Student beats the teacher: deep neural networks for lateral ventricles segmentation in brain MR

Mohsen Ghafoorian, Jonas Teuwen, Rashindra Manniesing, Frank-Erik de Leeuw, Bram van Ginneken, et al.

Mohsen Ghafoorian, Jonas Teuwen, Rashindra Manniesing, Frank-Erik de Leeuw, Bram van Ginneken, Nico Karssemeijer, Bram Platel, "Student beats the teacher: deep neural networks for lateral ventricles segmentation in brain MR," Proc. SPIE 10574, Medical Imaging 2018: Image Processing, 105742U (2 March 2018); doi: 10.1117/12.2293569

SPIE.

Event: SPIE Medical Imaging, 2018, Houston, Texas, United States

Student Beats the Teacher: Deep Neural Networks for Lateral Ventricles Segmentation in Brain MR

Mohsen Ghafoorian[a,c,*], Jonas Teuwen[a,d,*], Rashindra Manniesing[a], Frank-Erik de Leeuw[b], Bram van Ginneken[a], Nico Karssemeijer[a], and Bram Platel[a]

[a]Radboud University Medical Center, Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Nijmegen, the Netherlands
[b]Donders Institute for Brain, Cognition and Behaviour, Department of Neurology, Radboud University Medical Center, Nijmegen, the Netherlands
[c]TomTom, Amsterdam, the Netherlands
[d]Optics Research Group, Imaging Physics Department, Delft University of Technology, the Netherlands

ABSTRACT

Ventricular volume and its progression are known to be linked to several brain diseases such as dementia and schizophrenia. Therefore accurate measurement of ventricle volume is vital for longitudinal studies on these disorders, making automated ventricle segmentation algorithms desirable. In the past few years, deep neural networks have shown to outperform the classical models in many imaging domains. However, the success of deep networks is dependent on manually labeled data sets, which are expensive to acquire especially for higher dimensional data in the medical domain. In this work, we show that deep neural networks can be trained on much-cheaper-to-acquire pseudo-labels (e.g., generated by other automated less accurate methods) and still produce more accurate segmentations compared to the quality of the labels. To show this, we use noisy segmentation labels generated by a conventional region growing algorithm to train a deep network for lateral ventricle segmentation. Then on a large manually annotated test set, we show that the network significantly outperforms the conventional region growing algorithm which was used to produce the training labels for the network. Our experiments report a Dice Similarity Coefficient (DSC) of 0.874 for the trained network compared to 0.754 for the conventional region growing algorithm ($p < 0.001$).

Keywords: lateral ventricles, segmentation, deep neural network, fully convolutional neural networks, noisy labels, pseudo-label, large dataset

*Equal contribution

1. INTRODUCTION

Lateral ventricles are anatomical parts of the ventricular system in the brain, where the cerebrospinal fluid is produced. Ventricular volume and its progression are associated with several brain diseases. In certain forms of dementia, the increase of lateral ventricular volume has been associated to decline in cognitive function.[1] Some psychiatric illnesses such as schizophrenia have also been linked to enlargement in ventricular volume.[2] Additionally, asymmetrical shapes between the left and the right lateral ventricles together with the size of the ventricles can be indicative of abnormalities in the brain.[3]

Even though a rough estimation of the ventricular volume such as the number of slices that the ventricles appear in, might be sufficient for some applications, more accurate quantitative measurements are necessary to longitudinally study subtle differences. It has also been shown that leveraging spatial information using ventricles as landmarks are beneficial for the detection of a number of pathologies in the brain including white matter hyperintensities[4] and lacunes.[5] Though manual annotation of lateral ventricles might be an option on

Send correspondence to Jonas Teuwen: jonas.teuwen@radboudumc.nl.

Medical Imaging 2018: Image Processing, edited by Elsa D. Angelini, Bennett A. Landman, Proc. of SPIE
Vol. 10574, 105742U · © 2018 SPIE · CCC code: 1605-7422/18/$18 · doi: 10.1117/12.2293569

smaller datasets and cross-sectional studies, this would not be feasible otherwise as the task is time-consuming, laborious and subjective. Therefore an accurate, objective and independent segmentation of the left and right ventricles is desirable in clinical practice.

With the success of deep neural networks[6,7] in visual pattern recognition, many studies have been successfully conducted in the medical image analysis domain during the past few years,[8,9] that have resulted in intelligent systems that reach or surpass the level of medical experts on different tasks and domains.[10–12]

Since the recent deep learning approaches follow a data-driven strategy to learn the optimal representations for the specific tasks at hand, these methods often require large sets of annotated data to train on. Several recent studies have shown strong implications of training dataset size on the quality of trained networks. For instance, it has been shown that even with gigantic datasets, the performance of the trained network linearly scales with logarithm of the size of the training data.[13]

Given the reasoning above, the computer vision community has created enormous labeled datasets using crowd sourcing methods, for instance using Amazon mechanical turk. However this solution is not feasible for medical datasets, as the labeling process requires specific expertise that is only possible with medical experts available. Therefore, the high costs of gathering large medical datasets have still hindered feasibility of gigantic datasets that fully leverage the high capacity of the deep neural networks on various medical image analysis domains.

Another strategy to provide large labeled datasets is to use (not necessarily very accurate) available methods for the task in order to provide pseudo-labels. Using this, one can provide arbitrarily large datasets as far as unlabeled data is available. This however, arises a few interesting questions to be answered: 1) Considering an imposed trade-off between the dataset size and its relative label accuracy, would that make sense to train neural networks with noisy but large datasets rather than smaller ones with more accurate labels, and 2) In case we opt for the latter, is the low accuracy of the provided pseudo-labels necessarily an upper-bound for the accuracy of a trained network?

In this study, we aim to answer the aforementioned rather important questions by reporting a deep neural network that achieves high accuracy in segmenting the left and right ventricles separately, being trained on noisy pseudo-labels. We also show that, though desirable, accurate manual labels are not mandatory to produce good results, given a large set of (unbiased) noisy-labeled images.

2. METHODS

2.1 Material

The data used in this work is obtained from the RUN DMC[14] (Radboud University Nijmegen Diffusion Tensor and Magnetic Resonance Imaging Cohort), which is a longitudinal study of small vessel disease and its progression. The imaging protocol includes a 3D T1-magnetization prepared rapid gradient-echo (MPRAGE) pulse sequence with voxel size of $1.0 \times 1.0 \times 1.0$ mm^3 and a fluid attenuation inversion recovery (FLAIR) pulse sequence with voxel size $0.5 \times 0.5 \times 5.0$ mm^3 with a slice gap of 1 mm, scanned using a 1.5T MR scanner (Magnetom Sonata, Siemens Medical Solution, Erlangen, Germany).

We selected a subset of 397 subjects which was randomly split into sets of 246, 99 and 52 subjects for training, validation and testing purposes respectively.

2.2 Preprocessing

For an accurate segmentation, we need to take into account the possible movement of the patient between the acquisition of the T1 and FLAIR modalities. To align the image coordinates of both modalities, we rigidly registered the T1 images with the images using FSL-FLIRT.[15] In order to make the processing easier, we exclude non-brain tissue such as the skull, eyes, etc. We computed the brain mask using FSL-BET[15] on the T1 images. The resulting masks are then transformed using the computed transformation to the FLAIR images. To correct for the spatial intensity variations on the MR images caused by inhomogeneities in the magnetic field, we perform a bias-field correction using FSL-FAST.[15] As a final preprocessing step we normalize each image by dividing it with the 95$^{\text{th}}$ percentile of all intensities within the same image.

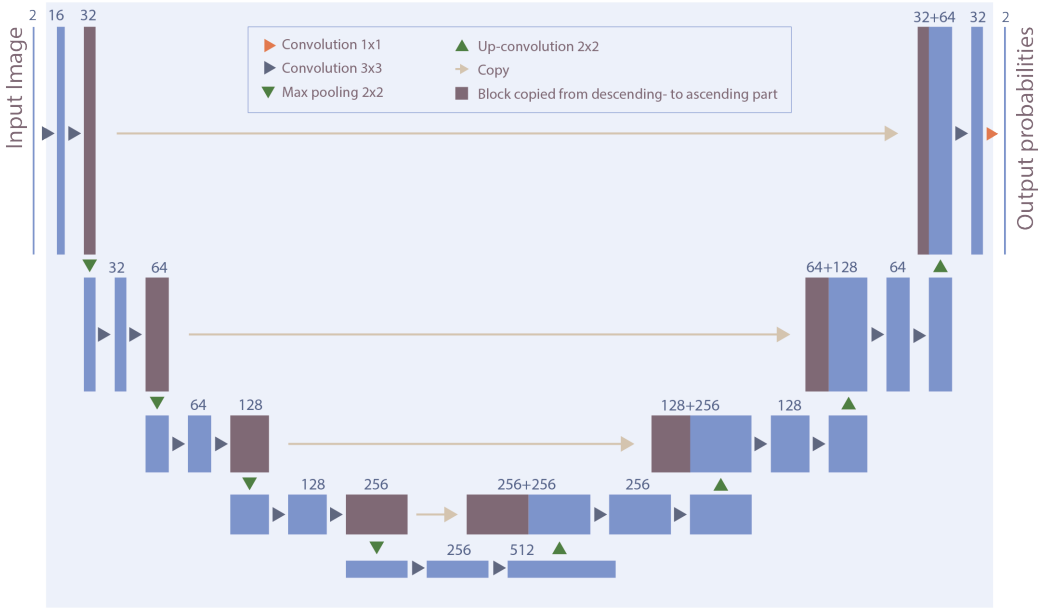

Figure 1: The fully convolutional network used for training our model. This is a U-net-like architecture[16,17] with slight modifications as described in the text.

2.3 Reference annotations

To generate the pseudo-labels for the training set, we used an in-house developed application where automatically selected seed points are used to perform a watershed-based segmentation algorithm on the T1 image to provide ventricle masks on the whole training, validation, and test sets. The algorithm is available in the commercial version of MeVisLab (MeVis Medical Solutions AG and Fraunhofer MEVIS; Bremen, Germany). The provided masks generated by the watershed region growing based algorithm are inaccurate in some cases or totally failing in some others. We therefore excluded 9 cases from the training set where the algorithm failed completely. In addition to this for evaluation purposes, the test set was independently manually segmented by an experienced reader on the registered T1 images, where the FLAIR images was used in cases of ambiguity.

2.4 Network architecture and training procedure

To segment the left and the right ventricles separately, we formulated the problem as a three-class segmentation of the background, left ventricle and right ventricle respectively. We utilized a fully convolutional network based on the U-net[16] architecture, with a depth of 5, applied slice-by-slice on a two-channel image composed of the T1 and FLAIR modalities. As in the analysis path of the standard U-net, we used 3×3 convolutional filters and 2×2 max pooling with $(2, 2)$ stride. We slightly deviated from the original architecture by using leaky ReLus with leakiness 0.01 and follow these by dropout with 0.1 probability and a batch normalization[18] layer. Additionally, we started the first convolutional layer with 16 filters and we already doubled the number of filters in each layer, before the max pooling to avoid bottlenecks.[17] We employed a similar scheme in the synthesis path. Details of the network architecture is illustrated in Figure 1.

We used the categorical cross-entropy loss function with L_2 regularization with $\lambda_2 = 10^{-5}$. To account for class imbalance, we weighted the loss function on the background by a factor of 0.01. The network weights were initialized from a Gaussian distribution $N(0, 2/\text{fan}_{\text{in}})$. To train the network, we used the Adam update rule[19] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. We trained our network for 200 epochs with an initial learning rate of 10^{-4} which was decreased in later epochs to 10^{-5}. The final model was selected as the model with the lowest validation loss. The network was trained using the imperfect segmentations made by the region growing method as described in section 2.3.

2.5 Evaluation of performance

We selected a threshold of 0.5 on the probability maps provided by the deep network to obtain binary segmentations. We compared the classical region growing method to the deep network thresholded output using the manual segmentations as the reference standard with the Dice Similarity Coefficient (DSC). The DSC is given by:

$$\mathrm{DSC}(R, X) = 2\frac{\sum_i |R_i \cap X_i|}{\sum_i |R_i| + |X_i|},$$

where R_i are the reference annotations for subject i, X_i are either the labels generated by the conventional approach or the thresholded network outputs and $|\cdot|$ is the size of the set.

We also report and compare receiver operating characteristic (ROC) curves that represent the methods with their true and false positive rates (sensitivity and 1−specificity) on various operating points. We use area under the ROC curve (AUC) as a single metric to quantitatively compare ROC curves. Furthermore, we use bootstrapping (over 1000 randomly created bootstraps) on the test set samples to report statistical significance test $p-$values. To be more specific, given "method A is no better than method B" as the null-hypothesis to reject, empirical $p-$value is reported as the proportion of bootstraps where method B results in a higher DSC.

3. RESULTS

Evaluating the methods on the test set, we obtained a DSC of 0.874 compared to 0.754 for the region growing based method, with respect to the manual annotations as the reference standard. The deep network significantly outperformed the region growing method ($p < 0.001$) on all three comparison scenarios of right, left and both ventricles segmentation, even though the network was trained on the outputs from that method. The DSC for segmentation of left and right ventricles and the whole lateral ventricles for the two methods are presented in Table 1, also shows a consistent improvement over the baseline method. Furthermore, an ROC analysis is illustrated on Figure 2 to provide a visual comparison between the two methods. It should be noted that since the region growing based model does not provide probabilistic outputs, its ROC analysis results consists of a single operating point, which is represented by a single point in the ROC graph. In Figure 3 we present a qualitative comparison between the different methods on a sample slice.

Table 1: Performance of the different methods on the left, right and both ventricles respectively.

	Left ventricle	Right ventricle	Both ventricles
DSC (region growing)	0.750	0.723	0.754
DSC (deep neural network)	0.881	0.867	0.874
AUC (deep neural network)	0.990	0.989	0.986

4. DISCUSSION AND CONCLUSIONS

Interestingly in the experiments, we observed that the deep model trained on imperfect ground truth could still get a decent training and outperform its ground truth generating method significantly. This is an interesting and important finding for the medical imaging domain where the high costs of generating large manually labeled datasets might seem to reject the feasibility of training deep neural networks that require gigantic training sets to achieve good performance. These results also show that a relatively low accuracy of the provided pseudo-labels is not necessarily an upper bound to the performance of the trained network.

For this to happen, there are two requirements that need to be satisfied: Firstly, the distribution of the samples with noisy labels should be adequately randomly scattered over the feature space. Otherwise, if the ground truth providing method is biased and constantly repeats the same error patterns the model would most likely learn the same error patterns. Secondly, the method should be regularized well enough to maintain its generalizability and not to overfit the noise patterns.

In this work, we presented a fully automated algorithm for the segmentation of the lateral ventricles on brain MR images that is well capable of discriminating between the left and right ventricles. Despite the noisy training labels, the network achieves a DSC of 0.874.

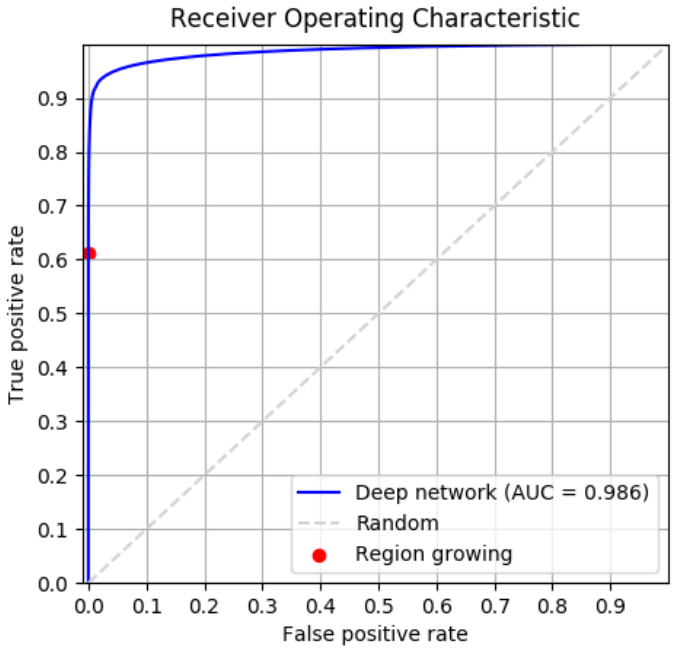

Figure 2: Receiver Operating Characteristic for the segmentation of both ventricles. Please note that the region growing algorithm is represented by a single point as the mentioned method is not probabilistic in contrast to the deep network.

REFERENCES

[1] J. Haxby, J. Gillette, D. Teichberg, *et al.*, "Longitudinal changes in lateral ventricular volume in datients with dementia of the alzheimer type," *Neurology* **42**(10), 2029–2029 (1992).

[2] I. C. Wright, S. Rabe-Hesketh, P. W. Woodruff, *et al.*, "Meta-analysis of regional brain volumes in schizophrenia," *American Journal of Psychiatry* **157**(1), 16–25 (2000).

[3] A. M. McKinney, *Enlargement or Asymmetry of the Lateral Ventricles Simulating Hydrocephalus*, 349–369. Springer International Publishing (2017).

[4] M. Ghafoorian, N. Karssemeijer, I. W. van Uden, *et al.*, "Automated detection of white matter hyperintensities of all sizes in cerebral small vessel disease," *Medical physics* **43**(12), 6246–6258 (2016).

[5] M. Ghafoorian, N. Karssemeijer, T. Heskes, *et al.*, "Deep multi-scale location-aware 3d convolutional neural networks for automated detection of lacunes of presumed vascular origin," *NeuroImage: Clinical* **14**, 391–399 (2017).

[6] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature* **521**, 436–444 (2015).

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Networks* **61**, 85–117 (2015).

[8] G. Litjens, T. Kooi, B. E. Bejnordi, *et al.*, "A survey on deep learning in medical image analysis," *Medical Image Analysis* **42**, 60 – 88 (2017).

[9] A. Rodriguez-Ruiz, J. Teuwen, S. Vreemann, *et al.*, "New reconstruction algorithm for digital breast tomosynthesis: better image quality for humans and computers," *Acta Radiologica* (2017).

[10] V. Gulshan, L. Peng, M. Coram, *et al.*, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama* **316**(22), 2402–2410 (2016).

[11] M. Ghafoorian, N. Karssemeijer, T. Heskes, *et al.*, "Location sensitive deep convolutional neural networks for segmentation of white matter hyperintensities," *Scientific Reports* **7** (2017).

[12] B. E. Bejnordi, M. Veta, P. J. van Diest, *et al.*, "Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer," *Jama* **318**(22), 2199–2210 (2017).

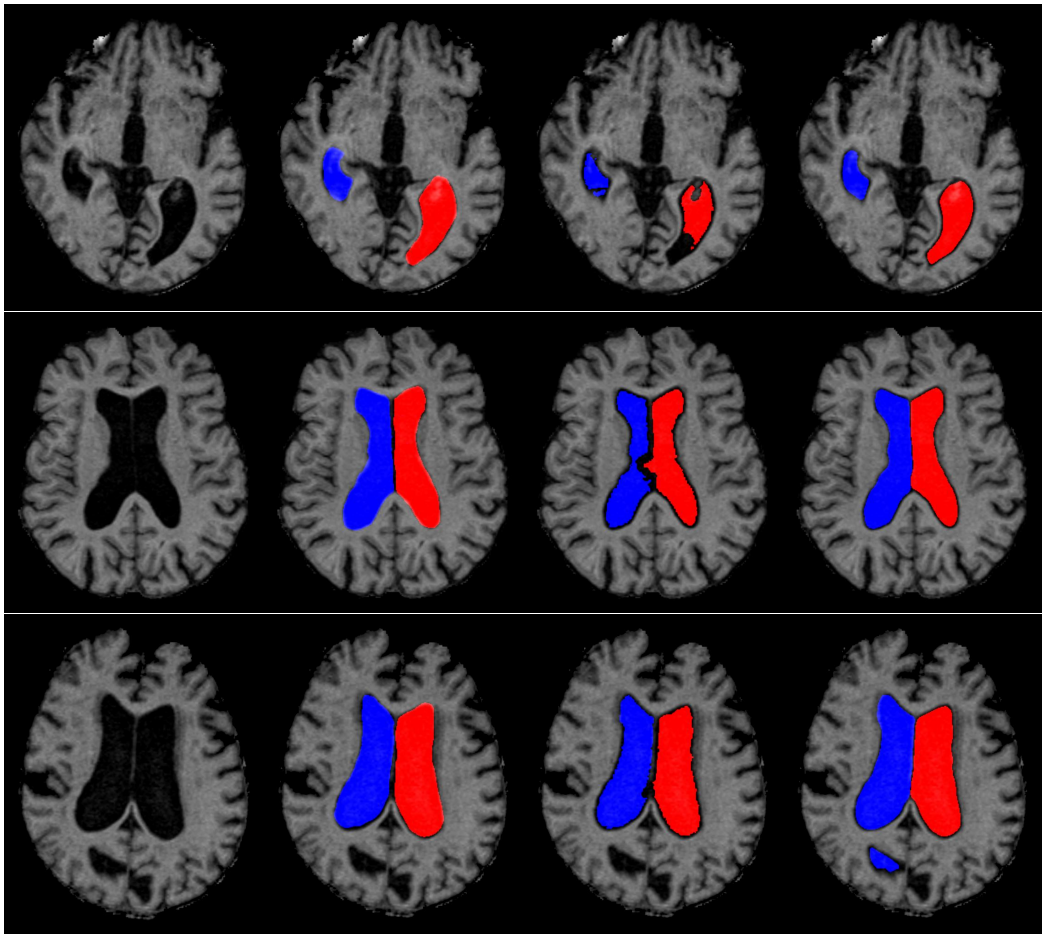

Figure 3: Sample slices of qualitatively representing results of our method. The four images in each represent the original T1 image, the manual segmentation, the output of the region growing algorithm used as reference standard and the output of the proposed method respectively. In the first two rows we have given examples where our method clearly outperforms the region growing algorithm. In the last row a case is given where our method makes false positives.

[13] C. Sun, A. Shrivastava, S. Singh, *et al.*, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *IEEE International Conference on Computer Vision (ICCV)*, (2017).

[14] A. G. van Norden, K. F. de Laat, R. A. Gons, *et al.*, "Causes and consequences of cerebral small vessel disease. the run dmc study: a prospective cohort study. study rationale and protocol," *BMC neurology* **11**(1), 29 (2011).

[15] M. Jenkinson, C. F. Beckmann, T. E. Behrens, *et al.*, "Fsl," *Neuroimage* **62**(2), 782–790 (2012).

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241, Springer (2015).

[17] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, *et al.*, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 424–432, Springer (2016).

[18] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International Conference on Machine Learning*, 448–456 (2015).

[19] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* (2014).