Automatic quality control of segmentation results using early epochs as data augmentation: application to choroid plexuses

Arya Yazdan-Panah, Bruno Stankoff, Olivier Colliot

Automatic quality control of segmentation results using early epochs as data augmentation: application to choroid plexuses

Arya Yazdan-Panah[a], Bruno Stankoff[b], and Olivier Colliot[a]

[a]Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM, CNRS, Inria, Inserm, AP-HP, Hôpital de la Pitié Salpêtrière, F-75013, Paris, France
[b]Sorbonne Université, Institut du Cerveau - Paris Brain Institute -, ICM, CNRS, Inserm, AP-HP, Hôpital Saint-Antoine, F-75012, Paris, France.

ABSTRACT

The establishment of automated image segmentation methods in medical imaging allows the analysis of very large datasets. However, visual quality control (QC) of segmentation results is impractical in large datasets, hence the need for automatic QC. In this paper, we introduce a novel automatic approach for QC of segmentation results. We developed a QC deep learning model (referred to as *QC model*) that, for a given patient, predicts the accuracy of the corresponding automatic segmentation (in our work the Dice score) provided by a deep learning segmentation model (referred to as *segmentation model*) in the absence of a ground truth annotation. To train the *QC model*, we introduce data augmentation by using the early epochs of the *segmentation model*. These early epochs allow us to feed the training of the *QC model* with examples of poor segmentation. We applied our approach to the QC of automatic segmentation of the choroid plexuses of the brain from MRI in controls and patients with multiple sclerosis. However, the method is generic and could be used with any segmentation model. The experiments showed that the proposed approach is very effective for predicting the segmentation accuracy with a correlation coefficient of 0.92, an R^2 of 0.763, a mean absolute error (MAE) of 0.078, and a mean squared error (MSE) of 0.009. Overall, this work shall provide a valuable tool for the automatic QC of segmentation results.

Keywords: Automatic Segmentation, Quality Control, Quality Assessment, Data Augmentation, Deep Learning

1. INTRODUCTION

Deep-learning-based methods are the state of the art for medical image segmentation. Such automatic methods allow the processing of very large imaging databases of several thousands of subjects. Even though the performances of the automatic segmentation models are remarkable, it remains necessary to perform quality control (QC) of segmentation results in order to identify poor segmentations which can bias subsequent analyses. However, visual QC is time-consuming and may thus be impractical for very large datasets. Therefore the need for automatic QC methods.

In medical imaging, automatic QC can refer to controlling raw images,[1–4] or post-processing results.[5–20] Here, we are concerned with the latter, more specifically segmentation results. Some methods for automatic QC of segmentation results rely on the manual annotation of segmentation quality in a set of samples (e.g. label them as "pass/fail"). These manual QC annotations are then used to train the QC model. This method requires manual labeling and suffers from a possible subjectivity of the annotator. Other approaches train a QC model to predict a segmentation accuracy metric (for instance the Dice accuracy). This approach has been followed in several studies. Non-exhaustively, methods include the registration of templates,[7,9] the extraction of manually defined features,[1,5] and deep learning methods.[8,10–12,15,16,19] Such an approach has the advantage of predicting a quality metric (e.g. the Dice accuracy) that is more informative than a simple "pass/fail". However, the

Further author information: (Send correspondence to Arya Yazdan-Panah)
Arya Yazdan-Panah: E-mail: arya.yazdan-panah@icm-institute.org or aypnaf@gmail.com

training requires a large and diverse set of segmentation outputs together with manual segmentation ground truth.

Another class of QC method includes reconstruction of labels using auto-encoder-like architectures.[21] Such methods take as input the labels and project them onto a "correct" label space. This has the advantage of not being restricted to a given metric used during training and thus the user can use any metric between the label and the reconstructed pseudo-ground truth. These types of methods may however not be best-suited to objects of interest that present highly variable spatial conformation between subjects and even between visits for the same subject.

In this paper, we propose a new method for automatic QC of segmentation results in medical imaging. To that purpose, we train a QC deep learning model (referred to as *QC model*) that predicts the accuracy of the corresponding automatic segmentation provided by a deep learning segmentation model (referred to as *segmentation model*). For training the model, we use a specific data augmentation technique that uses outputs of early epochs from the *segmentation model*. We developed the method for QC of automatic segmentation of the Choroid Plexuses (ChP),[22] but this concept could be easily applied to other tasks.

2. METHODS

2.1 Dataset and segmentation model

For this study, we utilized a previously published *segmentation model* that segments the choroid plexuses of the brain T1-weighted (T1w) from magnetic resonance imaging (MRI) scans[22] (please refer to this paper for details regarding the segmentation model). For the development of the model, we relied on data from 168 scans of individuals diagnosed with multiple sclerosis (MS, N=97), subjects with a radiologically isolated syndrome (RIS, N=27), and healthy controls (HC, N=44). The dataset was split at the subject level into three sets: training (72 subjects), validation (19 subjects), and testing (77 subjects). In the present paper, we use the same split in order to avoid data leakage.

2.2 Proposed automatic QC approach

Let us denote M_{seg} an automatic segmentation method, I an image to segment, $S_a = M_{seg}(I)$ the automatic segmentation of that image, and S_m the manual segmentation (ground-truth). The quality of the automatic segmentation is measured by comparison of S_a and S_m and denoted as $Q_S = score(S_a, S_m)$, with *score* being a metric chosen by the user and relevant to the task (e.g. Dice coefficient, Jaccard index, Hausdorff distance, ...). We aim at developing a model M_{QC}, such that $M_{QC}(I, S_a) = \hat{Q}_S \approx Q_S$. For this task, we chose *score* to be the Dice coefficient.

Let $M_{seg}^{k,e}$ be a segmentation model generated during the training of the segmentation algorithm such that

$$S_a^{k,e} = M_{seg}^{k,e}(I) \begin{cases} k \in [1,2,3,4,5] & \text{The cross-validation fold number} \\ e \in [1,2,\ldots,e_{max}] & \text{The training epoch} \end{cases} \quad (1)$$

The optimal segmentation model is selected at the epoch e_{opt} (in our case $e_{max} = 400$ and $e_{opt} = 200$ for all folds). We plotted the evolution of Q_S as a function of the epoch for each fold. We selected 10 additional epochs (within $[0, e_{opt}[$) of interest per fold to ensure variability of input Q_S for the training of M_{QC}. For each subject, one thus obtains: one image I, one manual segmentation S_m, and $5 \times (10 + 1) = 55$ automatic segmentations S_a (and therefore 55 scores Q_S). Finally, images were randomly flipped along the coronal plane with a 50% probability during training. To evaluate the effect of selecting additional epochs, we have performed training only with $S_a^{k,e_{opt}}$, denoted as without data augmentation (DA), as well as training with the full dataset.

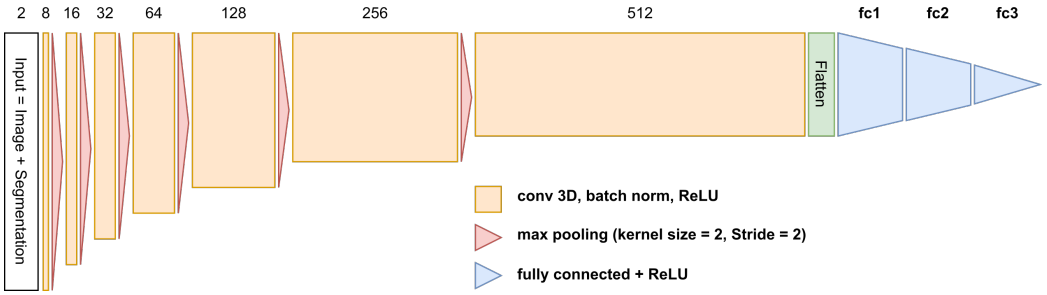

Figure 1. Model trained for predicting QC scores. Numbers above each convolution block represent the number of output features. The number of inputs of block "fc1" is dependent on the size of the input.

2.3 Proposed method

The *QC model* is a simple convolutional neural network composed of 7 convolutional and 3 fully connected layers (conv7fc3). Each convolutional layer (conv) is composed of a 3D convolution, batch normalization, and a ReLu. The output of the last conv is flattened. Each fully connected layer (FC) is followed by a ReLu. The output of the final FC is a single neuron. A schematic of the network is given in Figure 1. To avoid overfitting, we implemented dropout with a 0.5 probability.

We have tested several loss functions for the training of the network of the following form:

$$\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i|^p \begin{cases} \text{Mean Root Absolute Error (MRAEloss)} & \text{if } p=0.5 \\ \text{Mean Absolute Error (L1loss)} & \text{if } p=1 \\ \text{Mean Squared Error (MSEloss)} & \text{if } p=2 \end{cases} \tag{2}$$

where:

n : Batch size
y_i : True quality score of the i-th image-segmentation pair
$\hat{y}_i$: Predicted quality score of the i-th image-segmentation pair

2.4 Implementation details and training procedure

The experiments were conducted using PyTorch.[23] For optimization, we employed a weighted Adam optimizer[24] with an initial learning rate of 10^{-3}. This learning rate was halved when the validation loss exhibited less than 10^{-4} variation for an epoch, and further halving was restricted by a one-epoch cooldown period. The hardware setup consisted of Nvidia Tesla V100 32Go graphics cards, which allowed the use of a batch size of 17. To manage experiment tracking, visualization, and memory monitoring, we leveraged the Python package Weights&Biases.[25]

We used the same data split as in our previous publication: training (72 subjects, 3960 image-segmentation pairs), validation (19 subjects, 1045 pairs), and testing (77 subjects, 4235 pairs). During model training, two phases were employed: 10 epochs with DA and 100 epochs without. All training procedures were carried out using 5-fold cross-validation.

2.5 Performance evaluation

Performance was assessed using the mean absolute error (MAE), the mean square error (MSE), the coefficient of determination ($R2$), and Pearson's correlation coefficient (ρ). We plotted the true score Q_S against the predicted score $\hat{Q}_S$. We compared the results of the proposed method with data augmentation (DA) to those obtained without DA. We also studied the influence of the loss. We further compared results to a *trivial* model which consists in always predicting the average of the training set on the validation set (this model is created on the DA dataset).

3. EXPERIMENTS AND RESULTS

Performances are reported in Table 1. Results are reported as mean ± standard error. On the validation set, the best performances are obtained with the models trained with DA, notably with L1loss and MSEloss, while MRAEloss showed poorer performances. All models trained without DA performed poorly, even worse than the *trivial* model. Figure 2 shows the scatter plot of Q_S against $\hat{Q}_S$. On the test set, the performances of the proposed approach remained very good (correlation of 0.92, MAE of 0.078) but were lower than on the validation set.

Dataset	Model	Loss	DA	MAE	MSE	ρ	R^2
Validation	*trivial*	-	yes	0.165±0.004	0.040±0.002	-	0
	conv7fc3	MRAEloss	no	0.192±0.005	0.064±0.003	0.051	-0.624
	conv7fc3	L1loss	no	0.190±0.005	0.063±0.003	0.056	-0.594
	conv7fc3	MSEloss	no	0.194±0.005	0.066±0.003	0.022	-0.655
	conv7fc3	MRAEloss	yes	0.106±0.002	0.015±0.000	0.9	0.625
	conv7fc3	**L1loss**	**yes**	**0.052±0.001**	**0.005±0.000**	**0.941**	**0.885**
	conv7fc3	MSEloss	yes	0.059±0.001	0.005±0.000	0.94	0.873
Test	**conv7fc3**	**L1loss**	**yes**	**0.078±0.001**	**0.009±0.000**	**0.920**	**0.763**

Table 1. Performances of the different models on the validation and testing dataset on one fold. The best-performing model is denoted in boldface. The last line corresponds to the performances of the model on the independent testing set. For the *trivial* model, *rho* is not defined as it is a constant prediction of the mean. (DA: Data Augmentation; MAE: Mean Absolute Error; MSE: Mean Squared Error; ρ: Pearson's correlation coefficient; R^2: Coefficient of determination)

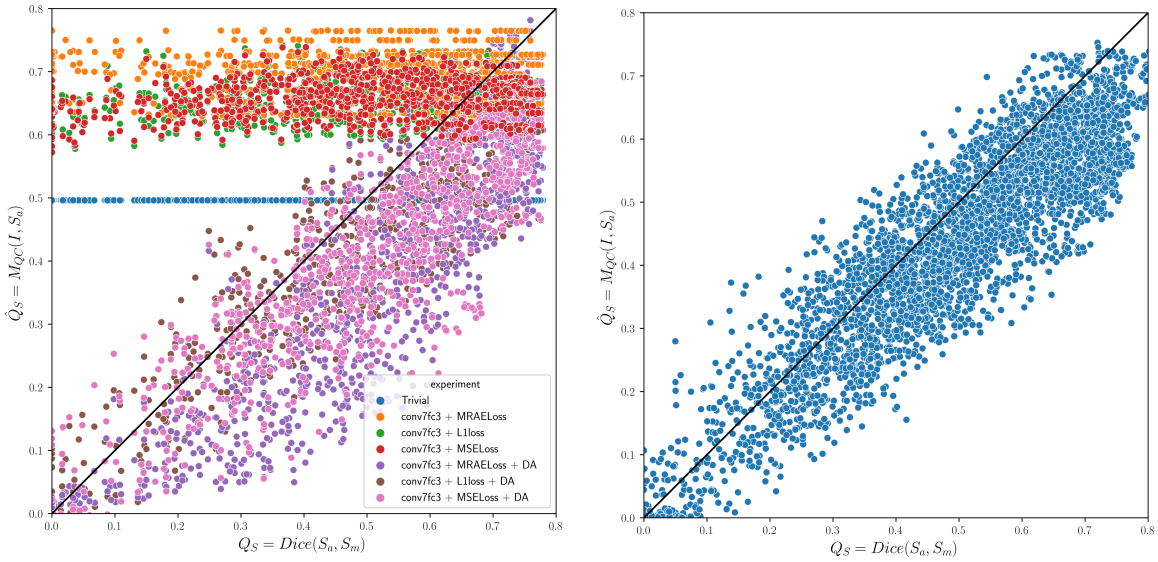

Figure 2. Scatter plot of Q_S against $\hat{Q}_S$. The identity line is marked in black. Left panel: comparison of the different approaches on the validation set. Right panel: evaluation of the proposed approach (conv7fc3 + L1loss + DA) on the test set.

4. DISCUSSION

We proposed a new approach for automatic QC of deep-learning-based segmentation. To that purpose, we used results from early epochs of the segmentation model for data augmentation, thereby providing a variety of possible segmentation quality and therefore ground truth annotations.

Our best-performing model, trained with DA, with the L1loss, reached a high correlation coefficient ($\rho = 0.92$) and coefficient of determination ($R^2 = 0.763$) indicating a strong relationship between predicted and true quality

scores. However, the performance was slightly lower than those obtained on the validation set. This requires further investigation as it may indicate a slight overfitting of the validation set by trying different approaches and losses. Nevertheless, the performance remained high and indicates that the proposed method can potentially be useful for automatic QC of large datasets.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the French government under the management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6), and from ICM under the Big Brain Theory program (project IMAGIN-DEAL in MS). This work was performed using HPC resources from GENCI-IDRIS (Grant 20XX-AD011013666R1).

REFERENCES

[1] Esteban, O., Birman, D., Schaer, M., Koyejo, O. O., Poldrack, R. A., and Gorgolewski, K. J., "Mriqc: Advancing the automatic prediction of image quality in mri from unseen sites," *PloS one* **12**(9), e0184661 (2017).

[2] Alfaro-Almagro, F., Jenkinson, M., Bangerter, N. K., Andersson, J. L., Griffanti, L., Douaud, G., Sotiropoulos, S. N., Jbabdi, S., Hernandez-Fernandez, M., Vallee, E., et al., "Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank," *Neuroimage* **166**, 400–424 (2018).

[3] Sujit, S. J., Coronado, I., Kamali, A., Narayana, P. A., and Gabr, R. E., "Automated image quality evaluation of structural brain mri using an ensemble of deep learning networks," *Journal of Magnetic Resonance Imaging* **50**(4), 1260–1267 (2019).

[4] Bottani, S., Burgos, N., Maire, A., Wild, A., Ströer, S., Dormont, D., and Colliot, O., "Automatic quality control of brain t1-weighted magnetic resonance images for a clinical data warehouse," *Medical Image Analysis* **75**, 102219 (2022).

[5] Kohlberger, T., Singh, V., Alvino, C., Bahlmann, C., and Grady, L., "Evaluating segmentation error without ground truth," in [*International Conference on Medical Image Computing and Computer-Assisted Intervention*], 528–536, Springer (2012).

[6] Li, K., Ye, C., Yang, Z., Carass, A., Ying, S. H., and Prince, J. L., "Quality assurance using outlier detection on an automatic segmentation method for the cerebellar peduncles," in [*Medical Imaging 2016: Image Processing*], **9784**, 398–404, SPIE (2016).

[7] Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B., "Reverse Classification Accuracy: Predicting Segmentation Performance in the Absence of Ground Truth," *IEEE Transactions on Medical Imaging* **36**, 1597–1606 (Aug. 2017).

[8] Robinson, R., Oktay, O., Bai, W., Valindria, V., Sanghvi, M., Aung, N., Paiva, J., Zemrak, F., Fung, K., Lukaschuk, E., Lee, A., Carapella, V., Kim, Y. J., Kainz, B., Piechnik, S., Neubauer, S., Petersen, S., Page, C., Rueckert, D., and Glocker, B., "Real-time prediction of segmentation quality," (2018).

[9] Robinson, R., Valindria, V. V., Bai, W., Oktay, O., Kainz, B., Suzuki, H., Sanghvi, M. M., Aung, N., Paiva, J. M., Zemrak, F., et al., "Automated quality control in image segmentation: application to the uk biobank cardiovascular magnetic resonance imaging study," *Journal of Cardiovascular Magnetic Resonance* **21**(1), 1–14 (2019).

[10] DeVries, T. and Taylor, G. W., "Leveraging uncertainty estimates for predicting segmentation quality," (2018).

[11] Liu, F., Xia, Y., Yang, D., Yuille, A. L., and Xu, D., "An alarm system for segmentation algorithm based on shape model," in [*Proceedings of the IEEE/CVF International Conference on Computer Vision*], 10652–10661 (2019).

[12] Zhou, L., Deng, W., and Wu, X., "Robust image segmentation quality assessment," *arXiv preprint arXiv:1903.08773* (2019).

[13] Klapwijk, E. T., Van De Kamp, F., Van Der Meulen, M., Peters, S., and Wierenga, L. M., "Qoala-t: A supervised-learning tool for quality control of freesurfer segmented mri data," *Neuroimage* **189**, 116–129 (2019).

[14] Arbelle, A., Elul, E., and Raviv, T. R., "Qanet – quality assurance network for image segmentation," (2019).

[15] Wang, S., Tarroni, G., Qin, C., Mo, Y., Dai, C., Chen, C., Glocker, B., Guo, Y., Rueckert, D., and Bai, W., "Deep generative model-based quality control for cardiac mri segmentation," in [*Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*], 88–97, Springer (2020).

[16] Xia, Y., Zhang, Y., Liu, F., Shen, W., and Yuille, A. L., "Synthesize then compare: Detecting failures and anomalies for semantic segmentation," in [*Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*], 145–161, Springer (2020).

[17] Fournel, J., Bartoli, A., Bendahan, D., Guye, M., Bernard, M., Rauseo, E., Khanji, M. Y., Petersen, S. E., Jacquier, A., and Ghattas, B., "Medical image segmentation automatic quality control: A multi-dimensional approach," *Medical Image Analysis* **74**, 102213 (2021).

[18] Gadewar, S., Zhu, A. H., Thomopoulos, S. I., Li, Z., Gari, I. B., Maiti, P., Thompson, P. M., and Jahanshad, N., "Region specific automatic quality assurance for mri-derived cortical segmentations," in [*2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*], 1288–1291, IEEE (2021).

[19] Li, K., Yu, L., and Heng, P.-A., "Towards reliable cardiac image segmentation: Assessing image-level and pixel-level segmentation quality via self-reflective references," *Medical Image Analysis* **78**, 102426 (2022).

[20] Sims, Z., Strgar, L., Thirumalaisamy, D., Heussner, R., Thibault, G., and Chang, Y. H., "Seg: Segmentation evaluation in absence of ground truth labels," *bioRxiv* , 2023–02 (2023).

[21] Galati, F. and Zuluaga, M. A., "Efficient model monitoring for quality control in cardiac image segmentation," in [*International Conference on Functional Imaging and Modeling of the Heart*], 101–111, Springer (2021).

[22] Yazdan-Panah, A., Schmidt-Mengin, M., Ricigliano, V. A. G., Soulier, T., Stankoff, B., and Colliot, O., "Automatic segmentation of the choroid plexuses: Method and validation in controls and patients with multiple sclerosis," *NeuroImage: Clinical* **38**, 103368 (Jan. 2023).

[23] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al., "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems* **32** (2019).

[24] Kingma, D. P. and Ba, J., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980* **1** (2014).

[25] Biewald, L. et al., "Experiment tracking with weights and biases," *Software available from wandb. com* **2**, 233 (2020).