Multi-task Fully Convolutional Network for Brain Tumour Segmentation

Haocheng Shen, Ruixuan Wang, Jianguo Zhang, and Stephen McKenna

Computing, School of Science and Engineering, Univeristy of Dundee, UK

Abstract. We propose a novel, multi-task, fully convolutional network (FCN) architecture for automatic segmentation of brain tumour. This network extracts multi-level contextual information by concatenating hierarchical feature representations extracted from multimodal MR images along with their symmetric-difference images. It achieves improved segmentation performance by incorporating different segmentation tasks directly into the loss function. The model was evaluated on BRATS13 and BRATS15 datasets where the proposed multi-task FCN outperforms single-task FCN in all sub-tasks. The method is among the most accurate available and has relatively low computational cost at test time.

Keywords: deep learning, tumour segmentation, multi-task learning

1 Introduction

Precise localization of brain tumours in 3D MR images of patients is clinically crucial to make treatment plans, guide surgery and monitor the rehabilitation progress of patients. Any unreliable segmentation might cause irreversible impact (e.g., the difficulty in speaking fluently) from surgery. Since manually segmenting brain tumour particularly in 3D MR images is a tedious and time-consuming process, computer-aided automatic and reliable segmentation of brain tumour is necessary and would save much of clinician's valuable time.

Among brain tumors, gliomas appear most frequent [1] in adult patients, either at high grade (HG) or low grade (LG) according to the aggressive form of the disease. Due to the diversity and variation of tumour size, shape, location, and appearance of gliomas, multimodal MRs are often taken from patients to enhance the contrast of potential tumor and its structures. Figure 1 shows a representative HG gliomas tumor and its sub-regions whose boundaries have been delineated by experts. Normally the tumore region could be divided into four different sub-regions or structures: edema (green), necrosis (red), non-enhancing (blue) and enhancing (yellow), where the combination of the last three structures is also called *tumor core*.

The automatic segmentation of gliomas and its sub-structures is often formulated as a patch-level or voxel-level classification problem, where each (either 2D or 3D) patch or voxel in the 3D MR is classified as one type of the sub-structures, and the collection of all patches' or voxels' classification would generate the final

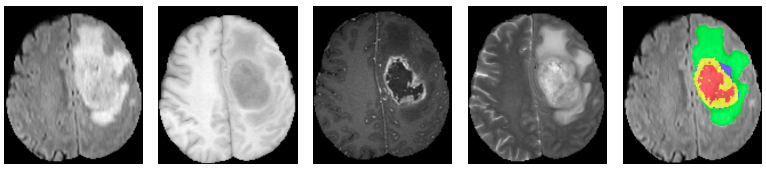

Fig. 1: An HG tumor. Left to right: Flair, T1, T1c, T2 and expert delineation; edema (green), necrosis (red), non-enhancing (blue), enhancing (yellow).

and whole segmentation of gliomas and its sub-structures. While hand-crafted features and conditional random field (CRF) incorporating class-label smoothness terms have been adopted for the voxel-level classification [1, 2], deep convolutional neural networks (CNNs), which have achieved substantial performance breakthroughs in several natural and medical image analysis benchmarks by automatically learning high-level discriminative feature representations, are not suprissingly achieving state-of-the-art results when applied to MRI brain tumor segmentation [3–5]. Specifically, Pereira et al [3] trained a traditional 2D CNN as a patch-level classifier, and Havaei et al [4] trained a 2D CNN to classify larger patches in a cascaded structure in order to capture both small and large-scale contextual information. Very recently, Kamnitsas et al [5] trained a 3D CNN directly on 3D instead of 2D patches and considered global contextual features via down-sampling. Note that all these methods are *patch*-level classification.

Different from traditional CNN models, fully convolutional networks (FCNs) were recently proposed by removing all the fully connected layers which often appear at the last several layers in CNNs, and have achieved promising results for natural image segmentation [10, 11] as well as medical image segmentation [12–14]. In FCNs, up-sampling (de-)convolutional layers are added on top of the traditional down-sampling convolutional layers, in order to gain the same spatial size at the network output as that of the original input. Compared to CNNs run on each sliding window of the input, FCNs run only once on the whole input and would generate the classification result for each voxel (or pixel). Therefore, FCNs as voxel-level classifiers are more computationally efficient than traditional CNNs as patch-level classifiers.

In this paper, we propose an new, tree-structured, multi-task FCN model for brain tumour segmentation. The main contributions of our work are: 1) to our best knowledge, we are the first to formulate and apply a tree-structured, multi-task FCN to multimodal brain tumour (and sub-structure) segmentation, which implicitly encodes the topology information of tumour subregion structures; 2) the tree-structured, multi-task FCN improves segmentation performance in all sub-tasks compared to single-task FC in both BRATS13 and BRATS15 datasets; 4) our model is ranked top on BRATS 2013 testing set, and more efficient than the other competitors.

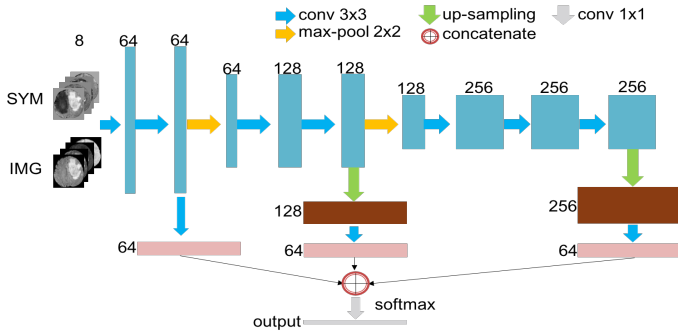

Fig. 2: Variant of FCN. Images and symmetry maps are concatenated as the input to the net [8]. Colored rectangles represent feature maps with numbers nearby being the number of feature maps. Best viewed in color.

2 Methodology

2.1 Variant of FCN

Our FCN variant includes a down-sampling path and three up-sampling paths (Fig. 2). The down-sampling path consists of three convolutional blocks separated by max pooling (see yellow arrows in Fig. 2). Each block includes 2∼3 convolutional layers as in the well-known VGG-16 network [7]. This down-sampling path extracts features ranging from small-scale low-level texture to larger-scale, higher-level features. For the three up-sampling paths, the FCN variant first up-samples the feature maps from the last convolutional layer of each convolutional block such that each up-sampled feature map (purple rectangles in Fig. 2) has the same spatial size as the input to the FCN. Then one convolutional layer is added to each up-sampling path to encode features at different scales. The output feature maps of the convolutional layer along the three up-sampling paths are concatenated before being fed to the final classification layer. We used ReLU activation functions and batch normalization after each convolutional layer. This FCN variant has been experimentally evaluated in a separate study [8].

2.2 Multi-task FCN

The above FCN can already produce good probability maps of tumor tissues. However, this FCN model predicts the class label for each pixel independently without encoding any high-order contextual information. In brain tumour context, we observe a topology in tumour structure: e.g., the complete tumour consists of tumour core and edema while tumour core consists of enhancing, non-enhancing and necrotic parts (see Fig. 1). Encoding this topological information into FCN framework will benefit tumour segmentation, for example, a core can never happen solely in the brain without edema region attached. To

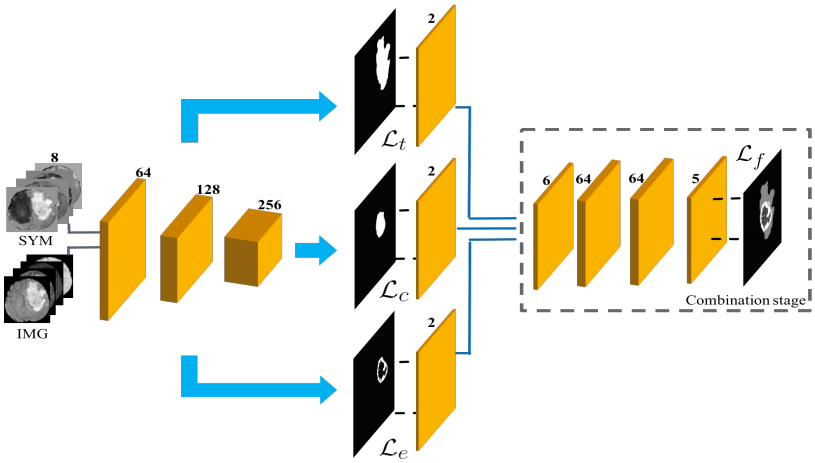

Fig. 3: The structure of multi-task FCN. The three up-sampling branches in the three FCNs are represented by blue arrows.

this end, we propose a deep multi-task network, which implicitly embeds the topology of tumour structure.

The structure of the proposed multi-task FCN is illustrated in Fig. 3. Instead of treating the segmentation task as a single pixel-wise classification problem, we formulate it within a multi-task learning framework. Three of the above FCN variants with shared down-sampling path and three different up-sampling branches (the blue arrows in Fig. 3) are applied for three separate tasks: *complete tumour*, *tumour core* and *enhancing tumour* classification. Then, the outputs (i.e., probability maps) from the three branches are concatenated and fed to a block of two convolutional layers followed by the final softmax classification layer ('combination stage' in Fig. 3). The combination stage considers the mutual inclusive information of all subregions in tumours estimated from the above three tasks. The 'combination stage' task is a 5-class classification task whereas the others are binary classification tasks. Cross-entropy loss is used for each task. Therefore, the total loss in our proposed multi-task FCN is

$$\mathcal{L}_{total}(\theta) = \sum_{m\in\{t,c,e,f\}} \mathcal{L}_m(\theta_m) = -\sum_{m\in\{t,c,e,f\}} \sum_n \sum_i \log P_m(l_m(x_{n,i}); x_{n,i}, \theta_m) \tag{1}$$

where $\{t, c, e, f\}$ are the tasks of *complete tumour*, *tumour core*, *enhancing core* and the final combination stage, respectively. $\theta = \{\theta_t, \theta_c, \theta_e, \theta_f\}$ is the set of weight parameters in the multi-task FCN. $\mathcal{L}_m$ refers to the loss function of each task. $x_{n,i}$ is the i-th voxel in the n-th image used for training, and P_m refers to the predicted probability of the voxel $x_{n,i}$ belonging to class l_m.

In the proposed multi-task FCN, 2D slices split from 3D MR volumes from axial view are used as part of input to the network. In addition, since adding

brain symmetry information has been proven helpful for FCN based tumor segmentation [8], 'symmetric intensity difference' maps are combined with original slices as the input, resulting in totally 8 input channels (see Fig 2 and 3) to the network.

3 Evaluation

Our model was evaluated on BRATS13 and BRATS15 datasets. BRATS13 contains 20 HG patients for training and 10 HGs for testing. (The 10 LG patients were not used for the HG segmentation task) . For BRATS15, we used 220 annotated HG patients' images in the training set. For each patient, there exist 4 modalities (T1, T1-contrast (T1c), T2 and Flair) which were skull-stripped and co-registered. Quantitative evaluation is performed on three sub-tasks: 1) the *complete* tumor (including all four tumor structures); 2) the tumour *core* (including all tumor structures except "edema"); 3) the *enhancing* tumor region (including only the "enhancing tumor" structure). For each sub-task, *Dice*, *Sensitivity* and *Positive Predictive Value* are computed.

Our model was implemented with the Keras library and Theano backend. For each MR image, voxel intensities were normalised to have zero mean and unit standard deviation. The network was trained with standard back-propagation using Adam optimizer. Learning rate was 0.001. The downsampling path was initialized with VGG-16 weights [7] while upsampling paths were initialized randomly using the strategy in [15].

3.1 Results on BRATS13 dataset

A 5-fold cross validation was performed on the 20 HG images in BRATS13 set. The training folds were augmented by scaling, rotating, flipping each image, before they were used for training the models.

We also compare a variant of the proposed multi-task model by replacing *core* task by *edema* task. The reason is: tumour *core* is a super-structure containing *enhancing*, *non-enhancing* and *necrotic* parts, which are different in texture appearance, e.g., in T1c (see Fig 1), *enhancing* parts show hyper-intensity signal whereas *necrosis* has low-intensity signal. The variability of *core* across patients is large as it depends the size of *enhancing* or *necrotic* parts. On the other hand, the texture appearance of *edema* is relatively consistent (e.g, hyper-intensity signal in Flair). Therefore, four models were compared on both validation set and test set: 1) variant of FCN (Fig.2), denoted **FCN**; 2) our implementation of FCN with topology loss proposed by [16], denoted **FCN+Topology**; 3) the multi-task FCN with core task, denoted as **mFCN_core**; 4) the multi-task FCN with edema task, denoted as **mFCN_edema**.

For the validation set, We plot *Dice* values of three tasks for the above four models at different epochs (every 5 epochs and up to 50) shown in the top row of Fig.4. It can be observed that although at the starting points (e.g., the fifth

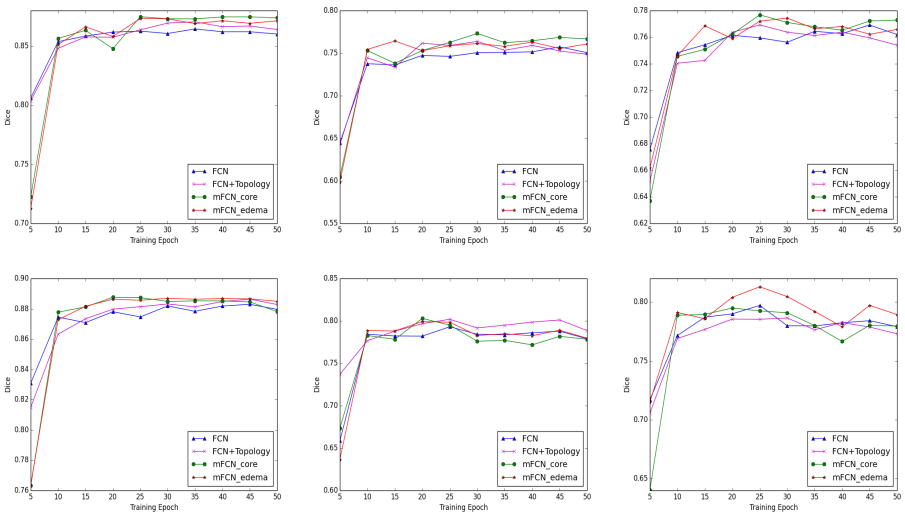

Fig. 4: From left to right: validation results of four models on *Complete*, *Core* and *Enhancing* tumor task. The vertical axis is Dice while horizontal axis is the number of epochs. The top row refers to BRATS13 while the bottom refers to BRATS15.

epoch), **mFCN_core** and **mFCN_edema** have lower performance due to the extra parameters in the network, they both outperforms baseline **FCN** in all tasks at most training epochs, especially for **mFCN_core**. **mFCN_edema** give competitive segmentation results in terms of *Complete* and *Enhancing* tasks while it is slightly worse in *Core* task compared to **mFCN_core**. We also observe using topology loss proposed by [16] give improvements over baseline FCN, indicating tumour segmentation benefits from encoding structure topology information. However, it is worse in all sub-tasks compared to **mFCN_core**, which confirms the efficacy of multi-task FCN framework. The validation performances of both mFCN_core and mFCN_edema models were saturated or even decreased around 30 epochs. Therefore, models trained at 30 epochs were used for benchmarking on test data.

We evaluated the proposed models on BRATS13 testing set. As it only contains HG images, we only use the 20 HG images for training. Both mFCN outperform FCN in most sub-tasks in terms of *Dice* and *Sensitivity*. This conclusion is consistent with the results on the validation set.

On BRATS13 testing set, our models are among the top-ranking in the state-of-the-arts (see Table 1). Specifically, Our models outperform all the best performers (Tustison, Meier and Reza) of BRATS13 challenge [1] and is also better than a semi-automatic method [9]. For CNN methods, our results are competitive to Pereira's [3] and better than Havaei's [4] while we report a 3 times faster

Table 1: Comparison with the state-of-the-arts on the testing set (ranked by VSD evaluation system [1])

Method	Dice			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
Pereira [3]	0.88	0.83	0.77	0.88	0.87	0.74	0.89	0.83	0.81
mFCN_core	0.88	0.83	0.75	0.86	0.85	0.70	0.91	0.85	0.83
mFCN_edema	0.88	0.82	0.76	0.85	0.82	0.72	0.92	0.85	0.82
FCN	0.87	0.82	0.75	0.85	0.87	0.72	0.89	0.79	0.80
Kwon [9]	0.88	0.83	0.72	0.92	0.90	0.74	0.84	0.78	0.72
Havaei [4]	0.88	0.79	0.73	0.89	0.79	0.68	0.87	0.79	0.80
Tustison [2]	0.87	0.78	0.74	0.85	0.74	0.69	0.89	0.88	0.83
Meier [1]	0.82	0.73	0.69	0.76	0.78	0.71	0.92	0.72	0.73
Reza [1]	0.83	0.72	0.72	0.82	0.81	0.70	0.86	0.69	0.76

Table 2: Performance on the BRATS15 44 testing set

Method	Dice			Positive Predictive Value			Sensitivity		
	Complete	Core	Enhancing	Complete	Core	Enhancing	Complete	Core	Enhancing
FCN	88.1	70.9	72.5	**92.2**	**82.7**	**79.7**	86.0	67.5	70.5
mFCN_edema	**88.5**	**71.0**	**73.1**	91.2	82.4	78.7	**87.5**	**67.9**	**71.4**

of average computational time (3 min) than the one (8 min) reported by Pereira's [3] due to the fast inference property of FCN. A direct comparison with 3D CNN [5] is not applicable as they did not report results on this dataset.

3.2 Results on BRATS15 dataset

We randomly split 220 HG images in BRATS15 training set into three subsets at a ratio of 6:2:2, resulting in 132 training, 44 validation and 44 test images. The performance curves along epochs are shown in the bottom row of Fig 4. We found **mFCN_edema** performs best on this dataset, especially for *Enhancing* task while slightly better for *Complete* task and comparable for *Core* task. However, **mFCN_core** performs worse compared to **mFCN_edema**. We attributes this inconsistency with BRATS13 to the relatively noise ground truth. It is worth noting that, in contrast to BRATS13 (where ground truth is the fusion of annotations from multiple radiologists), the ground truth of BRATS15 was produced by algorithms. Although it was verified by radiologists, it is less trustable compared to BRATS13, especially for the *Core* region, which has larger appearance variability.

On the 44 unseen test images, we trained the model by 25 epochs with both training and validation set. The results of baseline FCN and mFCN_edema are shown in Table 2. Consistent with the results on the validation set, The bmFCN_edema outperforms baseline FCN and in all sub-tasks in terms of *Dice* and *Sensitivity*, especially for *Enhancing* task.

4 Conclusion

In this paper, we introduced a multi-task FCN for brain tumor segmentation. Our approach formulates and jointly learns the *Complete*, *Core* and *Enhancing* of tumour segmentation tasks in a multi-task framework, which implicitly encode the topology information of subregions in tumour structure. This multi-task FCN achieved state-of-the-art results, and improved all sub-task segmentation performance on both BRATS13 and BRATS15 datasets compared to the single-task FCN. Our method is among the top ranked methods and has low computational cost.

References

1. Menze, Bjoern H., et al. The multimodal brain tumor image segmentation benchmark (BRATS). Medical Imaging 34.10 (2015): 1993-2024.
2. Tustison, Nicholas J., et al. Optimal symmetric multimodal templates and concatenated random forests for supervised brain tumor segmentation (simplified) with ANTsR. Neuroinformatics 13.2 (2015): 209-225.
3. Pereira, Srgio, et al. Brain tumor segmentation using convolutional neural networks in MRI images. Medical Imaging 35.5 (2016): 1240-1251.
4. Havaei, Mohammad, et al. Brain tumor segmentation with deep neural networks. Medical Image Analysis 35 (2017): 18-31.
5. Kamnitsas, Konstantinos, et al. Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Medical Image Analysis 36 (2017): 61-78.
6. Krhenbhl, P. and Koltun, V.: Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. In NIPS, pp. 109-117. 2011.
7. Simonyan, Karen, et al. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
8. Shen, Haocheng et al. Efficient Symmetry-Driven Fully Convolutional Network for Multimodal Brain Tumor Segmentation. Submitted to ICIP. 2017.
9. Kwon, Dongjin, et al. Combining generative models for multifocal glioma segmentation and registration. MICCAI. 2014.
10. Long, Jonathan, et al. Fully convolutional networks for semantic segmentation. CVPR. 2015.
11. Chen, Liang-Chieh, et al. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062 (2014).
12. Ronneberger, Olaf, et al. U-net: Convolutional networks for biomedical image segmentation. MICCAI, 2015.
13. Chen, Hao, et al. Deep contextual networks for neuronal structure segmentation. AAAI. 2016.
14. Chen, Hao, et al. Dcan: Deep contour-aware networks for accurate gland segmentation. CVPR. 2016.
15. He, Kaiming, et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. ICCV. 2015.
16. BenTaieb, Acha, et al. Topology Aware Fully Convolutional Networks for Histology Gland Segmentation. MICCAI. 2016.