Adult brain tissue segmentation in multi-modal images with a HyperDenseNet

Jose Dolz, Ismail Ben Ayed and Christian Desrosiers

Abstract

Accurate segmentation of brain tissue serves as the first step towards volumetric studies and quantitative analysis, being highly beneficial for diagnosis, treatment and follow up of many neurological diseases. Manual segmentation is the conventional approach in clinical routine, which is time-consuming and prone to inter- and intra- observer variability. Despite all the efforts devoted to the automation of this task, automatic precise segmentation still remains challenging. We propose a fully convolutional neural network with hyper dense connectivity, called *HyperDenseNet*, to tackle this problem. One of the main advantages of hyper-dense connections lies on its inherent ability to perform deep supervision and aiding the gradient flow during training due to the long connections. Further, by combining the feature maps of intermediate convolutional layers the architecture also injects multi-scale information into the final segmentation. To validate our network we have employed the MRBrainS dataset for brain segmentation on 3D magnetic resonance (MR) images. We compared the proposed method to well-known state-of-the-art networks outperforming them in the task at hand. Additionally, we have tested the performance of the proposed network with 2, and with 3 modalities to investigate the effect of having more independent streams on the network.

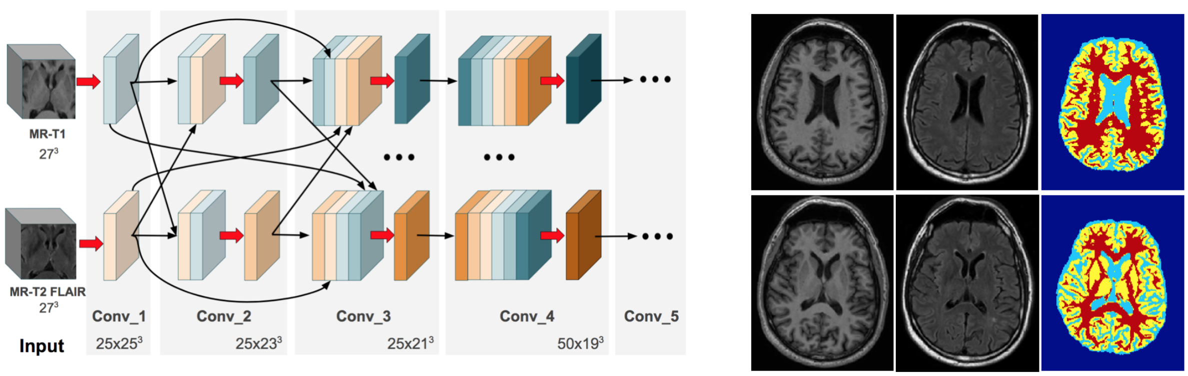

Figure 1: A section of the proposed HyperDenseNet. Each gray region represents a convolutional block. Red arrows correspond to convolutions and black arrows indicate dense connections between feature maps.

1 Method details

Inspired by the recent success of such densely-connected networks in medical image segmentation we propose a hyper-densely connected architecture, called HyperDenseNet, for the segmentation of multi-modal images. The presented architecture extends our recent work in [1], where we proposed a 3D fully CNN to segment subcortical brain structures. Particularly, we exploit the concept of dense connectivity in a multi-modal image setting. In this scenario, each modality is processed in independent paths, and dense connections occur not only between layers within the same path, but also between layers in different paths.

Let $\vec{x}_l$ be the output of the l^{th} layer. In CNNs, this vector is typically obtained from the output of the previous layer $\vec{x}_{l-1}$ by a mapping H_l composed of a convolution followed by a non-linear activation function:

$$\vec{x}_l \;=\; H_l(\vec{x}_{l-1}). \tag{1}$$

In a densely-connected network, connectivity follows a pattern that iteratively concatenates all feature outputs in a feed-forward manner, i.e.

$$\vec{x}_l \;=\; H_l([\vec{x}_{l-1}, \vec{x}_{l-2}, \ldots, \vec{x}_0]), \tag{2}$$

where [...] represents a concatenation operation. Pushing this idea further, *HyperDenseNet* considers a more sophisticated connectivity pattern that also links the output from layers in different streams, each one associated with a different image modality. Let consider a two image modalities scenario for simplicity, although extension to N

modalities is straightforward. Denote as $\vec{x}_l^1$ and $\vec{x}_l^2$ the outputs of the l^{th} layer in streams 1 and 2, respectively. The output of the l^{th} layer in a stream s can then be defined as

$$\vec{x}_l^s \;=\; H_l([\vec{x}_{l-1}^1, \vec{x}_{l-1}^2, \vec{x}_{l-2}^1, \vec{x}_{l-2}^2, \ldots, \vec{x}_0^1, \vec{x}_0^2]). \tag{3}$$

A section of the proposed architecture is depicted in Figure 1, where each gray region represents a convolutional block. For simplicity, we assume that red arrows indicate convolution operations only, and that black arrows represent direct connections between feature maps from different layers. Thus, the input of each convolutional block (maps before the red arrow) consists in the concatenation of the outputs (maps after red arrow) of all preceding layers from both paths. This version differs from the previous one in that the network is modified to accommodate three image modalities as input of the network, each of them processed independently in separated paths.

2 Experiments and results

Training and network parameters. Our CNN is composed of 13 layer levels in total: 9 convolutional layers in each path, 3 fully-connected layers, and the classification layer. The number of kernels in each convolutional layer, from shallow to deeper, is as follows: 25, 25, 25, 50, 50, 50, 75, 75 and 75, sizes of which are equal to $3\times3\times3$. The fully-connected layers are composed of 400, 200 and 150 hidden units, respectively, followed by a final classification layer. Momentum was set to 0.6 and the initial learning rate to 0.001, being reduced by a factor of 2 after every 5 epochs (starting from epoch 10). Weights in layer l were initialized based on a zero-mean Gaussian distribution of standard deviation $2/n_l$, where n_l denotes the number of connections to units in that layer. Patch size employed as input sample was equal to $27\times27\times27$ for training and $35\times35\times35$ for inference. Our 3D FCNN was trained for 15 epochs, each one composed of 20 subepochs. At each subepoch, a total of 1000 samples were randomly selected from the training images, and processed in batches of size 20. Training and testing were performed in a server equipped with a NVIDIA Tesla P100 GPU with 16 GB of RAM memory. Training our network took around 3 hours per epoch, and around 1 day for having a model trained. Segmentation of a whole 3D MR scan is performed in 4 minutes, as average. To train the networks, the provided labels were merged into gray matter, white matter and cerebrospinal fluid classes.

	Mean DSC		
	CSF	GM	WM
3D-ResNet [3, 4]	0.7710	0.8211	0.8693
3D-UNet [2]	0.8255	0.8496	0.8963
HyperDenseNet (T1-FLAIR)	0.8259	0.8620	0.8982
HyperDenseNet (T1-T1_IR)	0.7991	0.8226	0.8654
HyperDenseNet (T1_IR-FLAIR)	0.8094	0.8462	0.8901
HyperDenseNet (3-Modalities)	**0.8485**	**0.8658**	**0.9004**

Table 1: Comparison on three training subjects to different state of the art 3D networks. To evaluate performance in one subject, the 4 remaining subjects were employed for training.

Results. We first performed a cross-validation on the training set to compare the proposed network to some state-of-the-art deep architectures in medical segmentation: 3D UNet [2] and a 3D residual fully convolutional neural network based on [3, 4]. We employed the implementation of these networks provided in [5]. To perform this comparison 4 subjects were employed in training and one for validation. This process was repeated for 3 different subjects and an average was estimated. Table 2 depicts the results from this comparison, which shows that the proposed network outperforms these well-known networks. A noteworthy point is that both 3D-ResNet and 3D-UNet were trained employing an augmented dataset, generated by flipping the images along the sagittal axis and adding Gaussian noise with $\sigma = 0.2$ and offset equal to 0.5. On the other hand, no data augmentation was used to train the proposed *HyperDenseNet*. It can be also observed than by employing all three modalities, segmentation results are further improved with respect to the two image modality versions. A qualitative example of the segmentation result is depicted in Fig. 1, right.

References

[1] Dolz, J., et al. "3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study." NeuroImage (2017).

[2] Çiçek, Ö., et al. "3D U-Net: learning dense volumetric segmentation from sparse annotation." MICCAI. Springer International Publishing, 2016.

[3] Long, J et al. "Fully convolutional networks for semantic segmentation". CVPR 2015.

[4] He,K. et al. "Identity Mappings in Deep Residual Networks". ECCV 2016.

[5] Pawlowski, N., et al. "DLTK: State of the Art Reference Implementations for Deep Learning on Medical Images." arXiv preprint arXiv:1711.06853 (2017).