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

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

\textsuperscript{a}Biomedical Image Analysis Group, Imperial College London, UK
\textsuperscript{b}University Division of Anaesthesia, Department of Medicine, Cambridge University, UK
\textsuperscript{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 developed an efficient and effective dense training scheme which automatically adapts to the inherent class imbalance present in the data. The training makes use of the notion of image segments which joins multiple patches from the same image into one pass through the network. 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 and achieves a segmentation of a brain scan in less than six minutes. The source code of our implementation is made publicly available.

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

March 21, 2016
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, for example, of lesion volume may add insight into the functional outcome of patients (Ding et al. (2008); Moen et al. (2012)). 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)). Another example are brain tumors, where accurate delineation and estimation of the relative volume of a tumor’s sub-components are important for planning radiotherapy and treatment follow-up (Wen et al. (2010)). More generally, volume and number of lesions are biomarkers that have been shown to be related to cognitive deficits (Jokinen et al. (2009)) and are informative for assessing the effectiveness of different therapies (Kappos et al. (2007)). For example, the volume of white matter lesions (WMLs) is considered important in assessing stroke (Ikram et al. (2010)). The functional deficits caused by brain damage after stroke are likely to be associated with the extent of the damage to particular brain structures (Palmer and Carey (2013)). Further, in clinical research on multiple sclerosis (MS), lesion load and lesion count are extracted from brain MRI and used in models that aim to better understand disease progression and effectiveness of pharmaceutical treatment (Ge (2006); Rovira and León (2008)).

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 introduces 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 for 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 representable for 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 some 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 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 segmentation 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 (RF). 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); Dalca 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.\footnote{links: http://braintumorsegmentation.org/, www.isles-challenge.org}
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. Most related to our own approach are the CNN based tumor segmentation methods of Zhang et al. (2015); Zikic et al. (2014); Davy et al. (2014); Havaei et al. (2015); Pereira et al. (2015), 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 that 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.

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); Dvok and Menze (2015)). To incorporate 3D contextual information, early attempts used 2D CNNs on three orthogonal 2D patches (Prasoon et al. (2013); Roth et al. (2014, 2015); 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. (2013)), 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.

Segmentation performance of CNNs is significantly influenced by the strategy for extracting training samples. A commonly adopted approach is training on data that is equally sampled from each class. This, however, biases the classifier towards rare classes and may result in over-segmentation.
To counter this, Ciresan et al. (2013) and Havaei et al. (2015) proposed two-stage training schemes, employing uniform sampling in the second stage to capture the true class distribution. Alternatively, dense-training schemes (Long et al. (2015)) were used by Urban et al. (2014) to train on all voxels within a random part of an image and in (Brosch et al. (2015); Ronneberger et al. (2015)) for training on all voxels of an image per optimization step. Similarly to uniform-sampling, dense-training schemes introduce severe class-imbalance. Brosch et al. (2015); Ronneberger et al. (2015) proposed different weighted loss functions to alleviate this problem.

1.2. Contributions

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

1. We propose a novel, efficient hybrid training scheme, utilizing dense training (Long et al. (2015)) on large image segments, which automatically adapts 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. Last but not least, our method outperforms the state-of-the-art on a variety of highly challenging segmentation tasks, with top ranking results in two MICCAI competitions, 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 (2012) 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 neighbourhoods while preserving low inference times.

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 dense CRF is made publicly available on https://biomedia.doc.ic.ac.uk/software/deepmedic/.

1.3. Overview

The rest of the article is organised as follows: In Sec. 2 we first describe the architectural choices we made to alleviate problems arising from the high computational requirements of 3D CNNs in our goal to build a more powerful network. We then describe the fully-connected CRF that we employ for structural refinement of the predicted segmentation. We conducted a series of experiments to individually quantify the main contributions and examine the components of our network architecture, which are presented in Sec. 3. In Sec. 4 we provide a quantitative evaluation of our segmentation system on datasets of TBI, brain tumour and stroke lesions. Finally, discussion of the results and insights we acquired during this work follows in Sec. 5.

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 $y_{lm} = f\left(\sum_{n=1}^{C_{l-1}} k_{l}^{m,n} \ast y_{l-1}^{n} + b_{l}^{m}\right)$. This is the result of convolving each of the previous layer’s channels with a 3-dimensional kernel $k_{l}^{m,n}$, adding a learned bias $b_{l}^{m}$ and applying a non-linearity $f$. Each kernel is a matrix of learned hidden weights $W_{l}^{m,n}$. The images $y_{0}^{n}$, 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 $k_{l} = (k_{l}^{m,1},...,k_{l}^{m,C_{l-1}})$ can be viewed as a 4-dimensional kernel convolving the concatenated channels $y_{l-1} = (y_{l-1}^{1},...,y_{l-1}^{C_{l-1}})$, 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 $p_{c}(x) = \exp(y_{L}^{c}(x))/\sum_{c=1}^{C_{L}} \exp(y_{L}^{c}(x))$ for each class $c$, which form soft segmentation maps with (pseudo-)probabilities. $y_{L}^{c}(x)$ is the activation of the $c$-th classification FM at position $x \in \mathbb{N}^{3}$. 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 neighbourhood of voxels in the input that influence the activation of a neuron is its receptive field. Its size increases with each subsequent layer and is given by the 3-dimensional vector:

$$f_{l}^{(x,y,z)} = f_{l-1}^{(x,y,z)} + (i_{l}^{(x,y,z)} - 1)s_{l}^{(x,y,z)},$$

(1)
where $\hat{k}_l, \hat{s}_l \in \mathbb{N}^3$ are vectors expressing the size of the kernels and stride of the receptive field at layer $l$. $\hat{s}_l$ 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 $\hat{s}_l = (1, 1, 1)$. The receptive field of the neurons in the classification layer corresponds to the image patch that influences the prediction of the central voxel. This is called the CNN’s receptive field, with $\hat{f}_{CNN} = \hat{f}_L$.

If input of size $\hat{d}_m$ is provided, the dimensions of the FMs in layer $l$ are given by:

$$\hat{d}_l^{x,y,z} = \left\lfloor (\hat{d}_m^{x,y,z} - \hat{f}_l^{x,y,z}) / \hat{s}_l^{x,y,z} + 1 \right\rfloor$$  \hspace{2cm} (2)

In the common patch-wise classification setting, an input patch of size $\hat{d}_m = \hat{f}_{CNN}$ 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 $\hat{f}_{CNN}$ is provided (Sermanet et al. (2013)). 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). 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 high number of FMs needs 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 $\hat{f}_{CNN}$ 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 $\Theta$, 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:
where the pair \((I^p, c^p), \forall p \in [1, B]\) is the \(p\)-th patch in the batch and the true label of its central voxel, while the scalar value \(p_{c^p}\) is the predicted posterior for class \(c^p\). 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, 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 \(f_{CNN}\) is given as input to our network, the output is a posterior probability for multiple voxels \(V = \prod_{i=x,y,z} d_{L}^{(i)}\). If the training batches are formed of \(B\) segments randomly extracted from the training images, the cost function (3) in the case of dense-training becomes:

\[
J_D(\Theta; I_s, c_s) = -\frac{1}{B \cdot V} \sum_{s=1}^{B} \sum_{v=1}^{V} \log(p_{c^v}(x^v)) ,
\]

where \(I_s\) and \(c_s\) are the \(s\)-th segment of the batch and the true labels of its \(V\) predicted voxels respectively. \(c^v\) is the true label of the \(v\)-th voxel, \(x^v\) the corresponding position in the classification FMs and \(p_{c^v}\) 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 dense training scheme on a whole image (Long et al. (2015)), which is not applicable to large 3D CNNs due to memory limitations, and the commonly used training on individual patches.
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.

The sampling of input segments provides also 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, 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 elegantly adapts 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, the 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 in relevance 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 $C_l C_{l-1} \prod_{i=x,y,z} k_l^{(i)}$ weights to the model. Remember, $C_l$
is the number of FMs in layer $l$ and $\hat{k}_{l}^{\{x,y,z\}}$ 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)). 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 to train can be designed by simply replacing each layer of common architectures with more layers that use smaller kernels (Fig. 4).

![Figure 4](image_url)

**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 ($Number \times Size$).

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 $n_{l}^{in} \cdot var(W_{l})$, where $n_{l}^{in} = C_{l-1} \prod_{i=x,y,z} \hat{k}_{l}^{(i)}$ is the number of weights through which a neuron of layer $l$ is connected to its input and $var(W_{l})$ 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 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 $n_{l}^{in} \cdot \text{var}(W_{l,t+1}) = n_{l}^{in} \cdot (\text{var}(W_{l,t}) + \epsilon_{l,t})$. 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 introduces spatial invariance, an undesirable characteristic for segmentation.

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 on 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 $F_{S}$. 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.

In order to preserve the capability of dense inference, the second pathway is integrated as follows: For every $F_{S}$ strides of the receptive field $f_{L1}$ at the
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 $(\text{Number} \times \text{Size})$.

The two input segments can then be extracted centered on the same image location. The FMs of $L_2$ are up-sampled to match the dimensions of $L_1$’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.

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 then downsampling 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. Ad-
ditionally, 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 Dense CRF for Structured Prediction

Because neighboring voxels usually share substantial spatial context, the soft segmentation label maps produced by the CNN tend to be very smooth, even though neighborhood dependencies are not modeled directly. However, local minima during training and noise in the input images can still result in some spurious outputs, with small isolated regions or holes in the label map. We employ a fully connected CRF (Krähenbühl and Koltun (2012)) as a post-processing step in order to “clean-up” the CNN results and 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 $I$ and the label configuration (segmentation) $z$, the Gibbs energy in a CRF model is given by

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

The unary potential is the negative log-likelihood $\psi_u(z_i) = -log P(z_i|I_i)$, where in our case $P(z_i|I_i)$ is the output of the CNN for voxel $i$. In a fully connected CRF, the pairwise potential is of form $\psi_p(z_i, z_j) = \mu(z_i, z_j)k(f_i, f_j)$ between any pair of voxels, regardless of their spatial distance. The Pott’s Model is commonly used as the label compatibility function between two voxels, 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 $f_i, f_j$ being the feature vectors of the pair of voxels. Krähenbühl and Koltun (2012) made the important observation that if the penalty function is defined as a linear combination of Gaussian kernels, $k(f_i, f_j) = \sum_{m=1}^{M} w^{(m)} k^{(m)}(f_i, f_j)$, 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 $f_i, f_j$.

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, $k^{(1)}(f_i, f_j) = \exp\left(-\sum_{d=\{x,y,z\}} \frac{|p_{i,d} - p_{j,d}|^2}{2\sigma_{\alpha,d}^2}\right)$, is defined by a diagonal covariance matrix with elements the configurable parameters $\sigma_{\alpha,d}$, which express the size of neighbourhoods that homogeneous labels should be encouraged. The appearance kernel $k^{(2)}(f_i, f_j) = \exp\left(-\sum_{d=\{x,y,z\}} \frac{|p_{i,d} - p_{j,d}|^2}{2\sigma_{\beta,d}^2} - \sum_{c=1}^{M} \frac{|I_{i,c} - I_{j,c}|^2}{2\sigma_{\gamma,c}^2}\right)$ is defined similarly. The additional parameters $\sigma_{\gamma,c}$ can be interpreted as how strongly to enforce homogeneous appearance in the various 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.

Inference with this CRF model requires convolutions in a $3 + C$ dimensional feature space. In our implementation we utilize the same highly efficient data structure for optimizing the high-dimensional filtering as in the original work, which is based on permutohedral lattices (Adams et al. (2010)). 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 not found necessary in the scope of this work.

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 finally evaluate the influence of the multi-scale dual pathway (cf. Sec. 2.4).

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.

3.2. Effect of Dense Training on Image Segments

![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.

We compare our proposed dense training method with two other commonly used training schemes on the 5-layers baseline CNN (see Fig. 2; for further configuration details see appendix). The first common scheme trains on 17³ patches extracted uniformly from the brain region, and the second scheme samples patches equally from the lesion 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 accuracy on the validation samples since the majority is background, but not particularly high DSC scores due to under-segmentation (low sensitivity).

To evaluate our dense training scheme, we trained multiple models with varying sized image segments, equally sampled from lesions and background. The tested sizes of the segments go from 19³ upwards to 29³. The models are
Figure 7: Mean accuracy over validation samples and DSC for the segmentations of the validation images, as obtained from the shallower and deeper network variants. Training of the plain “Deep” variant fails due to the problem of signal vanishing at the initial stage of training (cf. Sec. 3.3). Deep+ performs significantly better than Shallow+ with similar computation time.

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 were adjusted to have a similar memory footprint and lead to similar training times as compared to training on $P_{uni}$ and $P_{eq}$ \(^2\). We observe a significant 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 very much improved segmentation as expressed by the DSC.

The segment size is a hyper-parameter in our model, for which 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. For the remaining experiments, all models were trained on segments of size $25^3$.

3.3. Effect of Deeper Networks

The 5-layers baseline CNN, here referred to as the “Shallow” model, is extended to 9-layers by replacing each convolutional layer that uses $5^3$ ker-

\(^2\)Note that dense training on a whole image was impossible to apply on this 3D CNN due to memory limitations, but was previously shown to give similar results as training on uniformly sampled patches (Long et al. (2015)).
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 capturing greater context yields best performance, while BigDeep+ seems to suffer from over-fitting.

Nets with two layers that use $3^3$ kernels. This model is 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 much higher values (see appendix) and overcomes this problem. Further preservation of the signal can be 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 performance improvement of Deep+ over Shallow+, as shown in Fig. 7, is the result of the greater representational power of the deeper network. Regarding computational time, the two models perform very similarly. 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

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. 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 limited receptive field of $33^2$ voxels.

Figure 8 shows the improvement achieved with DeepMedic 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, to confirm that the performance increase can be accounted to the additional context and not the additional capacity of the network, we built a single-scale model with twice as many FMs as Deep+ and two additional hidden layers, referred to as “BigDeep+”. This model did not improve the performance, while showing signs of over-fitting.

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 different 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
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.

Research Ethics Committee (LREC 97/290) and written assent via consultant agreement was obtained for all patients. The structural MRI sequences that are used in this work are isotropic MPRAGE (1mm×1mm×1mm), axial FLAIR, T2 and Proton Density (PD) (0.7mm×0.7mm×5mm), and Gradient-Echo (GE) (0.86mm×0.86mm×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)), and all images were resampled to an isotropic 1 mm$^3$ resolution, with dimensions 193×229×193, 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 add images with shifted intensities where each MR channel $c$ of a training segment is modified according to $i_c = r_c \sigma_c$. $r_c \sim \mathcal{N}(0,0.1)$ and $\sigma_c$ is the standard deviation of intensities under the brain mask. 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 three layers. The network is evaluated with 5-fold cross-validation on the 61 subjects. Each training session requires approximately two days on an NVIDIA GTX Titan X GPU.

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 CNN architecture. The 15 subjects are reshuffled into the 5-folds used for subsequent evaluation.
Random Forest 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. Training data points are approximately equally sampled from lesion and background classes. Two hundred randomized cross-channel box features are evaluated at each split node with maximum offsets and box sizes of 20 mm. The same folds of training and test sets are used as for our CNN approach.

4.1.3. Results

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 |
|------------|-------|-----------|-------------|-------|-----------|
| Random Forest | 51.1(20.0) | 50.1(24.4) | 60.1(15.8) | 8.29(6.76) | 64.17(15.98) |
| 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)** |

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 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 for both the single network and the ensemble. Fig. 11 shows two example cases.

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). Continuous lines represent mean values. The outlying subject 8 presents small TBI lesions, which are successfully segmented by the network, but also vascular ischemia. Because it is the only case in the database with the latter pathology, the network fails to segment it as it has not seen such lesion during training.

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 into different sets of whole tumor (all four classes), the core (classes 1,3,4),
and the enhancing tumor (class 4). 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 $1\text{mm}^3$ resolution. Dimensions of each volume are $240 \times 240 \times 155$. 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). We train the network with the same configuration as described in Sec. 3.4. We enrich the dataset with sagittal reflections. Opposite to the experiments on TBI, we do not employ the intensity perturbation and dropout on convolutional layers, because the network should not require as much regularisation thanks to the large size of the tumor database. The network is trained on image segments
extracted with equal probability centred on the whole tumor and healthy tissue.

In order to examine the behaviour of our network, we first evaluated 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. The training time for each run is approximately two days.

**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, in comparison to other submissions, as displayed on the online evaluation platform 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   | 77   | 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 influence of the latter two is statistically significant, although 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 top performing automatic method, the 2D CNN of Pereira et al. (2015) (peres1). Visual segmentation examples 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. Important to note is that the benefit of removing the false positive predictions with the CRF is much greater in this case.

Table 3: Average performance of our system on the test data 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*       | 85.5  | 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. In the training phase of the challenge, 28 datasets have been made available, along with manual segmentations.
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. (middle and bottom) A common mistake is the false segmentation of FLAIR hyper-intensities as oedema.

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 for all training images. Corresponding expert segmentation 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 described 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.

[^3]: link:www.isles-challenge.org
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 entry. Values correspond to the mean (and standard deviation).

|            | DSC   | Precision | Sensitivity | ASSD   | Hausdorff   |
|------------|-------|-----------|-------------|--------|-------------|
| 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) |
| halnh1     | 47(32)| 47(34)    | 56(33)      | 14.61(20.17) | 46.26(34.81) |

indicating superior performance on this challenging task among 14 submissions. Table 5 shows our results, along with the other two top entries (Feng et al. (2015); Halme et al. (2015)). 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)). The efficient architecture of DeepMedic allows models to be trained on GPUs with only 3GB of memory, however, the dimensions of the volumes in the processed databases do not allow dense training on whole volumes. Because inference requires only a forward-pass and thus less memory, dense inference on a whole volume is still possible in less than one minute, but requires 12GB NVIDIA GTX Titan X GPU. Tiling the volume into multiple segments of size $35^3$ allows inference on 3GB GPUs in less than three minutes.

Our 3D fully connected CRF is implemented by extending the original source code by Krähenbühl and Koltun (2012). The CPU implementation is capable of processing a five-channel brain scan in under three minutes, with potential improvement when considering a GPU implementation.

5. Discussion and Conclusion

We have presented DeepMedic, an automatic lesion segmentation system, that surpasses state-of-the-art on challenging data. The proposed novel training scheme is not only computationally efficient but also offers an adaptive
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.

way of alleviating the inherent class-imbalance of segmentation problems. We analyzed the benefits of using small convolutional 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 to avoid overfitting. 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.
Figure 14: (First row) GE scan and DeepMedic’s segmentation. (Second row) FMs of earlier and (third row) deeper layers of the normal 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.
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. 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 achieves the highest reported accuracy when applied on a cohort of patients with severe TBI. As a comparison, we improved over the reported performance of the pipeline presented in Rao et al. (2014). Important to note is that their work focused only on segmentation of contusions, while our system has been shown capable of segmenting even small and diffused pathologies. Additionally, our pipeline achieves 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, for example non-brain applications, the part that needs most attention is the reconfiguration of the CRF. The model improved our system’s performance with statistical significance in all investigated tasks, proving its flexibility and potential. Finding optimal parameters for each task, however, can be challenging in certain applications. This problem became most obvious on the task of multi-class tumor segmentation. Because the substructures of the tumor vary significantly in appearance, we encountered difficulties in finding a global set of parameters for the refinement that yields improvements on all classes. Instead, we applied the CRF in a binary fashion, refining only the boundaries of the whole tumor while suppressing spurious false positives. This CRF model can be configured with a separate set of parameters for each class. However the increased size of the parameter space would make its configuration particularly challenging. 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, and 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 within DeepMedic on the lesion segmentation tasks, we explored what patterns have been learned automatically. For this, we visualize the activations of DeepMedic’s FMs when processing a subject from our TBI database. Many of the appearing patterns are difficult to interpret, especially in deeper layers. However, in Fig. 14 we provide some examples that seem to 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 the differentiation of the tissue type is beneficial for lesion segmentation. This is in line with findings in the literature, where the 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 by the neurons of the low-resolution pathway. As they process greater image context, these neurons gain additional localization capabilities. The activations of certain FMs form fields in the surrounding areas of the brain, patterns which are preserved in the deepest hidden layers, which indicates that these are beneficial for the final segmentation (see two last rows of Fig. 14). We believe that such cues can be useful for providing 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 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 complex, 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, better ways for making the inference process interpretable are required, which would also allow us to better understand 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, but rather a value automatically calibrated to maximize the cross entropy over the training samples. 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 alleviating this problem is to learn a generative model for the data acquisition process, and use this model in the data augmentation step. That way, the CNN could be made invariant to the data heterogeneity. 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
Acknowledgements

This work is partially funded under the 7th Framework Programme by the European Commission (TBIcare: http://www.tbicare.eu/; CENTER-TBI: https://www.center-tbi.eu/). This work was further supported by a Medical Research Council (UK) Program Grant (Acute brain injury: heterogeneity of mechanisms, therapeutic targets and outcome effects [G9439390 ID 65883]), the UK National Institute of Health Research Biomedical Research Centre at Cambridge and Technology Platform funding provided by the UK Department of Health. KK is supported by the Imperial College London PhD Scholarship Programme. VFJN is supported by a Health Foundation/Academy of Medical Sciences Clinician Scientist Fellowship. DKM is supported by an NIHR Senior Investigator Award. We gratefully acknowledge the support of NVIDIA Corporation with the donation of two Titan X GPUs for our research.

References

Adams, A., Baek, J., Davis, M. A., 2010. Fast high-dimensional filtering using the permutohedral lattice. Computer Graphics Forum 29 (2), 753–762.

Badrinarayanan, V., Handa, A., Cipolla, R., 2015. Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293.

Bakas, S., Zeng, K., Sotiras, A., Rathore, S., Akbari, H., Gaonkar, B., Rozycki, M., Pati, S., Davatzikos, C., 2015. Segmentation of gliomas in multimodal magnetic resonance imaging volumes based on a hybrid generative-discriminative framework. in proc of BRATS-MICCAI.

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.

Brebiison, 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.

Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., LeCun, Y., 2015. The Loss Surfaces of Multilayer Networks. Aistats 38, 192–204.

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.

Ciresan, 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.

Dalca, A. V., Sridharan, R., Cloonan, L., Fitzpatrick, K. M., Kanakis, A., Furie, K. L., Rosand, J., Wu, O., Sabuncu, M., Rost, N. S., et al., 2014. Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2014. Springer, pp. 773–780.

Davy, A., Havaei, M., Warder-Farley, D., Biard, A., Tran, L., Jodoin, P.-M., Courville, A., Larochelle, H., Pal, C., Bengio, Y., 2014. Brain tumor segmentation with deep neural networks. In proc of BRATS-MICCAI.

Ding, K., de la Plata, C. M., Wang, J. Y., Humphrey, 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.
Dvok, P., Menze, B., 2015. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. In proc of BRATS-MICCAI.

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 stroke lesions in multispectral mr images using bias correction embedded fcm and three phase level set. In proc of ISLES-MICCAI.

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

Ge, Y., 2006. Multiple sclerosis: the role of mr imaging. American Journal of Neuroradiology 27 (6), 1165–1176.

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 tumor 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., Korvenoja, A., Salli, E., 2015. Segmentation of stroke lesions using spatial normalization, random forest classification and contextual clustering. in proc of ISLES-MICCAI.

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.
Jokinen, H., Kalska, H., Ylikoski, R., Madureira, S., Verdelho, A., Van Der Flier, W. M., Scheltens, P., Barkhof, F., Visser, M. C., Fazekas, F., et al., 2009. Longitudinal cognitive decline in subcortical ischemic vascular disease—the ladis study. Cerebrovascular diseases 27 (4), 384–391.

Kamnitsas, K., Chen, L., Ledig, C., Rueckert, D., Glocker, B., 2015. Multi-scane 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., Radii, 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., 2012. Efficient inference in fully connected crfs with gaussian edge potentials. arXiv preprint arXiv:1210.5644.

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.

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.

Palmer, S., Carey, L. M., 2013. Biologically reliable white matter fiber tractography: issues and solutions. Future Neurology 8 (6), 613–616.

Parisot, S., Duffau, H., Chemouny, S., Paragios, N., 2012. Joint tumor segmentation and dense deformable registration of brain mri images. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2012. Springer, pp. 651–658.
Pereira, S., Adriano, P., Alves, V., Silva, C., 2015. Deep convolutional neural networks for the segmentation of gliomas in multi-sequence mri. in proc of BRATS-MICCAI.

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., Farag, A., Shin, H.-C., Liu, J., Turkbey, E. B., Summers, R. M., 2015. Deeporgan: Multi-level deep convolutional networks for automated pancreas segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015. Springer, pp. 556–564.

Roth, H. R., Lu, L., Seff, A., Cherry, K. M., Hoffman, J., Wang, S., Liu, J., Turkbey, E., Summers, R. M., 2014. A new 2.5 d 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., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229.

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.

Zhang, W., Li, R., Deng, H., Wang, L., Lin, W., Ji, S., Shen, D., 2015. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 108, 214–224.

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.

Appendix

Network Configuration details: The core description of our system is presented in Sec. 2. All models 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 $m = 0.6$. We train the networks with dense-training on batches of 10 segments unless otherwise specified. Segment size used is $25^3$ unless otherwise specified. $L_1 = 10^{-6}$ and $L_2 = 10^{-4}$ regularisation is applied. The weights of our shallow, 5-layers networks are initialized by sampling from a normal distribution $N(0, 0.01)$. Initial learning rate of shallow networks is set to $a = 10^{-4}$. Deeper models use the weight initialisation scheme of He et al. (2015). The scheme increases the signal’s variance in our settings, which leads to RMSProp decreasing the learning rate. To counter this, initial learning $a$ is set higher for the deeper models, to $10^{-3}$. Learning rate of all models is halved when convergence plateaus. Dropout with 50% rate is employed on the two last hidden layers of 11-layers deep models. The low-resolution features were upsampled by repetition in our experiments and was found sufficient, since the following hidden layers learn to combine the multi-scale features. Other learned upsampling methods have been recently proposed, which could be explored and easily integrated (Long et al. (2015); Ronneberger et al. (2015); Badrinarayanan et al. (2015)).