}, which results in improved segmentation as expressed by the DSC.

The segment size is a hyper-parameter in our model. We observe that the increase in performance with increasing segment size quickly levels off, and similar performance is obtained for a wide range of segment sizes, which allows for easy configuration. For the remaining experiments, all models were trained on segments of size 25^3.
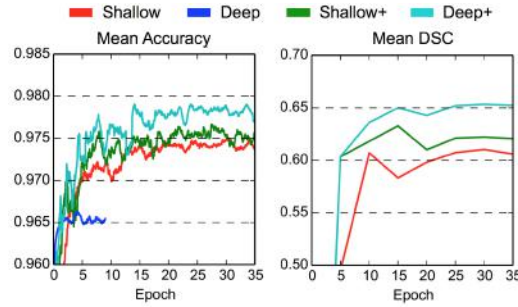
3.3. Effect of Deeper Networks

Figure 7: Mean accuracy over validation samples and DSC for the segmentations of the validation images, as obtained from the "Shallow" baseline and "Deep" variant with smaller kernels. Training of the plain deeper model fails (cf. Sec. 3.3). This is overcome by adopting the initialization scheme of (He et al. (2015)), which further combined with Batch Normalization leads to the enhanced (+) variants. Deep+ performs significantly better than Shallow+ with similar computation time, thanks to the use of small kernels.

The 5-layers baseline CNN (Fig. 2), here referred to as the "Shallow" model, is extended to 9-layers by replacing each convolutional layer that uses 5^3 kernels with two layers that use 3^3 kernels (Fig. 4). This model is

---

[2]Dense training on a whole volume was inapplicable in these experimental settings due to memory limitations but was previously shown to give similar results as training on uniformly sampled patches (Long et al. (2015)).

referred to as "Deep". Training the latter, however, utterly fails with the model making only predictions corresponding to the background class. This problem is related to the challenge of preserving the signal as it propagates through deep networks and its variance gets multiplied with the variance of the weights, as previously discussed in Sec. 2.3. One of the causes is that the weights of both models have been initialized with the commonly used scheme of sampling from the normal distribution $\mathcal{N}(0, 0.01)$ (cf. Krizhevsky et al. (2012)). In comparison, the initialization scheme by He et al. (2015), derived for preserving the signal in the initial stage of training, results in higher values and overcomes this problem. Further preservation of the signal is obtained by employing Batch Normalization. This results in an enhanced 9-layers model which we refer to as "Deep+", and using the same enhancements on the Shallow model yields "Shallow+". The significant performance improvement of Deep+ over Shallow+, as shown in Fig. 7, is the result of the greater representational power of the deeper network. The two models need similar computational times, which highlights the benefits of utilizing small kernels in the design of 3D CNNs. Although the deeper model requires more sequential (layer by layer) computations on the GPU, those are faster due to the smaller kernel size.

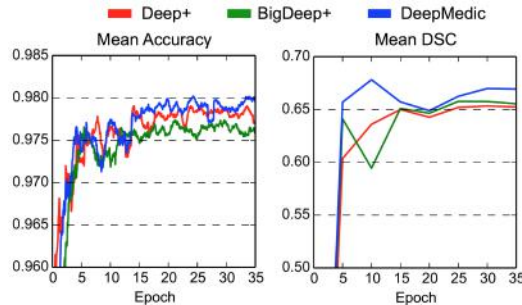*3.4. Effect of the Multi-Scale Dual Pathway*

Figure 8: Mean accuracy over validation samples and DSC for the segmentation of the validation images, as obtained by a single-scale model (Deep+) and our dual pathway architecture (DeepMedic). We also trained a single-scale model with larger capacity (BigDeep+), similar to the capacity of DeepMedic. DeepMedic yields best performance by capturing greater context, while BigDeep+ seems to suffer from over-fitting.

20

The final version of the proposed network architecture, referred to as "DeepMedic", is built by extending the Deep+ model with a second convolutional pathway that is identical to the first one. Two hidden layers are added for combining the multi-scale features before the classification layer, resulting in a deep network of 11-layers (cf. Fig. 5). The input segments to the second pathway are extracted from the images down-sampled by a factor of three. Thus, the network is capable of capturing context in a 51^3 area of the original image through the 17^3 receptive field of the lower-resolution pathway, while only doubling the computational and memory requirements over the single pathway CNN. In comparison, the most recent 2D CNN systems proposed for lesion segmentation (Havaei et al. (2015); Pereira et al. (2015)) have a receptive field limited to 33^2 voxels.

Figure 8 shows the improvement DeepMedic achieves over the single pathway model Deep+. In Fig. 9 we show two representative visual examples of this improvement when using the multi-scale CNN. Finally, we confirm that the performance increase can be accounted to the additional context and not the additional capacity of DeepMedic. To this end, we build a big single-scale model by doubling the FMs at each of the 9-layers of Deep+ and adding two hidden layers. This 11-layers deep and wide model, referred to as "BigDeep+", has the same number of parameters as DeepMedic. The performance of the model is not improved, while showing signs of over-fitting.

3.5. Processing 3D in comparison to 2D Context

Acquired brain MRI scans are often anisotropic. Such is the case for most sequences in our TBI dataset, which have been acquired with lower axial resolution, except for the isotropic MPRAGE. We perform a series of experiments to investigate the behaviour of 2D networks and assess the benefit of processing 3D context in this setting.

DeepMedic can be converted to 2D by setting the third dimension of each kernel to one. This way only information from the surrounding context on the axial plane influences the classification of each voxel. If 2D segments are given as input, the dimensionality of the feature maps decreases and so does the memory required. This allows developing 2D variants with increased width, depth and size of training batch with similar requirements as the 3D version, which are valid candidates for model selection in practical scenarios. We assess various configurations and present some representatives in Table B.1b along with their performance. Best segmentation among investigated 2D variants is achieved by a 19-layers, multi-scale network, reaching
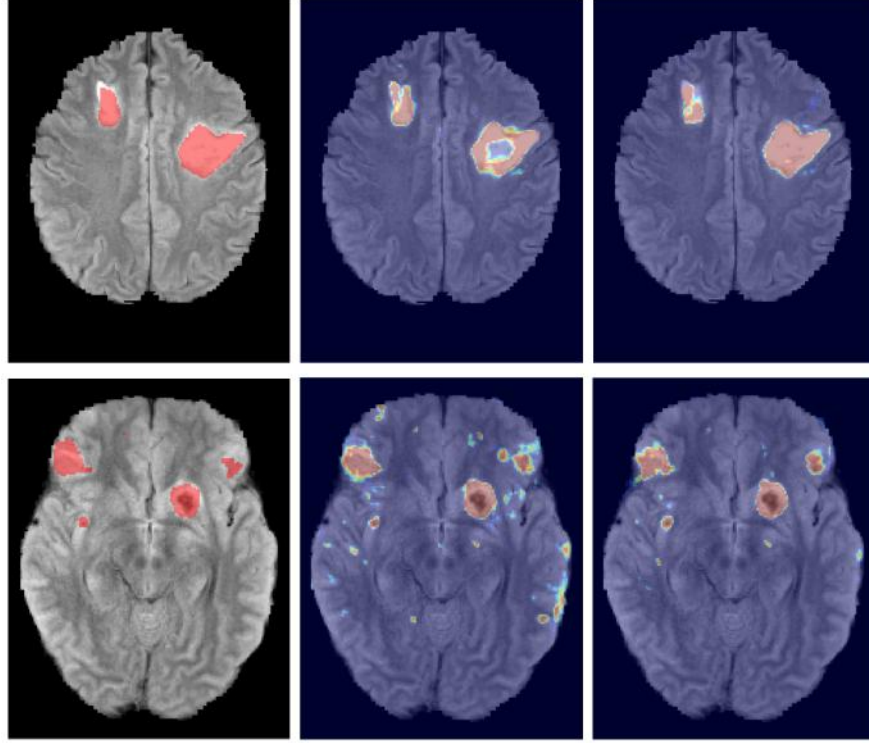
21

Figure 9: (Rows) Two cases from the severe TBI dataset, showing representative improvements when using the multi-scale CNN approach. (Columns) From left to right: the MRI FLAIR sequence with the manually labeled lesions, predicted soft segmentation map obtained from a single-scale model (Deep+) and the prediction of the multi-scale DeepMedic model. The incorporation of greater context enables DeepMedic to identify when it processes an area within larger lesions (top). Spurious false positives are significantly reduced across the image on the bottom.

61.5% average DSC on the validation fold. The decline from the 66.6% DSC achieved by the 3D version of DeepMedic indicates the importance of processing 3D context even in settings where most acquired sequences have low resolution along a certain axis.

4. Evaluation on Clinical Data

The proposed system consisting of the DeepMedic CNN architecture, optionally coupled with a fully connected CRF, is evaluated on three lesion

segmentation tasks including challenging clinical data from patients with traumatic brain injuries, brain tumors, and ischemic stroke. Quantitative evaluation and comparisons with state-of-the-art are reported for each of the tasks.

4.1. Traumatic Brain Injuries

4.1.1. Material and Pre-Processing

Sixty-six patients with moderate-to-severe TBI who required admission to the Neurosciences Critical Care Unit at Addenbrooke's Hospital, Cambridge, UK, underwent imaging using a 3-Tesla Siemens Magnetom TIM Trio within the first week of injury. Ethical approval was obtained from the Local Research Ethics Committee (LREC 97/290) and written assent via consultee agreement was obtained for all patients. The structural MRI sequences that are used in this work are isotropic MPRAGE ($1mm{\times}1mm{\times}1mm$), axial FLAIR, T2 and Proton Density (PD) ($0.7mm{\times}0.7mm{\times}5mm$), and Gradient-Echo (GE) ($0.86mm{\times}0.86mm{\times}5mm$). All visible lesions were manually annotated on the FLAIR and GE sequences with separate labeling for each lesion type. In nine patients the presence of hyperintense white matter lesions that were felt to be chronic in nature were also annotated. Artifacts, for example, signal loss secondary to intraparenchymal pressure probes, were also noted. For the purpose of this study we focus on binary segmentation of all abnormalities within the brain tissue. Thus, we merged all classes that correspond to intra-cerebral abnormalities into a single "lesion" label. Extra-cerebral pathologies such as epidural and subdural hematoma were treated as background. We excluded two datasets because of corrupted FLAIR images, two cases because no lesions were found and one case because of a major scanning artifact corrupting the images. This results in a total of 61 cases used for quantitative evaluation. Brain masks were obtained using the ROBEX tool (Iglesias et al. (2011)). All images were resampled to an isotropic $1mm^3$ resolution, with dimensions $193{\times}229{\times}193$ and affinely registered (Studholme et al. (1999)) to MNI space using the atlas by Grabner et al. (2006). No bias field correction was used as preliminary results showed that this can negatively affect lesion appearance. Image intensities were normalized to have zero-mean and unit variance, as it has been reported that this improves CNN results (Jarrett et al. (2009)).

4.1.2. Experimental Setting

Network configuration and training: The network architecture corresponds to the one described in Sec. 3.4, i.e. a dual-pathway, 11-layers deep CNN. The training data is augmented by adding images reflected along the sagittal axis. To make the network invariant to absolute intensities we also shift the intensities of each MR channel c of every training segment by $i_c = r_c \sigma_c$. r_c is sampled for every segment from $\mathcal{N}(0, 0.1)$ and σ_c is the standard deviation of intensities under the brain mask in the corresponding image. The network is regularized using dropout (Hinton et al. (2012)) with a rate of 2% on all convolutional layers, which is in addition to a 50% rate used on the last two layers. The network is evaluated with 5-fold cross-validation on the 61 subjects.

CRF configuration: The parameters of the fully connected CRF are determined in a configuration experiment using random-search and 15 randomly selected subjects from the TBI database with predictions from a preliminary version of the corresponding model. The 15 subjects are reshuffled into the 5-folds used for subsequent evaluation.

Random Forest baseline: We have done our best to set up a competitive baseline for comparison. We employ a context-sensitive Random Forest, similar to the model presented by Zikic et al. (2012) for brain tumors except that we apply the forest to the MR images without additional tissue specific priors. We train a forest with 50 trees and maximum depth of 30. Larger size did not improve results. Training data points are approximately equally sampled from lesion and background classes, with the optimal balance empirically chosen. Two hundred randomized cross-channel box features are evaluated at each split node with maximum offsets and box sizes of 20mm. The same folds of training and test sets are used as for our CNN approach.

4.1.3. Results

Table 1 summarizes the results on TBI. Our CNN significantly outperforms the Random Forest baseline, while the relatively overall low DSC values indicate the difficulty of the task. Due to randomness during training the local minima where a network converges are different between training sessions and some errors they produce differ (Choromanska et al. (2015)). To clear the unbiased errors of the network we form an *ensemble* of three similar networks, aggregating their output by averaging. This ensemble yields better performance in all metrics but also allows us to investigate the behaviour of our network focusing only on the biased errors. Fig. 10 shows the DSC obtained

Table 1: Performance of *DeepMedic* and an *ensemble* of three networks on the TBI database. For comparison, we provide results for a Random Forest baseline. Values correspond to the mean (and standard deviation). Numbers in bold indicate significant improvement by the CRF post-processing, according to a two-sided, paired t-test on the DSC metric ($*p < 5 \cdot 10^{-2}$, $**p < 10^{-4}$).

	DSC	Precision	Sensitivity	ASSD	Haussdorf
R. Forest	51.1(20.0)	50.1(24.4)	60.1(15.8)	8.29(6.76)	64.17(15.98)
R. Forest+CRF	**54.8(18.5)****	58.6(23.1)	56.9(17.4)	6.71(5.01)	59.45(15.52)
DeepMedic	62.3(16.4)	65.3(18.8)	64.4(16.3)	4.24(2.64)	56.50(15.88)
DeepMedic+CRF	**63.0(16.3)****	67.7(18.2)	63.2(16.7)	4.02(2.54)	55.68(15.93)
Ensemble	64.2(16.2)	67.7(18.3)	65.3(16.3)	3.88(2.33)	54.38(15.45)
Ensemble+CRF	**64.5(16.3)***	69.8(17.8)	63.9(16.7)	3.72(2.29)	52.38(16.03)

by the ensemble on each subject in relation to the manually segmented and predicted lesion volume. The network is capable of segmenting cases with very small lesions, although, performance is less robust in these cases as even small errors have large influence on the DSC metric. Investigation of the predicted lesion volume, which is an important biomarker for prognostication, shows that the network is neither biased towards the lesion nor background class, with promising results even on cases with very small lesions. Furthermore, we separately evaluate the influence of the post-processing with the fully connected CRF. As shown in Table 1, the CRF yields improvements over all classifiers. Effects are more prominent when the performance of the primary segmenter degrades, which shows the robustness of this regulariser. Fig. 11 shows three representative cases.

4.2. Brain Tumor Segmentation

4.2.1. Material and Pre-Processing

For brain tumors, we evaluate our system on the data from the 2015 Brain Tumor Segmentation Challenge (BRATS) (Menze et al. (2015)). The training set consists of 220 cases with high grade (HG) and 54 cases with low grade (LG) glioma for which corresponding reference segmentations are provided. The segmentations include the following tumor tissue classes: 1) necrotic core, 2) edema, 3) non-enhancing and 4) enhancing core. The test set consists of 110 cases of both HG and LG but the grade is not revealed. Reference segmentations for the test set are hidden and evaluation is carried out via an online system. For evaluation, the four predicted labels are merged
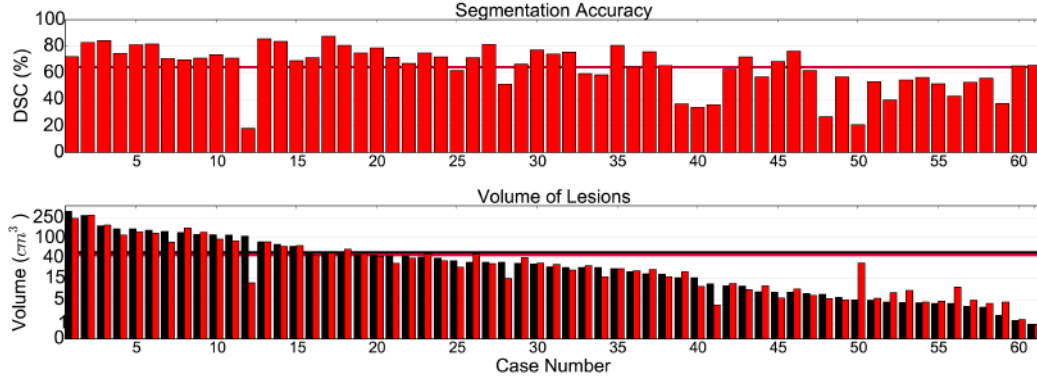
Figure 10: (Top) DSC achieved by our ensemble of three networks on each of the 61 TBI datasets. (Bottom) Manually segmented (black) and predicted lesion volumes (red). Note here the logarithmic scale. Continuous lines represent mean values. The outlying subject 12 presents small TBI lesions, which are successfully segmented, but also vascular ischemia. Because it is the only case in the database with the latter pathology, the networks fail to segment it as such lesion was not seen during training.

into different sets of whole tumor (all four classes), the core (classes 1,3,4), and the enhancing tumor (class 4)[3]. For each subject, four MRI sequences are available, FLAIR, T1, T1-contrast and T2. The datasets are pre-processed by the organizers and provided as skull-stripped, registered to a common space and resampled to isotropic $1mm^3$ resolution. Dimensions of each volume are $240\times240\times155$. We add minimal pre-processing of normalizing the brain-tissue intensities of each sequence to have zero-mean and unit variance.

4.2.2. Experimental Setting

Network configuration and training: We modify the DeepMedic architecture to handle multi-class problems by extending the classification layer to five feature maps (four tumor classes plus background). The rest of the configuration remains unchanged. We enrich the dataset with sagittal reflections. Opposite to the experiments on TBI, we do not employ the inten-

---

[3]For interpretation of the results note that, to the best of our knowledge, cases where the "enhancing tumor" class is not present in the manual segmentation are considered as zeros for the calculation of average performance by the evaluation platform, lowering the upper bound for this class.
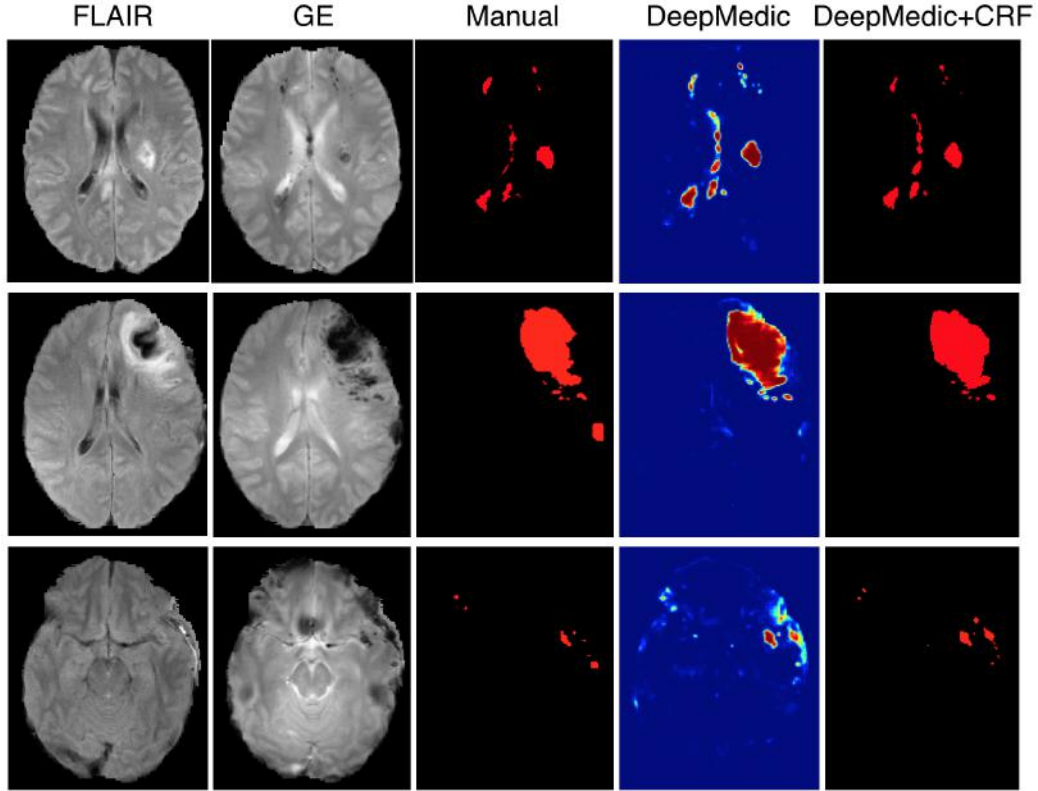
Figure 11: Three examples from the application of our system on the TBI database. It is capable of precise segmentation of both small and large lesions. Second row depicts one of the common mistakes observed. A contusion near the edge of the brain is under-segmented, possibly mistaken for background. Bottom row shows one of the worst cases, representative of the challenges in segmenting TBI. Post-surgical sub-dural debris is mistakenly captured by the brain mask. The network partly segments the abnormality, which is not a celebral lesion of interest.

sity perturbation and dropout on convolutional layers, because the network should not require as much regularisation with this large database. The network is trained on image segments extracted with equal probability centred on the whole tumor and healthy tissue. The distribution of the classes captured by our training scheme is provided in Appendix C.

To examine our network's behaviour, we first evaluate it on the training

data of the challenge. For this, we run a 5-fold cross validation where each fold contains both HG and LG images. We then retrain the network using all training images, before applying it on the test data.

CRF configuration: For the multi-class problem it is challenging to find a global set of parameters for the CRF which can consistently improve the segmentation of all classes. So instead we merge the four predicted probability maps into a single "whole tumor" map for CRF post-processing. The CRF then only refines the boundaries between tumor and background and additionally removes isolated false positives. Similarly to the experiments on TBI, the CRF is configured on a random subset of 44 HG and 18 LG training images, which are then reshuffled into the subsequent 5-fold cross validation.

4.2.3. Results

Table 2: Average performance of our system on the training data of BRATS 2015 as computed on the online evaluation platform and comparison to other submissions visible at the time of manuscript submission. Presenting only teams that submitted more than half of the 274 cases. Numbers in bold indicate significant improvement by the CRF, according to a two-sided, paired t-test on the DSC metric ($*p < 5 \cdot 10^{-2}$, $**p < 10^{-3}$).

	DSC			Precision			Sensitivity			
	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.	Cases
Ensemble+CRF	**90.1***	75.4	**72.8***	91.9	85.7	75.5	89.1	71.7	74.4	274
Ensemble	90.0	75.5	72.8	90.3	85.5	75.4	90.4	71.9	74.3	274
DeepMedic+CRF	**89.8****	75.0	**72.1***	91.5	84.4	75.9	89.1	72.1	72 .5	274
DeepMedic	89.7	75.0	72.0	89.7	84.2	75.6	90.5	72.3	72.5	274
bakas1	88	77	68	90	84	68	89	76	75	186
peres1	87	73	68	89	74	72	86	77	70	274
anon1	84	67	55	90	76	59	82	68	61	274
thirs1	80	66	58	84	71	53	79	66	74	267
peyrj	80	60	57	87	79	59	77	53	60	274

Quantitative results from the application of the DeepMedic, the CRF and an ensemble of three similar networks on the training data are presented in Table 2. The latter two offer an improvement, albeit fairly small since the performance of DeepMedic is already rather high in this task. Also shown are results from previous works, as reported on the online evaluation platform. Various settings may vary among submissions, such as the pre-processing pipeline or the number of folds used for cross-validation. Still it appears that our system performs favourably compared to previous state-of-the-art,

including the semi-automatic system of Bakas et al. (2015) (bakas1) who won the latest challenge and the method of Pereira et al. (2015) (peres1), which is based on grade-specific 2D CNNs and requires visual inspection of the tumor and identification of the grade by the user prior to segmentation. Examples of segmentations obtained with our method are shown in Fig. 12. DeepMedic behaves very well in preserving the hierarchical structure of the tumor, which we account to the large context processed by our multi-scale network.

Table 3 shows the results of our method on the BRATS test data. Results of other submissions are not accessible. The decrease in performance is possibly due to the the inclusion of test images that vary significantly from the training data, such as cases acquired in clinical centers that did not provide any of the training images, something that was confirmed by the organisers. Note that performance gains obtained with the CRF are larger in this case. This indicates not only that its configuration has not overfitted to the training database but also that the CRF is robust to factors of variation between acquisition sites, which complements nicely the more sensitive CNN.

Table 3: Average performance of our system on the 110 test cases of BRATS 2015, as computed on the online evaluation platform. Numbers in bold indicate significant improvement by the CRF, according to a two-sided, paired t-test on the DSC metric ($^*p < 5 \cdot 10^{-2}$, $^{**}p < 10^{-3}$). The decrease of the mean DSC by the CRF and the ensemble for the "Core" class was not found significant.

	DSC			Precision			Sensitivity		
	Whole	Core	Enh.	Whole	Core	Enh.	Whole	Core	Enh.
DeepMedic	83.6	67.4	62.9	82.3	84.6	64.0	88.5	61.6	65.6
DeepMedic+CRF	**84.7****	67.0	62.9	85.0	84.8	63.4	87.6	60.7	66.2
Ensemble	84.5	66.7	63.3	83.3	86.1	63.2	88.9	59.9	67.3
Ensemble+CRF	**84.9****	66.7	**63.4***	85.3	86.1	63.4	87.7	60.0	67.4

4.3. Ischemic Stroke Lesion Segmentation

4.3.1. Material and Pre-Processing

We participated in the 2015 Ischemic Stroke Lesion Segmentation (ISLES) challenge, where our system achieved the best results among all participants on sub-acute ischemic stroke lesions (Maier et al. (2017)). In the training phase of the challenge, 28 datasets have been made available, along with
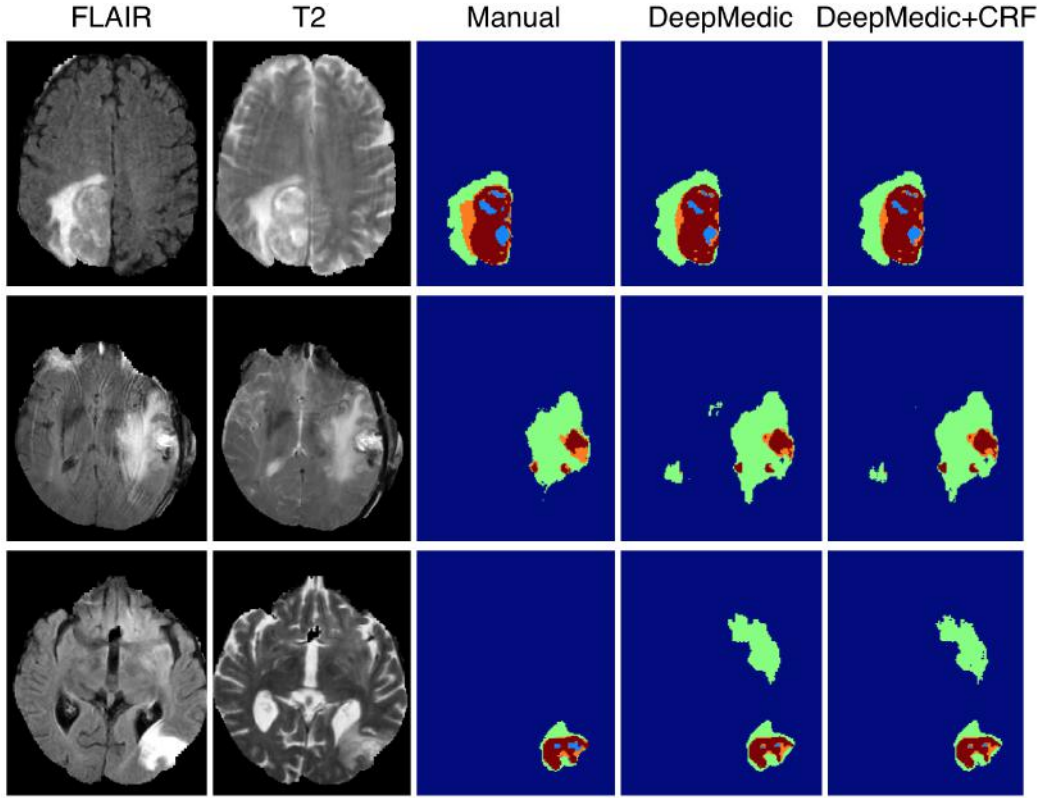
Figure 12: Examples of DeepMedic's segmentation from its evaluation on the training datasets of BRATS 2015. cyan: necrotic core, green: oedema, orange: non-enhancing core, red: enhancing core. (top and middle) Satisfying segmentation of the tumor, regardless motion artefacts in certain sequences. (bottom) One of the worst cases of over-segmentation observed. False segmentation of FLAIR hyper-intensities as oedema constitutes the most common error of DeepMedic.

manual segmentations. Each dataset included T1, T1-contrast, FLAIR and DWI sequences. All images were provided as skull-stripped and resampled to isotropic $1mm^3$ voxel resolution. Each volume is of size $230{\times}230{\times}154$. In the testing stage, teams were provided with 36 datasets for evaluation. The test data were acquired in two clinical centers, with one of them being the same that provided all training images. Corresponding expert segmentations were hidden and results had to be submitted to an online evaluation

platform. Similar to BRATS, the only pre-processing that we applied is the normalization of each image to the zero-mean and unit variance.

4.3.2. Experimental Setting

Network Configuration and Training: The configuration of the network employed is described in Kamnitsas et al. (2015). The main difference with the configuration used for TBI and tumors as employed above is the relatively smaller number of FMs in the low-resolution pathway. This choice should not significantly influence accuracy on the generally small SISS lesions but it allowed us to lower the computational cost.

Similar to the other experiments, we evaluate our network with a 5-fold cross validation on the training datasets. We use data augmentation with sagittal reflections. For the testing phase of the challenge, we trained an ensemble of three networks on all training cases and aggregate their predictions by averaging.

CRF configuration: The parameters of the CRF were configured via a random search on the whole training dataset.

4.3.3. Results

The performance of our system on the training data is shown in Table 4. Significant improvement is achieved by the structural regularisation offered by the CRF, although it could be partially accounted for by overfitting the training data during the CRF's configuration. Examples for visual inspection are shown in Fig. 13.

Table 4: Performance of our system on the training data of the ISLES-SISS 2015 competition. Values correspond to the mean (and standard deviation). Numbers in bold indicate significant improvement by the CRF, according to a two-sided, paired t-test on the DSC metric ($p < 10^{-2}$).

	DSC	Precision	Sensitivity	ASSD	Haussdorf
DeepMedic	64(23)	68(24)	65(23)	6.99(9.91)	73.32(26.03)
DeepMedic+CRF	**66(24)**	77(24)	63(25)	5.00(10.33)	55.93(28.55)

For the testing phase of the challenge we formed an ensemble of three networks, coupled with the fully connected CRF. Our submission ranked first, indicating superior performance on this challenging task among 14 submissions. Table 5 shows our results, along with the other two top entries (Feng

Table 5: Our ensemble of three networks, coupled with the fully connected CRF obtained overall best performance among all participants in the testing stage of the ISLES-SISS 2015 challenge. Shown is the performance of our pipeline along with the second and third entries. Values correspond to the mean (and standard deviation).

	DSC	Precision	Sensitivity	ASSD	Haussdorf
kamnk1(ours)	59(31)	68(33)	60(27)	7.87(12.63)	39.61(30.68)
fengc1	55(30)	64(31)	57(33)	8.13(15.15)	25.02(22.02)
halmh1	47(32)	47(34)	56(33)	14.61(20.17)	46.26(34.81)

et al. (2015); Halme et al. (2015)). Among the other participating methods was the CNN of Havaei et al. (2015) with 3 layers of 2D convolutions. That method perfomed less well on this challenging task (Maier et al. (2017)). This points out the advantage offered by 3D context, the large field of view of DeepMedic thanks to multi-scale processing and the representational power of deeper networks. It is important to note the decrease of performance in comparison to the training set. All methods performed worse on the data coming from the second clinical center, including the method of Feng et al. (2015) that is not machine-learning based. This highlights a general difficulty with current approaches when applied on multi-center data.

4.4. Implementation Details

Our CNN is implemented using the Theano library (Bastien et al. (2012)). Each training session requires approximately one day on an NVIDIA GTX Titan X GPU using cuDNN v5.0. The efficient architecture of DeepMedic also allows models to be trained on GPUs with only 3GB of memory. Note that although dimensions of the volumes in the processed databases do not allow dense training on whole volumes for this size of network, dense inference on a whole volume is still possible, as it requires only a forward-pass and thus less memory. In this fashion segmentation of a volume takes less than 30 seconds but requires 12 GB of GPU memory. Tiling the volume into multiple segments of size 35^3 allows inference on 3 GB GPUs in less than three minutes.

Our 3D fully connected CRF is implemented by extending the original source code by Krähenbühl and Koltun (2011). A CPU implementation is fast, capable of processing a five-channel brain scan in under three minutes. Further speed-up could be achieved with a GPU implementation, but was
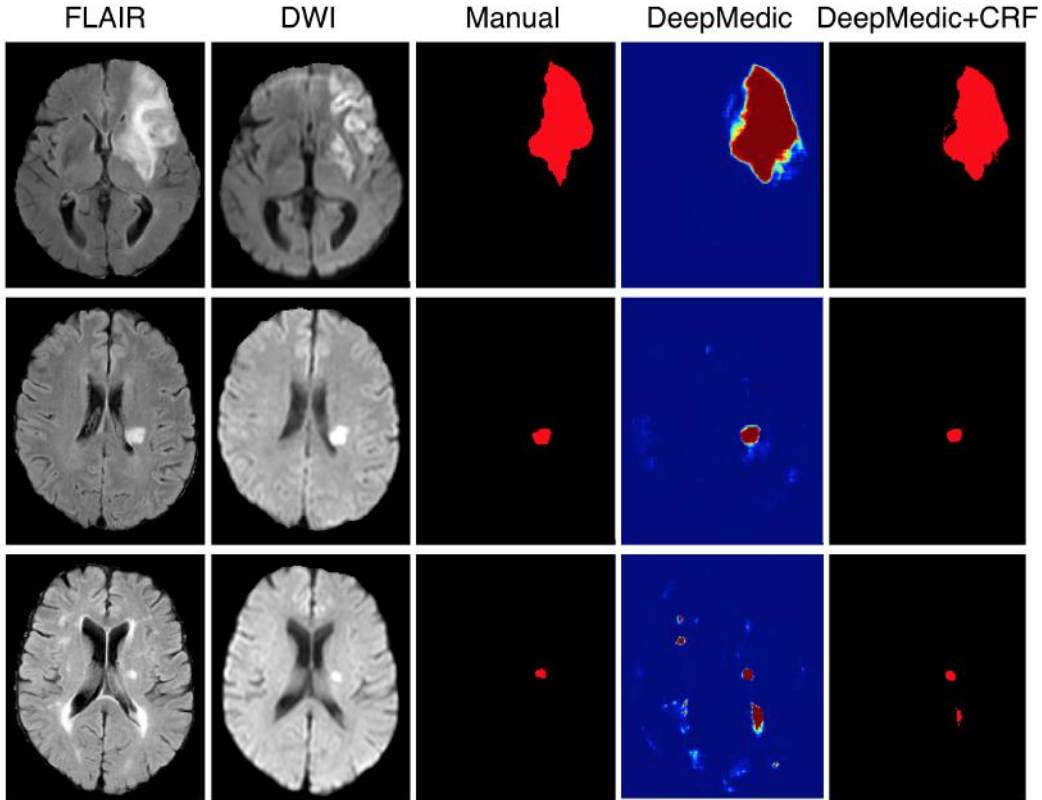
Figure 13: Examples of segmentations performed by our system on the training datasets of (SISS) ISLES 2015. (top and middle) The system is capable of satisfying segmentation of both large and smaller lesions. (bottom) Common mistakes are performed due to the challenge of differentiating stroke lesions from White Matter lesions.

not found necessary in the scope of this work.

5. Discussion and Conclusion

We have presented DeepMedic, a 3D CNN architecture for automatic lesion segmentation that surpasses state-of-the-art on challenging data. The proposed novel training scheme is not only computationally efficient but also offers an adaptive way of partially alleviating the inherent class-imbalance of segmentation problems. We analyzed the benefits of using small convolu-

tional kernels in 3D CNNs, which allowed us to develop a deeper and thus more discriminative network, without increasing the computational cost and number of trainable parameters. We discussed the challenges of training deep neural networks and the adopted solutions from the latest advances in deep learning. Furthermore, we proposed an efficient solution for processing large image context by the use of parallel convolutional pathways for multi-scale processing, alleviating one of the main computational limitations of previous 3D CNNs. Finally, we presented the first application of a 3D fully connected CRF on medical data, employed as a post-processing step to refine the network's output, a method that has also been shown promising for processing 2D natural images (Chen et al. (2014)). The design of the proposed system is well suited for processing medical volumes thanks to its generic 3D nature. The capabilities of DeepMedic and the employed CRF for capturing 3D patterns exceed those of 2D networks and locally connected random fields, models that have been commonly used in previous work. At the same time, our system is very efficient at inference time, which allows its adoption in a variety of research and clinical settings.

The generic nature of our system allows its straightforward application for different lesion segmentation tasks without major adaptations. To the best of our knowledge, our system achieved the highest reported accuracy on a cohort of patients with severe TBI. As a comparison, we improved over the reported performance of the pipeline in Rao et al. (2014). Important to note is that the latter work focused only on segmentation of contusions, while our system has been shown capable of segmenting even small and diffused pathologies. Additionally, our pipeline achieved state-of-the-art performance on both public benchmarks of brain tumors (BRATS 2015) and stroke lesions (SISS ISLES 2015). We believe performance can be further improved with task- and data-specific adjustments, for instance in the pre-processing, but our results show the potential of this generically designed segmentation system.

When applying our pipeline to new tasks, a laborious process is the reconfiguration of the CRF. The model improved our system's performance with statistical significance in all investigated tasks, most profoundly when the performance of the underlying classifier degrades, proving its flexibility and robustness. Finding optimal parameters for each task, however, can be challenging. This became most obvious on the task of multi-class tumor segmentation. Because the tumor's substructures vary significantly in appearance, finding a global set of parameters that yields improvements on all

classes proved difficult. Instead, we applied the CRF in a binary fashion. This CRF model can be configured with a separate set of parameters for each class. However the larger parameter space would complicate its configuration further. Recent work from Zheng et al. (2015) showed that this particular CRF can be casted as a neural network and its parameters can be learned with regular gradient descent. Training it in an end-to-end fashion on top of a neural network would alleviate the discussed problems. This will be explored as part of future work.

The discriminative power of the learned features is indicated by the success of recent CNN-based systems in matching human performance in domains where it was previously considered too ambitious (He et al. (2015); Silver et al. (2016)). Analysis of the automatically extracted information could potentially provide novel insights and facilitate research on pathologies for which little prior knowledge is currently available. In an attempt to illustrate this, we explore what patterns have been learned automatically for the lesion segmentation tasks. We visualize the activations of DeepMedic's FMs when processing a subject from our TBI database. Many appearing patterns are difficult to interpret, especially in deeper layers. In Fig. 14 we provide some examples that have an intuitive explanation. One of the most interesting findings is that the network learns to identify the ventricles, CSF, white and gray matter. This reveals that differentiation of tissue type is beneficial for lesion segmentation. This is in line with findings in the literature, where segmentation performance of traditional classifiers was significantly improved by incorporation of tissue priors (Van Leemput et al. (1999); Zikic et al. (2012)). It is intuitive that different types of lesions affect different parts of the brain depending on the underlying mechanisms of the pathology. A rigorous analysis of spatial cues extracted by the network may reveal correlations that are not well defined yet.

Similarly intriguing is the information extracted in the low-resolution pathway. As they process greater context, these neurons gain additional localization capabilities. The activations of certain FMs form fields in the surrounding areas of the brain. These patterns are preserved in the deepest hidden layers, which indicates they are beneficial for the final segmentation (see two last rows of Fig. 14). We believe these cues provide a spatial bias to the system, for instance that large TBI contusions tend to occur towards the front and sides of the brain (see Fig. 1c). Furthermore, the interaction of the multi-resolution features can be observed in FMs of the hidden layer that follows the concatenation of the pathways. The network learns to weight the
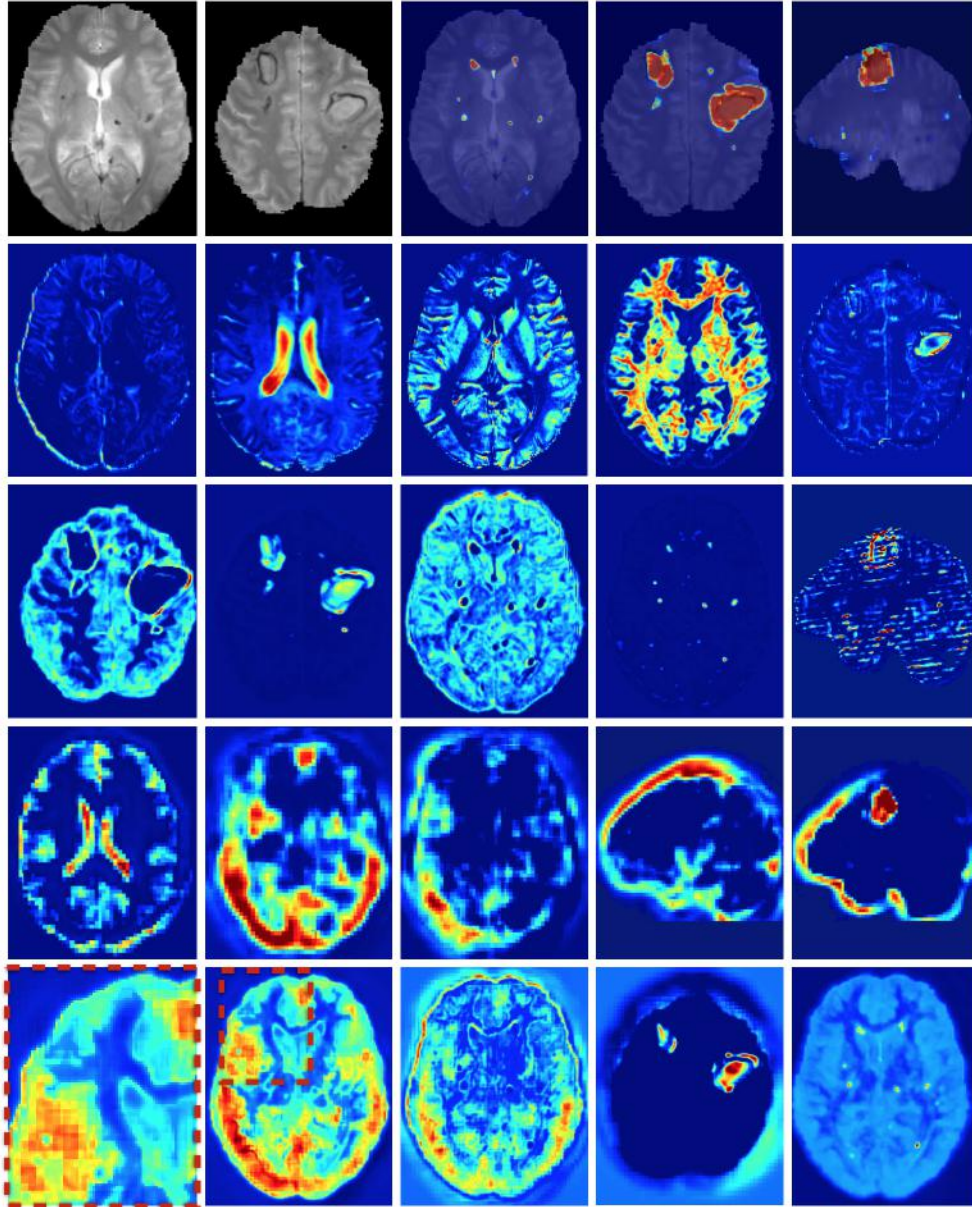
Figure 14: (First row) GE scan and DeepMedic's segmentation. (Second row) FMs of earlier and (third row) deeper layers of the first convolutional pathway. (Fourth row) Features learnt in the low-resolution pathway. (Last row) FMs of the two last hidden layers, which combine multi-resolution features towards the final segmentation.

output of the two pathways, preserving low resolution in certain parts and show fine details in others (bottom row of Fig. 14, first three FMs). Our assumption is that the low-resolution pathway provides a rough localization of large pathologies and brain areas that are challenging to segment, which reserves the rest of the network's capacity for learning detailed patterns associated with the detection of smaller lesions, fine structures and ambiguous areas.

The findings of the above exploration lead us to believe that great potential lies into fusing the discriminative power of the "deep black box" with the knowledge acquired over years of targeted biomedical research. Clinical knowledge is available for certain pathologies, such as spatial priors for white matter lesions. Previously engineered models have been proven effective in tackling fundamental imaging problems, such as brain extraction, tissue segmentation and bias field correction. We show that a network is capable of automatically extracting some of this information. It would be interesting, however, to investigate structured ways for incorporating such existing information as priors into the network's feature space, which should simplify the optimization problem while letting a specialist guide the network towards an optimal solution.

Although neural networks seem promising for medical image analysis, making the inference process more interpretable is required. This would allow understanding when the network fails, an important aspect in biomedical applications. Although the output is bounded in the $[0, 1]$ range and commonly referred to as probability for convenience, it is not a true probability in a Bayesian sense. Research towards Bayesian networks aims to alleviate this limitation. An example is the recent work of Gal and Ghahramani (2015) who show that model confidence can be estimated via sampling the dropout mask.

A general point should be made about the performance drop observed when our system is applied on test datasets of BRATS and ISLES in comparison to its cross-validated performance on the training data. In both cases, subsets of the test images were acquired in clinical centers different from the ones of training datasets. Differences in scanner type and acquisition protocols have significant impact on the appearance of the images. The issue of multi-center data heterogeneity is considered a major bottleneck for enabling large-scale imaging studies. This is not specific to our approach, but a general problem in medical image analysis. One possible way of making the CNN invariant to the data heterogeneity is to learn a generative model for

37

the data acquisition process, and use this model in the data augmentation step. This is a direction we explore as part of future work.

In order to facilitate further research in this area and to provide a baseline for future evaluations, we make the source code of the entire system publicly available.

Acknowledgements

Appendix A. Additional Details on Multi-Scale Processing

The integration of multi-scale parallel pathways in architectures that use solely unary kernel strides, such as the proposed, was described in Sec. 2.4. The required up-sampling of the low-resolution features was performed with simple repetition in our experiments. This was found sufficient, with the following hidden layers learning to combine the multi-scale features. In the case of architectures with strides greater than unary, the last convolutional layers of the two pathways, $L1$ and $L2$, have receptive fields $\boldsymbol{\varphi}_{L1}$ and $\boldsymbol{\varphi}_{L2}$ with strides $\boldsymbol{\tau}_{L1}$ and $\boldsymbol{\tau}_{L2}$ respectively. To preserve spatial correspondence of the multi-scale features and enable the network for dense inference, the dimensions of the input segments should be chosen such that the FMs in $L2$ can be brought to the dimensions of the FMs in $L1$ after sequential resampling by $\uparrow \boldsymbol{\tau}_{L2}, \uparrow F_D, \downarrow \boldsymbol{\tau}_{L1}$ or equivalent combinations. Here $\uparrow$ and $\downarrow$ represent up- and down-sampling by the given factor. Because they are more reliant on

38

these operations, utilization of more elaborate, learnt upsampling schemes (Long et al. (2015); Ronneberger et al. (2015); Noh et al. (2015)) should be beneficial in such networks.

Appendix B. Additional Details on Network Configurations

3D Networks: The main description of our system is presented in Sec. 2. All models discussed in this work outside Sec. 3.5 are fully 3D CNNs. Their architectures are presented in Table B.1a. They all use the PReLu non-linearity (He et al. (2015)). They are trained using the RMSProp optimizer (Tieleman and Hinton (2012)) and Nesterov momentum (Sutskever et al. (2013)) with value $m = 0.6$. $L1 = 10^{-6}$ and $L2 = 10^{-4}$ regularisation is applied. We train the networks with dense-training on batches of 10 segments, each of size 25^3. Exceptions are the experiments in Sec 3.2, where the batch sizes were adjusted along with the segment sizes, to achieve similar memory footprint and training time per batch. The weights of our shallow, 5-layers networks are initialized by sampling from a normal distribution $\mathcal{N}(0, 0.01)$ and their initial learning rate is set to $a = 10^{-4}$. Deeper models (and the "Shallow+" model in Sec 3.3) use the weight initialisation scheme of He et al. (2015). The scheme increases the signal's variance in our settings, which leads to RMSProm decreasing the effective learning rate. To counter this, we accompany it with an increased initial learning rate $a = 10^{-3}$. Throughout training, the learning rate of all models is halved whenever convergence plateaus. Dropout with 50% rate is employed on the two last hidden layers of 11-layers deep models.

2D Networks: Table B.1b presents representative examples of 2D configurations that were employed for the experiments discussed in Sec. 3.5. Width, depth and batch size were adjusted so that total required memory was similar to the 3D version of DeepMedic. Wider or deeper variants than the ones presented did not show greater performance. A possible reason is that this number of filters is enough for the extraction of the limited 2D information and that the field of view of the deep multi-scale variant is already sufficient for the application. The presented 2D models were regularized with $L1 = 10^{-8}$ and $L2 = 10^{-6}$ since they have less parameters than the 3D variants. All but Dm2dPatch were trained with momentum $m = 0.6$ and initial learning rate $a = 10^{-3}$, while the rest with $m = 0.9$ and $a = 10^{-2}$ as this setting increased performance. The rest of the hyper parameters are the same as for the 3D DeepMedic.

Table B.1: Network architectures investigated in Sec. 3 and final validation accuracy achieved in the corresponding experiments. (a) 3D and (b) 2D architectures. Columns from left to right: model's name, number of parallel identical pathways and number of feature maps at each of their convolutional layers, number of feature maps at each hidden layer that follows the concatenation of the pathways, dimensions of input segment to the normal and low resolution pathways, batch size and, finally, average DSC achieved on the validation fold. Further configuration details provided in Appendix B.

(a) 3D Network Architectures

	#Pathways: FMs/Layer	FMs/Hidd.	Seg.Norm.	Seg.Low	B.S.	DSC(%)
Shallow(+)	1: 30,40,40,50	-	25x25x25	-	10	60.2(61.7)
Deep(+)	1: 30,30,40,40,40,40,50,50	-	25x25x25	-	10	00.0(64.9)
BigDeep+	1: 60,60,80,80,80,80,100,100	150,150	25x25x25	-	10	65.2
DeepMedic	2: 30,30,40,40,40,40,50,50	150,150	25x25x25	19x19x19	10	66.6

(b) 2D Network Architectures

	#Pathways: FMs/Layer	FMs/Hidd.	Seg.Norm.	Seg.Low	B.S.	DSC(%)
Dm2dPatch*	2: 30,30,40,40,40,40,50,50	150,150	17x17x1	17x17x1	540	58.8
Dm2dSeg	2: 30,30,40,40,40,40,50,50	150,150	25x25x1	19x19x1	250	60.9
Wider2dSeg	2: 60,60,80,80,80,80,100,100	200,200	25x25x1	19x19x1	100	61.3
Deeper2dSeg	2: 16 layers, linearly 30 to 50	150,150	41x41x1	35x35x1	100	61.5
Large2dSeg	2: 12 layers, linearly 45 to 80	200,200	33x33x1	27x27x1	100	61.3

* Sampling was manually calibrated to achieve similar class balance as models that are trained on image segments. Model underperformed otherwise.

Appendix C. Distribution of Tumor Classes Captured in Training

Table C.1

Table C.1: Real distribution of the classes in the training data of BRATS 2015, along with the distribution captured by our proposed training scheme, when segments of size 25^3 are extracted centred on the tumor and healthy tissue with equal probability. Relative distribution of the foreground classes is closely preserved and the imbalance in comparison to the healthy tissue is automatically alleviated.

	Healthy	Necrosis	Edema	Non-Enh.	Enh.Core
Real	92.42	0.43	4.87	1.02	1.27
Captured	58.65	2.48	24.98	6.40	7.48

References

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., Rozycki, M., Pati, S., Davatzikos, C., 2015. GLISTRboost: Combining multimodal MRI segmentation, registration, and biophysical tumor growth modeling with gradient boosting machines for glioma segmentation. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, pp. 144–155.

Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I. J., Bergeron, A., Bouchard, N., Bengio, Y., 2012. Theano: new features and speed improvements. Deep Learning and Unsupervised Feature Learning NIPS 2012 Workshop.

Brebisson, A., Montana, G., 2015. Deep neural networks for anatomical brain segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 20–28.

Brosch, T., Yoo, Y., Tang, L. Y., Li, D. K., Traboulsee, A., Tam, R., 2015. Deep convolutional encoder networks for multiple sclerosis lesion segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 3–11.

Cardoso, M. J., Sudre, C. H., Modat, M., Ourselin, S., 2015. Template-based multimodal joint generative model of brain data. In: Information Processing in Medical Imaging. Springer, pp. 17–29.

Carey, L. M., Seitz, R. J., Parsons, M., Levi, C., Farquharson, S., Tournier, J.-D., Palmer, S., Connelly, A., 2013. Beyond the lesion: neuroimaging foundations for post-stroke recovery. Future Neurology 8 (5), 507–527.

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A. L., 2014. Semantic image segmentation with deep convolutional nets and fully connected CRFs. arXiv preprint arXiv:1412.7062.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., LeCun, Y., 2015. The loss surfaces of multilayer networks. In: AISTATS.

Ciresan, D., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2012. Deep neural networks segment neuronal membranes in electron microscopy images. In: Advances in neural information processing systems. pp. 2843–2851.

Cireşan, D. C., Giusti, A., Gambardella, L. M., Schmidhuber, J., 2013. Mitosis detection in breast cancer histology images with deep neural networks. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 411–418.

Ding, K., de la Plata, C. M., Wang, J. Y., Mumphrey, M., Moore, C., Harper, C., Madden, C. J., McColl, R., Whittemore, A., Devous, M. D., et al., 2008. Cerebral atrophy after traumatic white matter injury: correlation with acute neuroimaging and outcome. Journal of Neurotrauma 25 (12), 1433–1440.

Doyle, S., Vasseur, F., Dojat, M., Forbes, F., 2013. Fully automatic brain tumor segmentation from multiple MR sequences using hidden Markov fields and variational EM. Procs. NCI-MICCAI BRATS, 18–22.

Erihov, M., Alpert, S., Kisilev, P., Hashoul, S., 2015. A cross saliency approach to asymmetry-based tumor detection. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 636–643.

Feng, C., Zhao, D., Huang, M., 2015. Segmentation of ischemic stroke lesions in multi-spectral MR images using weighting suppressed FCM and three phase level set. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, pp. 233–245.

Gal, Y., Ghahramani, Z., 2015. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. arXiv preprint arXiv:1506.02142.

Geremia, E., Menze, B. H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N., 2010. Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010. Springer, pp. 111–118.

Glorot, X., Bengio, Y., 2010. Understanding the difficulty of training deep feedforward neural networks. In: International conference on artificial intelligence and statistics. pp. 249–256.

Gooya, A., Pohl, K. M., Bilello, M., Biros, G., Davatzikos, C., 2011. Joint segmentation and deformable registration of brain scans guided by a tu-

mor growth model. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011. Springer, pp. 532–540.

Grabner, G., Janke, A. L., Budge, M. M., Smith, D., Pruessner, J., Collins, D. L., 2006. Symmetric atlasing and model based segmentation: An application to the hippocampus in older adults. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2006. Springer, pp. 58–66.

Halme, H.-L., Korvenoja, A., Salli, E., 2015. ISLES (SISS) challenge 2015: segmentation of stroke lesions using spatial normalization, Random Forest classification and contextual clustering. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, pp. 211–221.

Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., Larochelle, H., 2015. Brain tumor segmentation with deep neural networks. arXiv preprint arXiv:1505.03540.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1026–1034.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. R., 2012. Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Iglesias, J. E., Liu, C.-Y., Thompson, P. M., Tu, Z., 2011. Robust brain extraction across datasets and comparison with publicly available methods. Medical Imaging, IEEE Transactions on 30 (9), 1617–1634.

Ikram, M. A., Vrooman, H. A., Vernooij, M. W., den Heijer, T., Hofman, A., Niessen, W. J., van der Lugt, A., Koudstaal, P. J., Breteler, M. M., 2010. Brain tissue volumes in relation to cognitive function and risk of dementia. Neurobiology of Aging 31 (3), 378–386.

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167.

Irimia, A., Wang, B., Aylward, S. R., Prastawa, M. W., Pace, D. F., Gerig, G., Hovda, D. A., Kikinis, R., Vespa, P. M., Van Horn, J. D., 2012. Neuroimaging of structural pathology and connectomics in traumatic brain injury: Toward personalized outcome prediction. NeuroImage: Clinical 1 (1), 1–17.

Jarrett, K., Kavukcuoglu, K., Ranzato, M., LeCun, Y., 2009. What is the best multi-stage architecture for object recognition? In: Computer Vision, 2009 IEEE 12th International Conference on. IEEE, pp. 2146–2153.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B., 2015. Multi-scale 3D convolutional neural networks for lesion segmentation in brain MRI. in proc of ISLES-MICCAI.

Kappos, L., Freedman, M. S., Polman, C. H., Edan, G., Hartung, H.-P., Miller, D. H., Montalbán, X., Barkhof, F., Radü, E.-W., Bauer, L., et al., 2007. Effect of early versus delayed interferon beta-1b treatment on disability after a first clinical event suggestive of multiple sclerosis: a 3-year follow-up analysis of the BENEFIT study. The Lancet 370 (9585), 389–397.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected CRFs with gaussian edge potentials. Adv. Neural Inf. Process. Syst.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. ImageNet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86 (11), 2278–2324.

Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V. F., Makropoulos, A., Lötjönen, J., Menon, D. K., Rueckert, D., 2015. Robust whole-brain segmentation: Application to traumatic brain injury. Medical image analysis 21 (1), 40–58.

Lehtonen, S., Stringer, A. Y., Millis, S., Boake, C., Englander, J., Hart, T., High, W., Macciocchi, S., Meythaler, J., Novack, T., et al., 2005. Neuropsychological outcome and community re-integration following traumatic brain injury: the impact of frontal and non-frontal lesions. Brain Injury 19 (4), 239–256.

Li, R., Zhang, W., Suk, H.-I., Wang, L., Li, J., Shen, D., Ji, S., 2014. Deep learning based imaging data completion for improved brain disease diagnosis. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 305–312.

Liu, X., Niethammer, M., Kwitt, R., McCormick, M., Aylward, S., 2014. Low-rank to the rescue – atlas-based analyses in the presence of pathologies. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 97–104.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3431–3440.

Lyksborg, M., Puonti, O., Agn, M., Larsen, R., 2015. An ensemble of 2D convolutional neural networks for tumor segmentation. In: Image Analysis. Springer, pp. 201–211.

Maas, A. I., Menon, D. K., Steyerberg, E. W., Citerio, G., Lecky, F., Manley, G. T., Hill, S., Legrand, V., Sorgner, A., Participants, C.-T., Investigators, et al., 2015. Collaborative european neurotrauma effectiveness research in traumatic brain injury (CENTER-TBI): A prospective longitudinal observational study. Neurosurgery 76 (1), 67–80.

Maier, O., Menze, B. H., , et al., 2017. ISLES 2015 – a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Medical Image Analysis 35, 250–269.

Menze, B. H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al., 2015. The multimodal brain tumor image segmentation benchmark (BRATS). Medical Imaging, IEEE Transactions on 34 (10), 1993–2024.

Mitra, J., Bourgeat, P., Fripp, J., Ghose, S., Rose, S., Salvado, O., Connelly, A., Campbell, B., Palmer, S., Sharma, G., et al., 2014. Lesion segmentation from multimodal MRI using Random Forest following ischemic stroke. NeuroImage 98, 324–335.

Moen, K. G., Skandsen, T., Folvik, M., Brezova, V., Kvistad, K. A., Rydland, J., Manley, G. T., Vik, A., 2012. A longitudinal MRI study of traumatic

axonal injury in patients with moderate and severe traumatic brain injury. Journal of Neurology, Neurosurgery & Psychiatry, jnnp–2012.

Noh, H., Hong, S., Han, B., 2015. Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1520–1528.

Parisot, S., Duffau, H., Chemouny, S., Paragios, N., 2012. Joint tumor segmentation and dense deformable registration of brain MR images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 651–658.

Pereira, S., Pinto, A., Alves, V., Silva, C. A., 2015. Deep convolutional neural networks for the segmentation of gliomas in multi-sequence MRI. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries. Springer, pp. 131–143.

Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., Nielsen, M., 2013. Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 246–253.

Prastawa, M., Bullitt, E., Ho, S., Gerig, G., 2004. A brain tumor segmentation framework based on outlier detection. Medical image analysis 8 (3), 275–283.

Rao, A., Ledig, C., Newcombe, V., Menon, D., Rueckert, D., 2014. Contusion segmentation from subjects with traumatic brain injury: A Random Forest framework. In: Biomedical Imaging (ISBI), 2014 IEEE 11th International Symposium on. IEEE, pp. 333–336.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 234–241.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R. M., 2014. A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 520–527.

Rovira, À., León, A., 2008. MR in the diagnosis and monitoring of multiple sclerosis: an overview. European journal of radiology 67 (3), 409–414.

Schmidt, P., Gaser, C., Arsic, M., Buck, D., Förschler, A., Berthele, A., Hoshi, M., Ilg, R., Schmid, V. J., Zimmer, C., et al., 2012. An automated tool for detection of FLAIR-hyperintense white-matter lesions in multiple sclerosis. Neuroimage 59 (4), 3774–3783.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2014. Overfeat: Integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations.

Sharp, D. J., Beckmann, C. F., Greenwood, R., Kinnunen, K. M., Bonnelle, V., De Boissezon, X., Powell, J. H., Counsell, S. J., Patel, M. C., Leech, R., 2011. Default mode network functional and structural connectivity after traumatic brain injury. Brain 134 (8), 2233–2247.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al., 2016. Mastering the game of Go with deep neural networks and tree search. Nature 529 (7587), 484–489.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

Springenberg, J. T., Dosovitskiy, A., Brox, T., Riedmiller, M., 2014. Striving for simplicity: The all convolutional net. arXiv preprint arXiv:1412.6806.

Studholme, C., Hill, D. L., Hawkes, D. J., 1999. An overlap invariant entropy measure of 3D medical image alignment. Pattern recognition 32 (1), 71–86.

Sutskever, I., Martens, J., Dahl, G., Hinton, G., 2013. On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th international conference on machine learning (ICML-13). pp. 1139–1147.

Tieleman, T., Hinton, G., 2012. Lecture 6.5-RMSprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning.

Tustison, N., Wintermark, M., Durst, C., Brian, A., 2013. ANTs and Arboles. in proc of BRATS-MICCAI.

Urban, G., Bendszus, M., Hamprecht, F., Kleesiek, J., 2014. Multi-modal brain tumor segmentation using deep convolutional neural networks. in proc of BRATS-MICCAI.

Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P., 1999. Automated model-based tissue classification of MR images of the brain. Medical Imaging, IEEE Transactions on 18 (10), 897–908.

Warner, M. A., de la Plata, C. M., Spence, J., Wang, J. Y., Harper, C., Moore, C., Devous, M., Diaz-Arrastia, R., 2010. Assessing spatial relationships between axonal integrity, regional brain volumes, and neuropsychological outcomes after traumatic axonal injury. Journal of neurotrauma 27 (12), 2121–2130.

Weiss, N., Rueckert, D., Rao, A., 2013. Multiple sclerosis lesion segmentation using dictionary learning and sparse coding. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 735–742.

Wen, P. Y., Macdonald, D. R., Reardon, D. A., Cloughesy, T. F., Sorensen, A. G., Galanis, E., DeGroot, J., Wick, W., Gilbert, M. R., Lassman, A. B., et al., 2010. Updated response assessment criteria for high-grade gliomas: response assessment in neuro-oncology working group. Journal of Clinical Oncology 28 (11), 1963–1972.

Ye, D. H., Zikic, D., Glocker, B., Criminisi, A., Konukoglu, E., 2013. Modality propagation: Coherent synthesis of subject-specific scans with data-driven regularization. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013. Springer, pp. 606–613.

Yuh, E. L., Cooper, S. R., Ferguson, A. R., Manley, G. T., 2012. Quantitative CT improves outcome prediction in acute traumatic brain injury. Journal of neurotrauma 29 (5), 735–746.

Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P. H., 2015. Conditional random fields as recurrent neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1529–1537.

Zikic, D., Glocker, B., Konukoglu, E., Criminisi, A., Demiralp, C., Shotton, J., Thomas, O., Das, T., Jena, R., Price, S., 2012. Decision forests for tissue-specific segmentation of high-grade gliomas in multi-channel MR. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 369–376.

Zikic, D., Ioannou, Y., Brown, M., Criminisi, A., 2014. Segmentation of brain tumor tissues with convolutional neural networks. in proc of BRATS-MICCAI.