3560

Addressing the need for less MRI sequence dependent DL-based segmentation methods: model generalization to multi-site and multi-scanner data

Yasmina Al Khalil[1], Cristian Lorenz[2], Jürgen Weese[2], and Marcel Breeuwer[1,3]
[1]Biomedical Engineering Department, Eindhoven University of Technology, Eindhoven, Netherlands, [2]Philips Research Laboratories, Hamburg, Germany, [3]Philips Healthcare, MR R&D - Clinical Science, Best, Netherlands

Synopsis

The versatility of MRI acquisition parameters and sequences can have a substantial impact on the design and performance of medical image segmentation algorithms. Even though recent studies report excellent results of deep-learning (DL) based algorithms for tissue segmentation, their generalization capability and sequence dependence is rarely addressed, while being crucial for inclusion in clinical settings. This study attempts to demonstrate the lack of adaptation of such algorithms to unseen data from different sites and scanners. For this purpose, we use a 3D U-Net trained for brain tumor detection and test it site-wise to evaluate how well generalization can be achieved.

Introduction

A crucial component of MR image analysis is automated tissue segmentation, where state-of-the art methods using deep learning show an improvement in accuracy and speed compared to manual segmentation. Even though image acquisition and segmentation were for a long time considered and developed as two separate processes, MRI acquisition can have a substantial impact on segmentation performance[1, 2]. Variations in intrinsic acquisition parameters (relaxation constants and physiological parameters), scanning hardware, acquisition sequence implementation and imaging parameters cause significant differences in image quality, appearance and noise[3,4]. While crucial, the assessment of segmentation algorithms for their generalization capability to the afore-mentioned variations occurring across MRI datasets has rarely been performed. Generalizability of algorithms to unseen data has only recently been addressed by some studies, with a primary focus on applying unsupervised algorithms to alleviate the need for large annotated datasets[5, 6]. However, these methods require handcrafted parameter tuning and can be optimized only for particular tasks. Other studies focus on establishing good data pre-processing and augmentation pipelines, but systematic differences and site specific bias are still observable[2, 7].
In this case study we demonstrate how one of the most famous state-of-the-art algorithms for medical image segmentation deteriorates in performance if tested on unseen data coming from a different site or scanner. We perform this experiment on a subset of BRaTS 2015 dataset, which consists of computer-aided and manually corrected segmentation labels for the pre-operative multi-institutional scans of The Cancer Genome Atlas (TCGA) low grade glioma (LGG) collection[8]. While most reported results on this dataset are achieved by using the complete dataset for training in a heterogeneous scenario, we show how by simply changing the training protocol of the same network and evaluating it site-wise, the network performance reduces significantly.

Methods

The dataset used in this study consists of skull-stripped and co-registered multi-modal MRI volumes, containing a T1-weighted, a post-contrast T1-weighted, a T2-weighted and a FLAIR MRI for each patient (representative samples shown in Fig. 1). The dataset is heterogeneous in nature, originating from a number of institutions and acquired with different protocols, magnetic field strengths and MRI scanners. The provided ground truth segmentations contain tumor labels for edema, necrosis and non-enhancing tumor, and enhancing tumor. We split the data per site (4 sites in total), in order to evaluate how well a network trained for whole tumor segmentation generalizes to unseen data coming from a different site, simulating a realistic clinical scenario. We compare these results to a traditional cross-validation training procedure with 5 folds, where data is randomly split into train and test data.
We utilize a 3D U-Net[9], which we setup and train in line with the nnU-Net[10] recommendations, using an input patch size of 128x128x128 and a batch size of 2 per training iteration. We use the Adam optimizer with an initial learning rate of $2 * 10^{-4}$ and an l2 weight decay of 10^{-5}, as well as a combination of multi-class dice and cross-entropy loss. The exponential moving average of the training loss is used as an indicator of whether the learning rate should be reduced, where we take in account the last 30 epochs and reduce it by a factor of 0.2. In all cases, the network is trained in a five-fold cross-validation scenario, but for each case, one site is excluded from the training/validation set and used for inference only. We normalize the data using z-score normalization for each individual patient. We apply data augmentation in the form of random rotations, scaling and elastic deformations.

Results

Table 1 shows the experiment setup and segmentation performance of the trained 3D U-net across different sites. These results are compared to the baseline performance of the network trained and validated on all available sites (first row). We can observe that the segmentation performance significantly decreases in all cases compared to the baseline. However, with the addition of a small subset of data per site for each case, as in Table 2, the performance of the network significantly improves. Moreover, we have done further experiments by including only specific sites in the train set, while using others for testing, as shown in Table 3. We observe that the performance of the network significantly depends on the type of data available for training.

Discussion and Conclusion

The results of this study demonstrate some of the challenges of utilizing automated segmentation algorithms in realistic clinical settings. Combining multi-site and multi-vendor data is considered as a potential solution to improve algorithm generalization, as well as increase their statistical ability to detect small but crucial changes in anatomy[11, 12]. However, collecting sufficient data for every scenario is unrealistic, as medical data is generally hard to obtain and requires manual annotation. Moreover, recent studies show that building a large training database could affect algorithm performance negatively, suggesting that a better approach could be achieved by including more difficult cases in the dataset, rather than simply increasing database size with multi-site data[13]. Finally, the results in this study indicate that there is an evident need for changing the way algorithms are evaluated and results reported, especially considering their potential use in clinical settings.

Acknowledgements

This research is a part of the OpenGTN project, supported by the European Union in the Marie Curie Innovative Training Networks (ITN) fellowship program under project No. 764465.

References

1. Bai W, Sinclair M, Tarroni G, Oktay O, Rajchl M, Vaillant G, Lee AM, Aung N, Lukaschuk E, Sanghvi MM, Zemrak F. Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of Cardiovascular Magnetic Resonance. 2018 Dec;20(1):65.

2. Chen C, Bai W, Davies RH, Bhuva AN, Manisty C, Moon JC, Aung N, Lee AM, Sanghvi MM, Fung K, Paiva JM. Improving the generalizability of convolutional neural network-based segmentation on CMR images. arXiv preprint arXiv:1907.01268. 2019 Jul 2.3.

3. Karani N, Chaitanya K, Baumgartner C, Konukoglu E. A lifelong learning approach to brain mr segmentation across scanners and protocols. In International Conference on Medical Image Computing and Computer-Assisted Intervention 2018 Sep 16 (pp. 476-484). Springer, Cham.

4. Rajiah P, Bolen MA. Cardiovascular MR imaging at 3 T: opportunities, challenges, and solutions. Radiographics. 2014 Oct 13;34(6):1612-35.

5. Sun B, Saenko K. Deep coral: Correlation alignment for deep domain adaptation. InEuropean Conference on Computer Vision 2016 Oct 8 (pp. 443-450). Springer, Cham.

6. Long M, Cao Y, Wang J, Jordan MI. Learning transferable features with deep adaptation networks. arXiv preprint arXiv:1502.02791. 2015 Feb 10.

7. Kostro D, Abdulkadir A, Durr A, Roos R, Leavitt BR, Johnson H, Cash D, Tabrizi SJ, Scahill RI, Ronneberger O, Klöppel S. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. NeuroImage. 2014 Sep 1;98:405-15.

8. Bakas S, Akbari H, Sotiras A, Bilello M, Rozycki M, Kirby JS, Freymann JB, Farahani K, Davatzikos C. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data. 2017 Sep 5;4:170117.

9. Isensee F, Kickingereder P, Wick W, Bendszus M, Maier-Hein KH. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. InInternational MICCAI Brainlesion Workshop 2017 Sep 14 (pp. 287-297). Springer, Cham.

10. Isensee F, Petersen J, Kohl SA, Jäger PF, Maier-Hein KH. nnU-Net: Breaking the Spell on Successful Medical Image Segmentation. arXiv preprint arXiv:1904.08128. 2019 Apr 17.

11. Bento M, Souza R, Salluzzi M, Frayne R. Reliability of computer-aided diagnosis tools with multi-center MR datasets: impact of training protocol. InMedical Imaging 2019: Computer-Aided Diagnosis 2019 Mar 13 (Vol. 10950, p. 1095008). International Society for Optics and Photonics.

12. Thompson PM, Stein JL, Medland SE, Hibar DP, Vasquez AA, Renteria ME, Toro R, Jahanshad N, Schumann G, Franke B, Wright MJ. The ENIGMA Consortium: large-scale collaborative analyses of neuroimaging and genetic data. Brain imaging and behavior. 2014 Jun 1;8(2):153-82.

13. Zheng B, Wang X, Lederman D, Tan J, Gur D. Computer-aided detection: the effect of training databases on detection of subtle breast masses. Academic radiology. 2010 Nov 1;17(11):1401-8.

Figures

Test site	Number of subjects	Train set size	Test set size	Whole tumor Dice	IoU
All	65	58	7	0.859	0.767
CS	11	54	11	0.741	0.651
HT	13	52	13	0.697	0.562
FG	6	59	6	0.806	0.703

Table 1: Segmentation performance of the 3D U-Net across different sites. All scores are the mean Dice and IoU scores per each site.

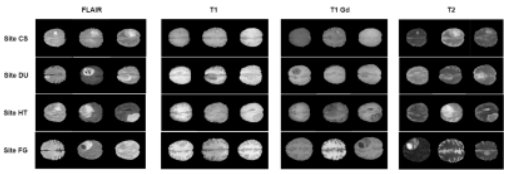

Figure 1: Representative depiction of contrast differences due to variation in scan parameters and vendors between the datasets from four different sites per each MRI modality. These are acquired for the same slice across all volumes.

Train site	Test site	Train set size	Test set size	Whole tumor Dice	IoU
All+CS	CS	57	8	0.812	0.703
All+HT	HT	55	10	0.741	0.613
All+FG	FG	61	4	0.831	0.72

Table 2: Segmentation performance of the network with a small subset of testing set added to the train set. Precisely, 3 volumes from site CS (row 1), 3 volumes from site HT (row 2) and 2 volumes from site FG (row 3) were added to the overall training set. We can observe improvement in segmentation performance compared to results obtained in Table 1.

Train site	Test site	Train set size	Test set size	Whole tumor Dice	IoU
DU	CS	35	11	0.584	0.406
DU	HT	35	13	0.522	0.392
DU	FG	35	6	0.791	0.641
DU + CS	CS	38	8	0.682	0.517
DU + HT	HT	38	10	0.639	0.483
DU + FG	FG	37	4	0.827	0.662

Table 3: Segmentation performance of the network trained on site DU (chosen for this experiment as it contains the most subjects) and tested across other sites (rows 1 to 3). Rows 4 to 6 demonstrate that with addition of a small subset of the test data, the segmentation performance and generalization capability of the network significantly increases.

Proc. Intl. Soc. Mag. Reson. Med. 28 (2020)
3560