Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210168

Challenges in Building of Deep Learning Models for Glioblastoma Segmentation: Evidence from Clinical Data

Anvar KURMUKOV[a,b], Aleksandra DALECHINA[c,1], Talgat SAPAROV[a,d], Mikhail BELYAEV[e], Svetlana ZOLOTOVA[c], Andrey GOLANOV[c] and Anna NIKOLAEVA[c]
[a] *Institute for Information Transmission Problems of the Russian Academy of Sciences (Kharkevich Institute), Moscow, Russia*
[b] *Higher School of Economics - National Research University, Moscow, Russia*
[c] *N. N. Burdenko National Medical Research Center of Neurosurgery, Moscow, Russia*
[d] *Moscow Institute of Physics and Technology, Moscow, Russia*
[e] *Skolkovo Institute of Science and Technology, Moscow, Russia*

Abstract. In this article, we compare the performance of a state-of-the-art segmentation network (UNet) on two different glioblastoma (GB) segmentation datasets. Our experiments show that the same training procedure yields almost twice as bad results on the retrospective clinical data compared to the BraTS challenge data (in terms of Dice score). We discuss possible reasons for such an outcome, including inter-rater variability and high variability in magnetic resonance imaging (MRI) scanners and scanner settings. The high performance of segmentation models, demonstrated on preselected imaging data, does not bring the community closer to using these algorithms in clinical settings. We believe that a clinically applicable deep learning architecture requires a shift from unified datasets to heterogeneous data.

Keywords. Deep learning, segmentation, glioblastoma, clinical data

1. Introduction

Recently, deep learning methods show great results in medical image segmentation. Automatic segmentation based on convolutional neural networks (CNN) speeds up the process of both tumour and organ at risk delineation, improving efficiency of the contouring process and reducing level of inter and intra-rater variability [1], [2], [3].
Automatic segmentation of brain tumours, especially gliomas is of great research interest. Many methods for glioma segmentation were developed under the competitions like Brain Tumor Segmentation Challenge (BraTS) and on the unified prospective datasets [4], [5], some of them even achieving beyond human-level performance [6]. However, a large amount of retrospective data, for instance, stored in radiation treatment planning systems remains unused. Recently, Eijgelaar et al. demonstrated that a model trained only on a BraTS data reached a median Dice score of 0.81 on BraTS test data

[1] Aleksandra Dalechina, Radiosurgery and Radiation therapy Department, N. N. Burdenko National Medical Research Center of Neurosurgery, E-mail: avdalechina@gmail.com

meanwhile the results for the model trained on the clinical data were much worse (Dice of 0.49) [7].

In this study we compare performance of the state-of-the art deep learning architecture on the same task but two different datasets: first, the unified, preprocessed BraTS data; second, clinical data collected with different acquisition protocols. While we were able to achieve *almost* human-level performance on a BraTS dataset using standard U-net-based CNN [8], applying the same architecture to a clinical data yielded relatively poor results. We argue that further research direction should be shifted from training on (even large) unified datasets and achieving beyond human-level performance, towards developing algorithms robust to data variability.

2. Materials and Methods

2.1. Clinical data

Our clinical data consists of fluid-attenuated inversion recovery (T2/FLAIR, further FLAIR) MRI images from 185 patients with glioblastoma (GB) treated with radiation therapy (RT) at the N.N. Burdenko National Medical Research Center of Neurosurgery from 2014 to 2019. The dataset includes 98 FLAIR images with 2D axial orientation and 87 images with 2D sagittal orientation[2]. Scanners settings and models variability is presented on Figure 1. Prior to the analysis all private patient information were anonymized. Throughout the paper we use the term "image" to denote a whole 3D MRI scan and the term "slice" for 2D slice (either in axial or sagittal plane).

2.2. *Model data*

As a reference dataset we used the data provided during BraTS 2015 challenge. These data include postoperative MRI scans of GB and low-grade glioma from 30 patients (we exclude synthetic data provided by the challenge organizers). All images were preprocessed, i.e. co-registered to the same anatomical template, interpolated to the same resolution (1 mm^3) and skull-stripped, full details could be found in a corresponding paper [5]. For the purpose of this paper we only used FLAIR MRI from the BraTS dataset. As we show in the following sections, using this modality alone allows us to achieve *almost* human-level segmentation.

2.3. Segmentation of Gross Tumour Volume

Our goal is to predict the Gross Tumour Volume (GTV) mask of GB. To compare segmentation results on our dataset and BraTS 2015 data we have chosen FLAIR as a "target" MRI sequence for GTV segmentation. In the BraTS dataset, GTV contours were obtained by one to four raters, following the same annotation protocol, and their annotations were approved by the experienced neuro-radiologists.

It is worth noting that in contrast to the homogeneous BraTS 2015 dataset our clinical data demonstrates high level of variability in GTV contours. Changing in the contouring guidelines and the number of radiation oncologists who delineated the

[2] Images have the highest resolution in the corresponding plane.

treatment planning volume over a 5 year period determine the heterogeneity in the segmentation mask of our dataset. In some cases GTV might differ from the FLAIR hyperintense area due to the both various contouring techniques and correction of the volume using additional MRI sequences and imaging modalities (Figure 2).

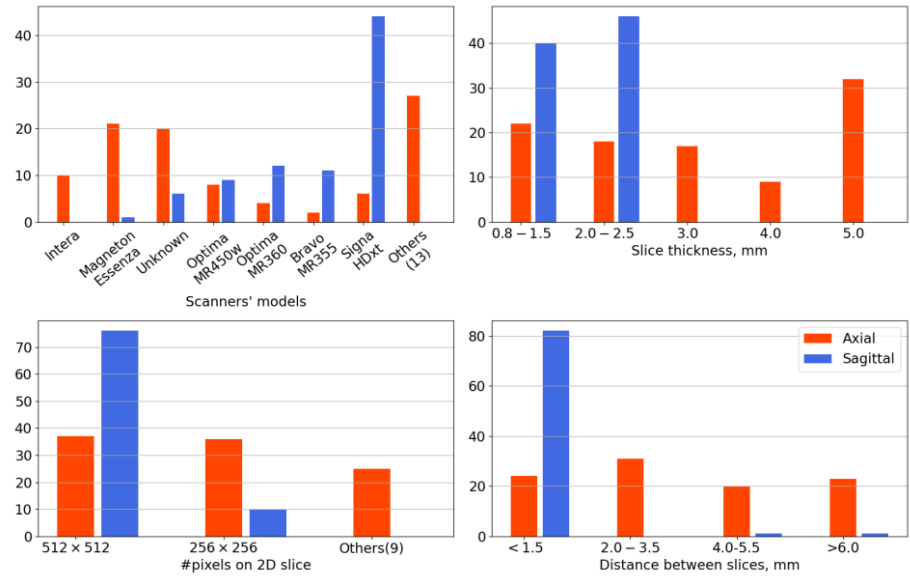

Figure 1. Clinical data scanners variability settings. Red - images with axial "main" plane; blue - images with sagittal main plane. Scanners include models from three manufacturers: Siemens, Philips and General Electric.

2.4. *Comparison metric, validation procedure*

To evaluate performance of the model Dice coefficient was used. Validation of the models was performed with five-fold cross-validation.

Due to the variability in segmentation masks of the clinical data (discussed in 2.3) we could not use the entire dataset in cross validation. We manually select 102 images with the GTV mask visually matching the FLAIR abnormality region (45 images from axial and 57 in the sagittal projections). In each of the subsamples for cross-validation, both these pre-selected images and the remaining images with "noise" in the segmentation masks were stratified. Thus, for training, all data were used, and for calculating the Dice metric on the test sample, only data without "noise" in the expert contours.

2.5. Training procedure

For our experiments we use a single CNN model: slightly modified U-Net [8] with residual blocks instead of plain convolutions, implemented within the PyTorch framework. We train all models for 200 epochs using batch SGD with Nesterov momentum, with x10 decrease of a learning rate after 75, 125 and 175 epochs (starting with 0.1) and binary cross-entropy as a loss function. We train all networks with patches of size 240 x 240 with 128 patches in a single batch. With the probability of 0.5 we sample the patch, so that it contains a full lesion, otherwise we sample it uniformly.

On clinical data we train two separate networks (of exactly the same architecture) on axial and sagittal planes. On the BraTS data we train three 2D networks: using axial 2D slices, using sagittal 2D slices, and on both axial and sagittal 2D slices. Since, all three networks yielded the same results we only report the Dice score for the latter. Additionally, we train a 3D U-Net on the BraTS dataset within the same training procedure, with the only difference in patch size (80 x 80 x 80) and convolution (3 x 3 x 3 instead of 3 x 3).

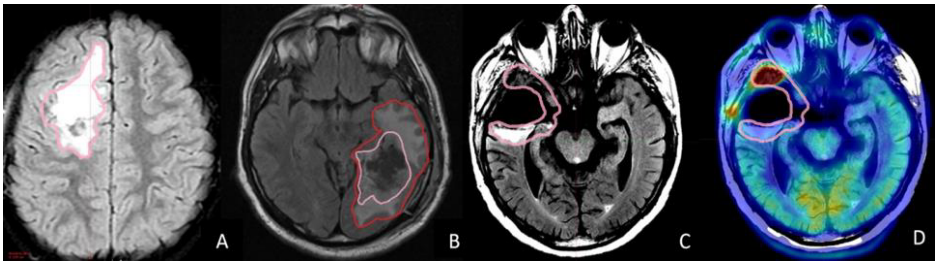

Figure 2. Examples of GTV contours in patients with GB. A - GTV includes FLAIR hyperintense area, B - GTV includes only the tumor bed (light pink), C - GTV includes FLAIR abnormalities but the contour was corrected according to PET (D).

3. Results

The 2D U-net achieved a mean Dice score of 0.72 (std 0.19) on the BraTS challenge data. On a sagittal subset of our clinical dataset we achieve mean Dice 0.71 (std 0.14). However, the same network architecture trained on an axial subset shows almost twice as bad results mean Dice score 0.47 (std 0.23). The 3D U-net model trained on the BraTS dataset achieves *almost* human-level performance with mean Dice score 0.80 (std 0.12).

Recall, that there is almost no variability of a sagittal subset of a clinical data Figure 1 (blue). Thus, we argue that the main reason for a poor performance on an axial subset is its high variability Figure 1 (red).

4. Discussion

In our work, we identify problems related to the successful building of a deep learning model based on clinical data for GB segmentation. The main source of these problems is intra-modal data variability. We demonstrate that variability in scanners' settings (intra-modal MRI FLAIR variability) hardly affects auto-segmentation models. We show that the CNN architecture capable of a human-level performance on unified and preprocessed data, failed to achieve the same results on a clinical data with high variability in scanners' settings, in a task of glioblastoma delineation on the postoperative MRI scans. Another issue related to clinical data is heterogeneity of "ground truth" manual labels. Thus, there is always "noise" in annotations of clinical dataset that prevents the building of a robust deep learning method in clinical settings.

3. Conclusion

Despite the neuroimaging improvements there is still no definitive consensus for RT treatment volume in GB. Even though several research groups demonstrated high performance of the developed models in terms of Dice metrics (over 0.80) [4],[6],[9]. These results were obtained on the prospective or retrospective data with segmentation masks created under a single guideline. The MRI scans were acquired on the same tomograph or according to an established scanning protocol. However, to this day there are no successful examples of algorithms trained on heterogeneous retrospective data. We believe that diverse scanner parameters and different acquisitions protocols are the most compelling obstacles for the successful usage of auto-segmentation models in a clinical setting.

Acknowledgments

The results have been obtained under the support of the Russian Foundation for Basic Research grant 18-29-01054.

References

[1] Nikolov S, Blackwell S, Mendes R, De Fauw J, Meyer C, Hughes C, Askham H, Romera-Paredes B, Karthikesalingam A, Chu C, Carnell D. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy 2018 Sep 12.

[2] Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, van Elmpt W, Dekker A. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. Radiotherapy and Oncology. 2018 Feb 1;126(2):312-7.

[3] Shirokikh B, Dalechina A, Shevtsov A, Krivov E, Kostjuchenko V, Durgaryan A, Galkin M, Osinov I, Golanov A, Belyaev M. Deep Learning for Brain Tumor Segmentation in Radiosurgery: Prospective Clinical Evaluation. In International MICCAI Brainlesion Workshop, Springer, Cham. 2019 Oct 17; p. 119-128.

[4] Bakas S, Reyes M, Jakab A, Bauer S, Rempfler M, Crimi A, Shinohara RT, Berger C, Ha SM, Rozycki M, Prastawa M. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge; 2018 Nov 5.

[5] Menze BH, Jakab A, Bauer S, Kalpathy-Cramer J, Farahani K, Kirby J, Burren Y, Porz N, Slotboom J, Wiest R, Lanczi L. The multimodal brain tumor image segmentation benchmark (BRATS). IEEE transactions on medical imaging. 2014 Dec 4;34(10):1993-2024.

[6] Ermiş E, Jungo A, Poel R, Blatti-Moreno M, Meier R, Knecht U, Aebersold DM, Fix MK, Manser P, Reyes M, Herrmann E. Fully automated brain resection cavity delineation for radiation target volume definition in glioblastoma patients using deep learning. Radiation oncology. 2020 Dec;15:1-0.

[7] Eijgelaar RS, Visser M, Müller DM, Barkhof F, Vrenken H, van Herk M, Bello L, Conti Nibali M, Rossi M, Sciortino T, Berger MS. Robust Deep Learning–based Segmentation of Glioblastoma on Routine Clinical MRI Scans Using Sparsified Training. Radiology: Artificial Intelligence. 2020 Sep 30;2(5):e190103.

[8] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. International Conference on Medical image computing and computer-assisted intervention. Cham. Springer. 2015;234-241.

[9] Kickingereder P, Isensee F, Tursunova I, Petersen J, Neuberger U, Bonekamp D, Brugnara G, Schell M, Kessler T, Foltyn M, Harting I. Automated quantitative tumour response assessment of MRI in neuro-oncology with artificial neural networks: a multicentre, retrospective study. The Lancet Oncology. 2019 May 1;20(5):728-40.