Automatic Brain Structures Segmentation Using Deep Residual Dilated U-Net

Hongwei Li(✉), Andrii Zhygallo, and Bjoern Menze

Technical University of Munich, Munich, Germany
{hongwei.li,andrii.zhygallo,bjoern.menze}@tum.de

Abstract. Brain image segmentation is used for visualizing and quantifying anatomical structures of the brain. We present an automated approach using 2D deep residual dilated networks which captures rich context information of different tissues for the segmentation of eight brain structures. The proposed system was evaluated in the MICCAI Brain Segmentation Challenge (http://mrbrains18.isi.uu.nl/) and ranked 9^{th} out of 22 teams. We further compared the method with traditional U-Net using leave-one-subject-out cross-validation setting on the public dataset. Experimental results shows that the proposed method outperforms traditional U-Net (i.e. 80.9% *vs* 78.3% in averaged Dice score, 4.35 mm *vs* 11.59 mm in averaged robust Hausdorff distance) and is computationally efficient.

Keywords: Brain structure segmentation · Deep learning

1 Introduction

Brain MRI segmentation is an important task in many clinical applications. Various approaches for brain analysis rely on accurate segmentation of anatomical regions. For example, it is commonly used for measuring and visualizing different brain structures, for delineating lesions, for analysing brain development, and for characterization of brain disorders such as Alzheimers disease, epilepsy, schizophrenia, multiple sclerosis (MS), cancer, and infectious and degenerative diseases. Manual segmentation is the gold standard for in-vivo images. However, it requires outlining structures slice-by-slice by neuroradiologist, which is highly time-consuming and prone to rater-bias. Therefore, there is a need for automated segmentation approaches to provide accuracy close to that of expert raters with a high reproducibility.

Early works on segmentation of normal brain structures focus on white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF), which is important for studying early brain developments in infants and quantitative assessment of the brain tissue and intracranial volume in large scale studies. Atlas-based approaches [7,12], which match intensity information between an atlas and target images and pattern recognition approaches [10], which classify tissues based on a set of local intensity features, are the classical approaches that have been

A. Crimi et al. (Eds.): BrainLes 2018, LNCS 11383, pp. 385–393, 2019.
https://doi.org/10.1007/978-3-030-11723-8_39

used for brain tissue segmentation. The MRBrainS Challenge 2013 [8] was held to compare state-of-the-art segmentation algorithms on three brain structures in conjunction with the 16^{th} International Conference on Medical Image Computing and Computer Assisted Intervention. Deep-learning based approaches have shown superior performances to the traditional state-of-art methods on the segmentation of brain stroke lesions, brain white matter lesions and brain tumors [5,6,9].

In this paper, we presented a deep-learning based method for segmenting eight brain tissues including cortical gray matter (GM), basal ganglia, WM, white matter lesions/hyperintensities (WMH), CSF, ventricles, cerebellum and brain stem. Deep dilated residual U-Net was adopted to learn context and texture information of different brain tissues. Multi-sequence data including T1, T1-IR and FLAIR which captures complementary information of different brain structures. The proposed 2-D network was more computationally efficient than 3D network and traditional U-Net. Experimental results showed that the proposed method outperforms traditional U-Net.

2 Materials

2.1 Dataset and Protocols

Dataset. Thirty MRI scans were acquired on a 3.0T Philips Achieva MR scanner at the University Medical Center Utrecht (Netherlands). The following sequences were acquired and used for the evaluation framework: 3D T1 (TR: 7.9 ms, TE: 4.5 ms), T1-IR (TR: 4416 ms, TE: 15 ms, and TI: 400 ms), and T2-FLAIR (TR: 11000 ms, TE: 125 ms, and TI: 2800 ms). The sequences were aligned by rigid registration using Elastix [3] and bias correction was performed using SPM8. After registration, the voxel size within all provided sequences (T1, T1-IR, and T2-FLAIR) was $0.96 \times 0.96 \times 3.00\,\text{mm}^3$. Seven scans with annotations were released as a public training set, and the remaining twenty-three scans were used as hidden testing set. For more details on the method of ranking performance, please find the relevant information on the challenge website.

Evaluation Metric. Three types of measures were employed to evaluate the segmentation results. The Dice coefficient is used to determine the spatial overlap and is defined as:

$$Dice = \frac{2|G \cap P|}{|G| + |P|} \tag{1}$$

where G is the reference standard, P is the segmentation result.

The 95th-percentile of the Hausdorff distance is used to determine the distance between the segmentation boundaries. Hausdorff distance is defined as:

$$H(G,P) = max\{\sup_{x \in G} \inf_{y \in P} d(x,y), \sup_{y \in P} \inf_{x \in G} d(x,y)\} \tag{2}$$

where *d(x, y)* denotes the distance of x and y, *sup* denotes the supremum and *inf* for the infimum.

The third measure is the volumetric similarity. Let V_G and V_P be the volume of lesion regions in G and P respectively. Then the volumetric similarity (VS) in percentage is defined as:

$$VS = \frac{|V_G - V_P|}{V_G} \tag{3}$$

3 Methodology

3.1 Image Preprocessing

A patient-wise normalization of the image intensities was performed both during training and testing. For the scan of each patient, the mean value and standard deviation were calculated based on intensities of all voxels. Then each image volume was normalized to zero mean and unit standard deviation. Rotation, shearing, scaling along horizontal direction (x-scaling), and scaling along vertical direction (y-scaling) were employed for data augmentation. After data augmentation, a four times larger training dataset was obtained.

3.2 2D Dilated Residual U-Net

We used Dilated Residual U-Net (DRUNet), which was originally proposed in [1] for nerve head tissues segmentation in optical coherence tomography images. DRUNet exploits the inherent advantages of the U-Net skip connections [11], residual learning [2] and dilated convolutions [13] to capture rich context information and offer a robust brain structure segmentation with a minimal number of trainable parameters.

DRUNet architecture is presented in Fig. 1. The model consists of downsampling and upsampling parts. In turn, each part includes one standard block and two residual blocks. Corresponding blocks in downsampling and upsampling parts are connected through skip connections. Convolution layers in both block types have 32 filters of size 3 × 3. In total the entire network consists of 156,105 trainable parameters.

3.3 Combination of Modalities

Multi-sequence data including T1-weighted (T1), T1-weighted inversion recovery (T1-IR) and FLAIR which captures complementary information of different brain structures were used for training the network. In clinical practice, the combination of FLAIR and T1 is beneficial for segmenting white matter lesions while the combination of T1 and T1-IR is helpful for annotating cerebrospinal fluid. We feed different combinations of modalities for multiple networks.

3.4 Ensemble Model

To improve the robustness of our model, an ensemble method was used in the testing stage. Then when given a new testing subject, each subject will be segmented based on the averaged probability maps by the ensemble model.

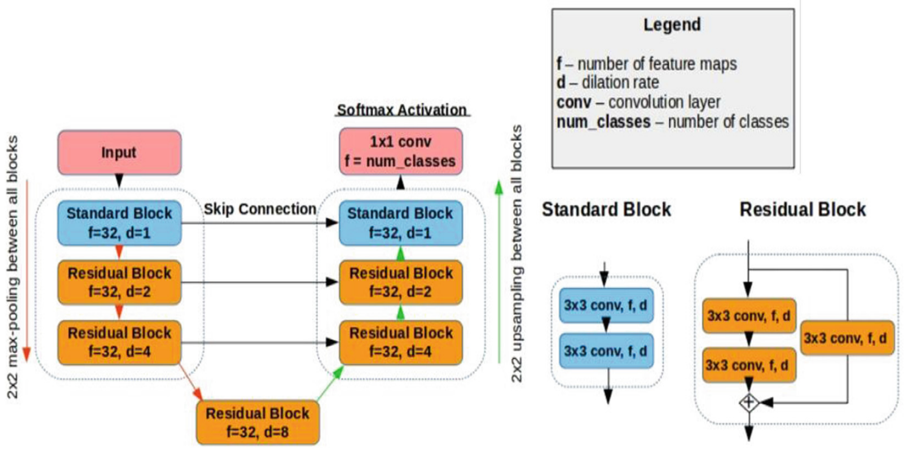


Fig. 1. Details of DRUnet architecture which contains residual blocks with dilated convolutions.

3.5 Our Submissions

Submission 1. We used only DRUNet for simultaneously segmenting the ten labels including infarction and pathologies were set to background label during the training of the network. We generated five DRUNet models with the same architecture but trained with shuffled batches. Then in testing stage, each subject was segmented based on the averaged probability maps by the ensemble models.

Submission 2. We used two Dilated Residual U-Nets (DRUNet) and one traditional U-Net for segmenting different labels. Since not all the labels were annotated in the same modalities, i.e., white matter lesions were annotated on the FLAIR scan and the outer border of the CSF was segmented using both the T1-weighted scan and the T1-weighted inversion recovery scan, we employed a multi-stage approach to segment different tissues from coarse to fine using different combinations of input modalities. Firstly, coarse segmentation including eight brain tissues (other labels including infarction and pathologies were set to background label) was performed using FLAIR and T1-weighted modalities by DRUNet (model 1). Secondly, CSF was independently segmented using T1 and T1-IR modalities by DRUNet (model 2). Thirdly, since segmentation of white matter lesions is a very challenging task, we used the pre-trained model of the winning method in MICCAI WMH challenge [4] (model 3) to perform segmentation independently. Finally we fused the multi-stage segmentation results. Five DRUNet models for model 1 and model 2, respectively, with the same architecture were trained with shuffled batches.

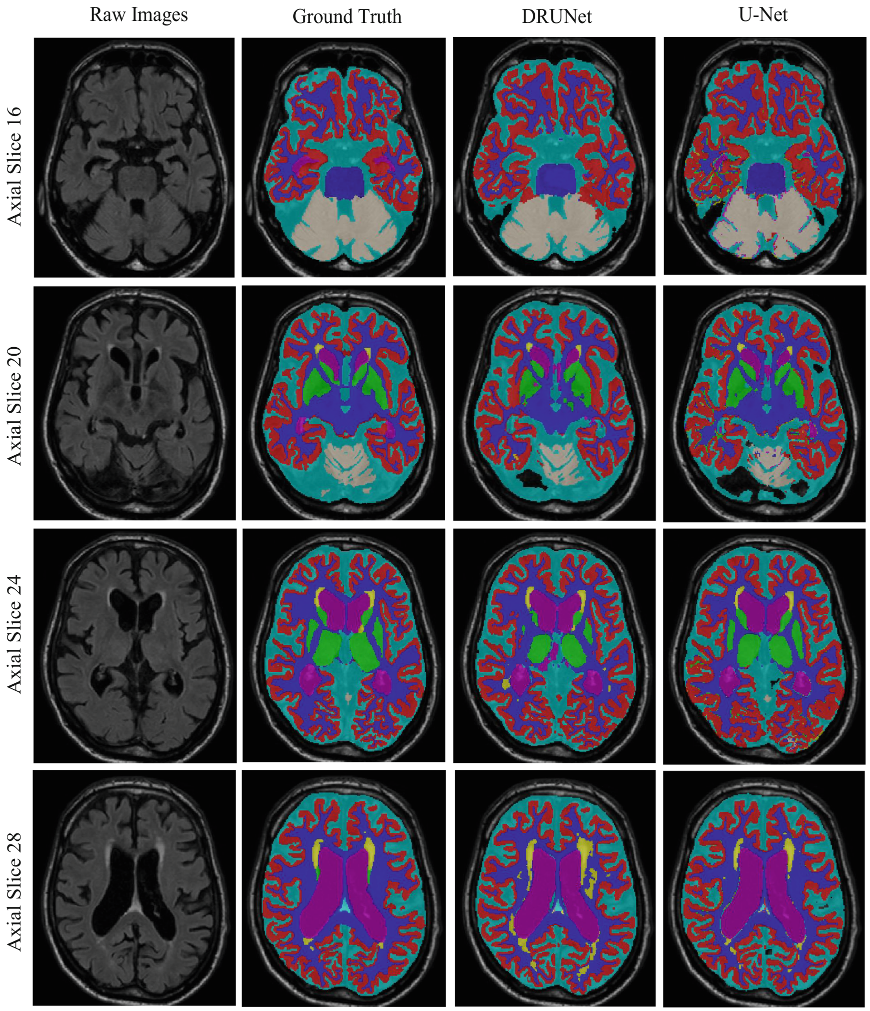


Fig. 2. Sample segmentation result on *Case 70*. From top to bottom: four axial slices of the same scan. From left to right: FLAIR MR images, the associated ground truth, segmentation result using DRUNet and segmentation result using U-Net. (Best viewed in colour). We can observed from the segmentation result of axial slice 16 that DRUNet achieved better performance on large continuous regions while U-Net generated some isolated false positives. It indicates that the dilated convolution in DRUNet helps to capture context information. On the other hand, for the segmentation of small tissues such as WMHs, DRUNet seems to generate more false positives than U-Net as observed from axial slice 28.

Table 1. Leave-one-subject-out evaluation of our submissions on the public training set containing seven subjects. The averaged Dice score, averaged H95, averaged volume similarity of eight tissues for each subject were shown in the table. The left and right values in each cell were the results of submission 1 and submission 2 respectively. The values in bold indicates the subject on which the two submissions has significant segmentation difference.

Metrics	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5	Subject 6	Subject 7
Dice	0.86/0.85	0.82/0.82	0.77/0.77	**0.73/0.77**	0.85/0.84	0.80/0.80	0.83/0.81
H95	2.98/2.43	3.15/2.33	6.07/6.56	**8.25/5.87**	3.42/2.39	4.17/6.58	2.49/8.07
VS	0.97/0.98	0.92/0.91	0.86/0.87	**0.82/0.88**	0.94/0.92	0.88/0.89	0.92/0.90

4 Results

4.1 Leave-One-Subject-Out Evaluation

To test the generalization performance of our systems across different subjects, we conducted an experiment on the public training datasets (seven subjects) in a leave-one-subject-out setting. Specifically, we used the subject IDs to split the public training dataset into training and validation sets. In each split, we used slices from six subjects for training, and the slices from the remaining subject for testing. This procedure was repeated until all of the subjects are used as testing. The results were shown in Table 1. There exists significant segmentation difference on subject 4. We further observed the brain structures of subject 4 and found it was a heathy brain scan without WMHs, infarctions and other lesions. The reason for the performance difference could be that the models in first submission were trained on 10 labels including infarctions and other lesions while the models in the second submission were trained on 8 main structures excluding two other labels. When testing on healthy scans, the models trained with 8 main healthy tissues could be more effective since the data distributions among training and testing were similar.

4.2 Comparison with U-Net

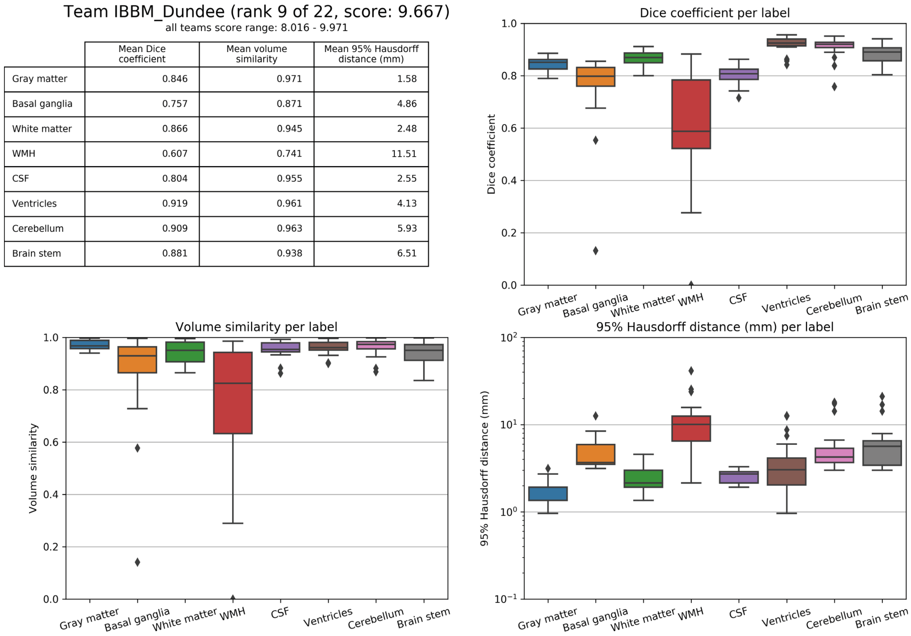
We further compared the performance of the proposed method (submission 1) with traditional U-Net using the state-of-the-art architecture proposed in [4]. As shown in Table 2, generally our approach outperformed traditional U-Net, especially in segmentation of WM and CSF, with an improvement of 8% and 11% in Dice score. WM and CSF are both large structures in brains. We concluded that the use of dilated convolutions is beneficial for capturing the context information of large target. Furthermore, our model is with much fewer trainable parameters (156,105 *vs* 8,748,609). Thus the training of the network is computationally efficient. The segmentation results of both DRUNet and U-Net on test *case 70* was shown in Fig. 2.

Table 2. Comparison on each class with traditional U-Net under leave-one-subject-out setting. The performance on each class was averaged over seven subjects. The values in bold indicated significant improvement over traditional U-Net.

Metrics	GM	BG	WM	WMH	CSF	Ventricles	Cerebellum	Brain stem	*Averaged*
Dice_{U-Net}	0.83	0.84	0.70	0.79	0.43	0.89	0.9	0.88	*0.783*
Dice_{DRUNet}	0.84	0.85	**0.78**	**0.81**	**0.54**	0.88	**0.92**	0.85	***0.809***
H95_{U-Net}	1.26	1.8	43.5	1.78	23.09	3	15.63	2.67	*11.59*
H95_{DRUNet}	1.29	1.67	**5.82**	1.61	**14.8**	3.15	**2.97**	3.45	***4.35***
VS_{U-Net}	0.95	0.95	0.84	0.93	**0.71**	0.94	0.96	0.94	*90.25*
VS_{DRUNet}	0.96	0.94	**0.89**	0.94	0.66	0.93	0.97	0.92	*90.13*

4.3 Results on Hidden Testing Cases

Our submissions were independently evaluated by the challenge organizer. Figures 3 and 4 show the box plots of performance on eight labels on 23 testing scans. Submission 1 and submission 2 ranked 9th and 12th respectively out of 22 teams.

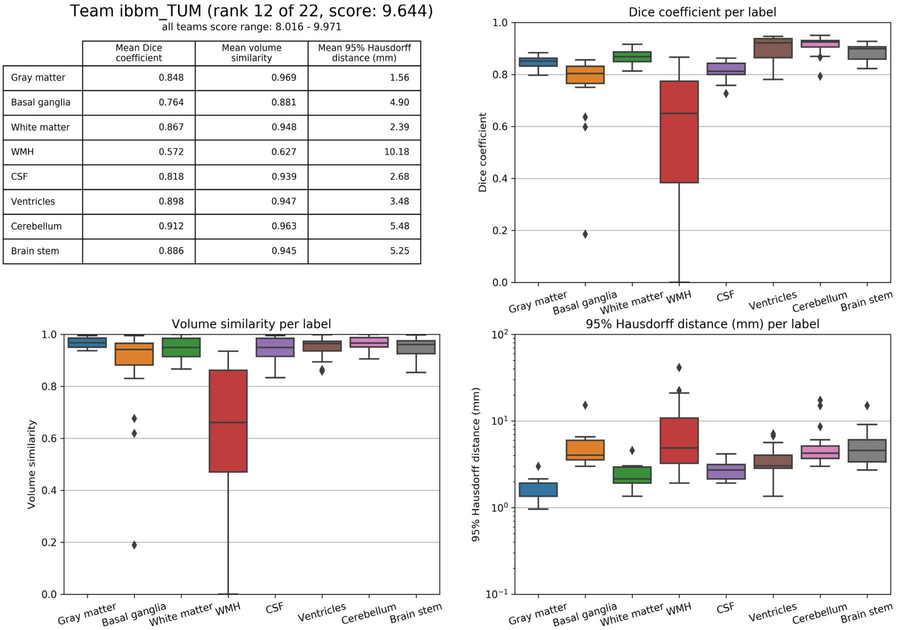


Team IBBM_Dundee (rank 9 of 22, score: 9.667)

all teams score range: 8.016 - 9.971

	Mean Dice coefficient	Mean volume similarity	Mean 95% Hausdorff distance (mm)
Gray matter	0.846	0.971	1.58
Basal ganglia	0.757	0.871	4.86
White matter	0.866	0.945	2.48
WMH	0.607	0.741	11.51
CSF	0.804	0.955	2.55
Ventricles	0.919	0.961	4.13
Cerebellum	0.909	0.963	5.93
Brain stem	0.881	0.938	6.51

Fig. 3. Result of our first submission on the 23 hidden testing set evaluated by the challenge organizers. Our method achieved Dice scores of more than 80% and volume similarity of more 90% on the major classes while the segmentation performance on WMHs is relatively poor. This is because the WMHs are in small volumes and thus the most difficult structure to be segmented.



Team ibbm_TUM (rank 12 of 22, score: 9.644)
all teams score range: 8.016 - 9.971

	Mean Dice coefficient	Mean volume similarity	Mean 95% Hausdorff distance (mm)
Gray matter	0.848	0.969	1.56
Basal ganglia	0.764	0.881	4.90
White matter	0.867	0.948	2.39
WMH	0.572	0.627	10.18
CSF	0.818	0.939	2.68
Ventricles	0.898	0.947	3.48
Cerebellum	0.912	0.963	5.48
Brain stem	0.886	0.945	5.25

Fig. 4. Result of our second submission on the 23 hidden testing set evaluated by the challenge organizers. The two submissions achieved comparable performance on major classes except the WMHs. Actually the second submission was designed to improve the segmentation performance of WMHs and integrated the state-of-the-art models from [4]. There may exist some implementation mistakes in the label fusion stage.

References

1. Devalla, S.K., et al.: DRUNET: a dilated-residual u-net deep learning network to digitally stain optic nerve head tissues in optical coherence tomography images. arXiv preprint arXiv:1803.00232 (2018)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Klein, S., Staring, M., Murphy, K., Viergever, M.A., Pluim, J.P.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
4. Li, H., et al.: Fully convolutional network ensembles for white matter hyperintensities segmentation in MR images. arXiv preprint arXiv:1802.05203 (2018)
5. Li, H., Zhang, J., Muehlau, M., Kirschke, J., Menze, B.: Multi-scale convolutional-stack aggregation for robust white matter hyperintensities segmentation. arXiv preprint arXiv:1807.05153 (2018)
6. Maier, O., et al.: ISLES 2015-a public evaluation benchmark for ischemic stroke lesion segmentation from multispectral MRI. Med. Image Anal. **35**, 250–269 (2017)
7. Makropoulos, A., et al.: Automatic whole brain MRI segmentation of the developing neonatal brain. IEEE Trans. Med. Imaging **33**(9), 1818–1831 (2014)

8. Mendrik, A.M., et al.: MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput. Intell. Neurosci. **2015**, 1 (2015)
9. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (brats). IEEE Trans. Med. Imaging **34**(10), 1993 (2015)
10. Moeskops, P., et al.: Automatic segmentation of MR brain images of preterm infants using supervised classification. NeuroImage **118**, 628–641 (2015)
11. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28. http://lmb.informatik.uni-freiburg.de/Publications/2015/RFB15a. arXiv:1505.04597
12. Vrooman, H.A., et al.: Multi-spectral brain tissue segmentation using automatically trained k-nearest-neighbor classification. Neuroimage **37**(1), 71–81 (2007)
13. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint arXiv:1511.07122 (2015)