Scalable multimodal convolutional networks for brain tumour segmentation

Lucas Fidon[1], Wenqi Li[1], Luis C. Garcia-Peraza-Herrera[1],
Jinendra Ekanayake[2,3], Neil Kitchen[2],
Sebastien Ourselin[1,3], and Tom Vercauteren[1,3]

[1] TIG, CMIC, University College London, London, UK
[2] NHNN, University College London Hospitals, London UK
[3] Wellcome/EPSRC Centre for Surgical and Interventional Science,
UCL, London, UK

Abstract. Brain tumour segmentation plays a key role in computer-assisted surgery. Deep neural networks have increased the accuracy of automatic segmentation significantly, however these models tend to generalise poorly to different imaging modalities than those for which they have been designed, thereby limiting their applications. For example, a network architecture initially designed for brain parcellation of monomodal T1 MRI can not be easily translated into an efficient tumour segmentation network that jointly utilises T1, T1c, Flair and T2 MRI. To tackle this, we propose a novel scalable multimodal deep learning architecture using new nested structures that explicitly leverage deep features within or across modalities. This aims at making the early layers of the architecture structured and sparse so that the final architecture becomes scalable to the number of modalities. We evaluate the scalable architecture for brain tumour segmentation and give evidence of its regularisation effect compared to the conventional concatenation approach.

1 Introduction

Gliomas make up 80% of all malignant brain tumours. Tumour-related tissue changes can be captured by various MR modalities, including T1, T1-contrast, T2, and Fluid Attenuation Inversion Recovery (FLAIR). Automatic segmentation of gliomas from MR images is an active field of research that promises to speed up diagnosis, surgery planning, and follow-up evaluations. Deep Convolutional Neural Networks (CNNs) have recently achieved state-of-the-art results on this task [1, 2, 6, 12]. Their success is partly attributed to their ability of automatically learning hierarchical visual features as opposed to conventional hand-crafted features extraction. Most of the existing multimodal network architectures handle imaging modalities by concatenating the intensities as an input. The multimodal information is implicitly fused by training the network discriminatively. Experiments show that relying on multiple MR modalities consistently is key to achieving highly accurate segmentations [3, 9]. However, using classical modality concatenation to turn a given monomodal architecture into a

multimodal CNN does not scale well because it either requires to dramatically augment the number of hidden channels and network parameters, or imposes a bottleneck on at least one of the network layers. This lack of scalability requires the design of dedicated multimodal architectures and makes it difficult and time-consuming to adapt state-of-the-art network architectures.

Recently, Havaei et al. [3] proposed an hetero-modal network architecture (HeMIS) that learns to embed the different modalities into a common latent space. Their work suggests that it is possible to impose more structure on the network. HeMIS separates the CNN into a backend that encodes modality-specific features up to the common latent space, and a frontend that uses high-level modality-agnostic feature abstractions. HeMIS is able to deal with missing modalities and shows promising segmentation results. However, the authors do not study the adaption of existing networks to additional imaging modalities and do not demonstrate an optimal fusion of information across modalities.

We propose a scalable network framework (ScaleNets) that enables efficient refinement of an existing architecture to adapt it to an arbitrary number of MR modalities instead of building a new architecture from scratch. ScaleNets are CNNs split into a backend and frontend with across-modality information flowing through the backend thereby alleviating the need for a one-shot latent space merging. The proposed scalable backend takes advantage of a factorisation of the feature space into imaging modalities (M-space) and modality-conditioned features (F-space). By explicitly using this factorisation, we impose sparsity on the network structure with demonstrated improved generalisation.

We evaluate our framework by starting from a high-resolution network initially designed for brain parcellation from T1 MRI [8] and readily adapting it to brain tumour segmentation from T1, T1c, Flair and T2 MRI. Finally, we explore the design of the modality-dependent backend by comparing several important factors, including the number of modality-dependent layers, the merging function, and convolutional kernel sizes. Our experiments show that the proposed networks are more efficient and scalable than the conventional CNNs and achieve competitive segmentation results on the BraTS 2013 challenge dataset.

2   Structural transformations across features/modalities

Concatenating multimodal images as input is the simplest and most common approach in CNN-based segmentation [2, 6]. We emphasise that the complete feature space FM can be factorised into a M-feature space M derived from imaging modalities, and a F-feature space F derived from scan intensity. However the concatenation strategy doesn't take advantage of it.

We propose to impose structural constraints that make this factorisation explicit. Let $V \subset \mathbb{R}^3$ be a discrete volume domain, and F (resp. M) be a finite F-features (resp. M-features) domain, the set of feature maps associated to (V, F, M) is defined as: $\mathcal{G}(V \times F \times M) = \{x : V \times F \times M \to \mathbb{R}\}$. This factorisation allows us to introduce new scalable layers that perform the transformation $\tilde{f}$ of the joint FM feature space in two steps (1). f (resp. g) typically uses convolutions
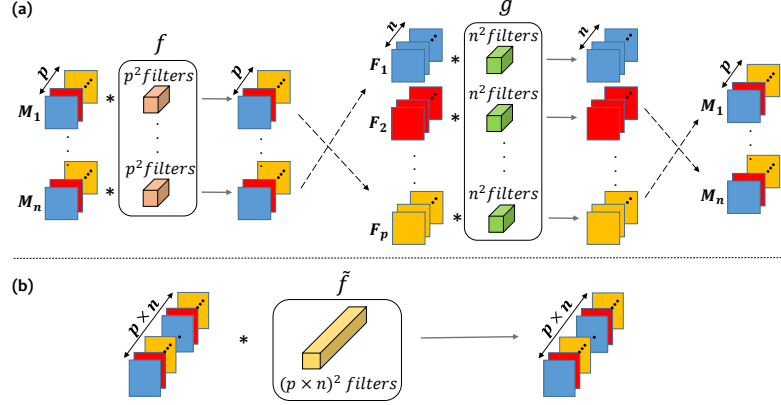
Fig. 1. (a) The proposed scalable multimodal layer. (b) A classic CNN layer with multimodal images concatenated as input. Volumes are represented as slices, the colours correspond to the F-features $(F_1, ..., F_p)$ and $(M_1, ..., M_n)$ correspond to the M-features. In (a) transformations across F-features f and across M-features g are explicitly separated (as illustrated by the rooted structure) while in (b) there are implicitly both applied in $\hat{f}$. The ratio of the number of parameters in (a) compared to (b) is $\frac{p+n}{p \times n}$.

across F-features (resp. across M-features). The proposed layer architecture, illustrated in Fig. 1, offers several advantages compared to classic ones: (1) cross F-feature layers remain to some extent independent of the number of modalities (2) cross M-feature layers allow the different modality branches to share complementary information (3) the total number of parameters is reduced. The HeMIS architecture [3], where one branch per modality is maintained until averaging merges the branches, is a special case of our framework where the cross M-features transformations g are identity mappings.

$$
\begin{array}{ccc}
\mathcal{G}(V \times F \times M) & \xdashrightarrow{\tilde{f}} & \mathcal{G}(V \times F' \times M') \\
& \searrow{\scriptstyle f} \qquad \nearrow{\scriptstyle g} & \\
& \mathcal{G}(V \times F' \times M) &
\end{array} \tag{1}
$$

Another important component of the proposed framework is the merging layer. It aims at recombining the F-features space and the M-features space together either by concatenating them or by applying a downsampling/pooling (averaging, maxout) on the M-features space to reduce its dimension to one:

$$
\mathcal{G}(V \times F \times M) \xrightarrow{concat} \mathcal{G}(V \times FM), \quad \mathcal{G}(V \times F \times M) \xrightarrow{pooling} \mathcal{G}(V \times F \times \{1\})
$$

As opposed to concatenation, relying on averaging or maxout for the merging layer at the interface between a backend and frontend makes the frontend structurally independent of the number of modalities and more generally of the entire backend. The proposed ScaleNets rely on such merging strategies to offer scalability in the network design.

3   ScaleNets implementation

The modularity of the proposed feature factorisation raises different questions: 1) Is the representative power of scalable F/M-structured multimodal CNN the same as classic ones? 2) What are the important parameters for the tradeoff between accuracy and complexity? 3) How can this modularity help readily transform existing architectures into scalable multimodal ones?

To demonstrate that our scalable framework can provide, to a deep network, the flexibility of efficiently being reused for different sets of image modalities, we adapt a model originally built for brain parcellation from T1 MRI [8]. As illustrated in Fig. 2, the proposed ScaleNets splits the network into two parts: (i) a backend and (ii) a frontend. In following experiments, we explore different backend architectures allowing to scale the monomodal network into a multimodal network. We also add a merging operation that allows plugging any backend into the frontend and makes the frontend independent from the number of modalities used. As a result, the frontend will be the same for all our architectures.

To readily adapt the backend from the monomodal network architecture [8] we duplicate the layers to get the across F-features transformations (one branch per M-features) and add an across M-features transformation after each of them (one branch per F-features) as shown in Fig. 2. In the frontend, only the number of outputs of the last layer is changed to match the number of classes for the new task. The proposed scalable models (SN31Ave1, SN31Ave2, SN31Ave3, SN33Ave2, SN31Max2) are named consistently. For example, SN31Ave2 stands for: "ScaleNet with 2 cross M-features residual blocks with 3^3 convolution and 1^3 convolution before averaging" and corresponds to the model (a) of Fig. 2.
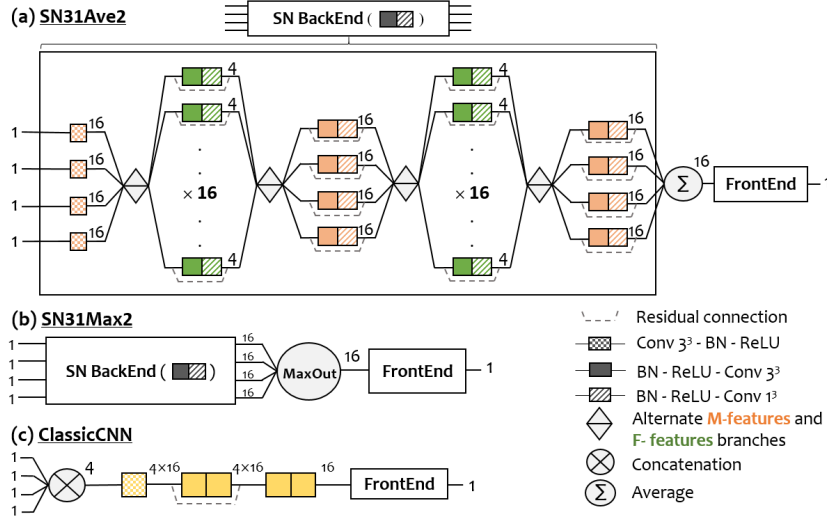

Fig. 2. Scalable and Classic CNN architectures. Numbers in bold are for the number of branches and the other correspond to the number of features.

Baseline monomodal architecture. The baseline architecture used for our experiments is a high-resolution, compact network designed for volumetric image segmentation [8]. It has proved to reach state-of-the-art results for brain parcellation of T1 scans. This fully convolutional neural network makes an end-to-end mapping from a monomodal image volume to a voxel-level segmentation map mainly with convolutional blocks and residual connections. It also takes advantage of dilated convolutions to incorporate image features at multiple scales while maintaining the spatial resolution of the input images. The maximum receptive field is $87 \times 87 \times 87$ voxels and is, therefore, able to catch multi-scale information in one path. By learning the variation between successive feature maps, the residual connections allow the initialisation of cross M-feature transformations closed to identity mappings. Thus it encourages information sharing across the modalities without changing their nature.

Brain tumour segmentation. We compare the different models on the task of brain tumour segmentation using BraTS'15 training set that is composed of 274 multimodal images (T1, T1c, T2 and Flair). We divide it into 80% for training, 10% for validation and 10% for testing. Additionally, we evaluate one of our scalable network model on the challenge BraTS'13 dataset, for which an online evaluation platform is available[4], to compare it to state-of-the-art (all the models were trained on the BraTS'15 though).

Implementation details. We maximise the soft Dice score as proposed by [10]. We train all the networks with Adam Optimization method [7] with a learning rate $lr = 0.01$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We also used early stopping on the validation set. Rotation of random small angles in the range $[-10°, 10°]$ are applied along each axis during training. All the scans of BraTS dataset are available after skull stripping, resampling to a 1mm isotropic grid and co-registration of all the modalities to the T1-weighted images for each patient. Additionally, we applied the histogram-based standardisation method [11]. The experiences have been performed using NiftyNet[5] and one GPU Nvidia GTX Titan.

Evaluation of segmentation performance. Results are evaluated using the Dice score of different tumour subparts: whole tumour, core tumour and enhanced tumour [9]. Additionally, we introduce a healthy tissue class to separate it from the background (zeroed out in the BraTS dataset).

4   Experiments and results

To demonstrate the usefulness of our framework, we compare two basic ScaleNets and a classic CNN. Tab. 1 highlights the benefits of ScaleNets in terms of number of parameters. We also explore some combinations of the important factors

---

[4] https://www.virtualskeleton.ch/BraTS/

[5] Our implementation of the ScaleNets and other CNNs used for comparison can be found at http://www.niftynet.io

Table 1. Comparison of ScaleNets and Classic concatenation-based CNNs for model adaptation on the testing set.

Method	# Param.	Mean(Std) Dice Score (%)		
		Whole Tumour	Core Tumour	Active Tumour
SN31Ave1	**0.83M**	87(8)	**73**(22)	**72(26)**
SN31Ave2	0.85M	87(7)	71(19)	70(28)
SN31Ave3	0.88M	**88(6)**	69(**17**)	71(27)
SN31Max2	0.85M	85(9)	67(**17**)	71(28)
SN33Ave2	0.92M	87(7)	70(18)	67(27)
HeMIS-like	0.89M	86(12)	70(20)	69(28)
Classic CNN	1.15M	81(18)	64(28)	65(28)

appearing in the choice of the architecture to try to address some key practical questions. How deep does the cross modalities layers have to be? When should we merge the different branches? Which merging operation should we use? Wilcoxon signed-rank p-values are reported to highlight significant improvements.

ScaleNet with basic merging and classic CNN. We compare three merging strategies (averaging: "SN31Ave2", maxout: "SN31Max2" and concatenation: "Classic CNN"). To be as fair as possible, we carefully choose the size of the kernels so that the maximum receptive field remain the same across all architectures. Quantitative Dice score results Tab. 1 show that both SN31Ave2 and SN31Max2 outperform Classic CNN on the segmentation of all tumour region. SN31Ave2 outperforms SN31Max2 for core tumour and get similar results on whole tumour and enhanced tumour.

We compare ScaleNets with resp. 1, 2 or 3 scalable multimodal layers before averaging (resp. named "SN31Ave1", "SN31Ave2", "SN31Ave3"). The results reported on Tab. 1 show similar performance for all of those models. This suggests that a short backend is enough to get a modality-agnostic sufficient repre-
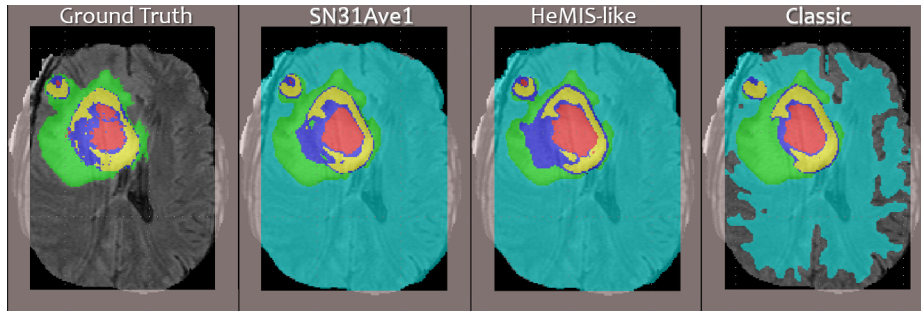

Fig. 3. Qualitative comparison of different models output on a particular testing case. Colours correspond to different tissue regions. red: necrotic core, yellow: enhancing tumour, blue: non-enhancing tumour, green: edema, cyan: healthy tissues.

Table 2. Dice score on Leaderboard and Challenge against BraTS'13 winners.

Method	Leaderboard			Challenge		
	Whole	Core	Enhanced	Whole	Core	Enhanced
Tustison	**79**	**65**	53	87	**78**	**74**
Zaho	**79**	59	47	84	70	65
Meier	72	60	53	82	73	69
SN31Ave1	77	64	**56**	**88**	77	72

sentation for Gliomas segmentation using T1, T1c, FLAIR and T2. Furthermore, SN31Ave1 outperforms Classic CNN on all tumour regions ($p \leq 0.001$).

Qualitative results in a testing case with artifact deformation (Fig. 3) and the decreasing of Dice score standard deviation for whole and core tumour (Tab. 1) demonstrate the robustness of ScaleNets compared to classic CNNs and show the regularisation effect of the proposed scalable multimodal layers Fig. 1.

Comparison to state-of-the-art. We validate the usefulness of the cross M-feature layers by comparing our proposed network to an implementation of ScaleNets aiming at replicating the characteristics of the HeMIS network [3] by removing the cross M-feature layers. We refer to this latest network as HeMIS-like. Dice score results in Tab. 1 illustrate improved results on the core tumour ($p \leq 0.03$) and similar performance on whole and active tumour. Qualitative comparison in Fig. 3 clearly confirmed this trend.

We compare our SN31Ave1 model to the state-of-the-art. The results obtained on Leaderboard and Challenge BraTS'13 dataset are reported in Tab. 2 and compared to the BraTS'13 Challenge Winners listed in [9]. We achieved similar results with no need of post-processing.

5   Conclusions

We have proposed a scalable deep learning framework that allows building more reusable and efficient deep models when multiple correlated sources are available. In the case of volumetric multimodal MRI for brain tumour segmentation, we proposed several scalable CNNs that integrate smoothly the complementary information about tumour tissues scattered across the different image modalities. ScaleNets impose a sparse structure to the backend of the architecture where cross features and cross modalities transformations are separated. It is worth noticing that ScaleNets are related to the recently proposed implicit Conditional Networks [5] and Deep Rooted Networks [4] that use sparsely connected architecture but do not suggest the transposition of branches and grouped features. Both of these frameworks have been shown to improve the computational efficiency of state-of-the-art CNNs by reducing the number of parameters, the amount of computation and increasing the parallelisation of the convolutions.

Using our proposed scalable layer architecture, we readily adapted a compact network for brain parcellation of monomodal T1 into a multimodal network for

brain tumour segmentation with 4 different image modalities as input. Scalable structures, thanks to their sparsity, have a regularisation effect. Comparison of classic and scalable CNNs shows that scalable networks are more robust and use fewer parameters while maintaining similar or better accuracy for medical image segmentation. Scalable network structures have the potential to make deep network for medical images more reusable. We believe that scalable networks will play a key enabling role for efficient transfer learning in volumetric MRI analysis.

References

1. Chen, H., Dou, Q., Yu, L., Qin, J., Heng, P.A.: Voxresnet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage (2017), http://www.sciencedirect.com/science/article/pii/S1053811917303348
2. Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.M., Larochelle, H.: Brain tumor segmentation with deep neural networks. Med. Image Anal. 35, 18–31 (2017)
3. Havaei, M., Guizard, N., Chapados, N., Bengio, Y.: HeMIS: Hetero-modal image segmentation. In: Proc. MICCAI'16. pp. 469–477. Springer (2016)
4. Ioannou, Y., Robertson, D., Cipolla, R., Criminisi, A.: Deep roots: Improving CNN efficiency with hierarchical filter groups. CVPR'17 (2017)
5. Ioannou, Y., Robertson, D., Zikic, D., Kontschieder, P., Shotton, J., Brown, M., Criminisi, A.: Decision forests, convolutional networks and the models in-between. arXiv:1603.01250 (2016)
6. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. Med. Image Anal. 36, 61–78 (2017)
7. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
8. Li, W., Wang, G., Fidon, L., Ourselin, S., Cardoso, M.J., Vercauteren, T.: On the compactness, efficiency, and representation of 3D convolutional networks: Brain parcellation as a pretext task. Proc. IPMI'17 (2017)
9. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BraTS). IEEE Trans. Med. Imag. 34(10), 1993–2024 (2015)
10. Milletari, F., Navab, N., Ahmadi, S.A.: V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proc. 3DV'16. pp. 565–571 (2016)
11. Nyul, L.G., Udupa, J.K., Zhang, X.: New variants of a method of mri scale standardization. IEEE Transactions on Medical Imaging 19(2), 143–150 (Feb 2000)
12. Pereira, S., Pinto, A., Alves, V., Silva, C.A.: Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans. Med. Imag. 35(5), 1240–1251 (2016)