Physics and Imaging in Radiation Oncology 29 (2024) 100528

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Correspondence

Response letter to Wahid *et al.* regarding our publication “A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations”

We would like to thank Wahid *et al.* for their insightful comments on our recently published paper [1]. In this paper our aim was to identify a network uncertainty-based metric that correlated strongly with deep learning-based auto-segmentation performance and could thus be used to detect auto-segmentations that require manual review. Wahid *et al.* rightly highlight that the calibration and interpretation of network uncertainty estimations are still points of contention in the field. We acknowledge that it is important to study network uncertainty at a fundamental level. In addition, we believe that the synergy between fundamental and pragmatic approaches is desirable in order to find a workable solution for the problem of auto-segmentation quality assurance in clinical practice.

In our work, we took a pragmatic, application-oriented approach to this problem. Our approach was pragmatic in two ways: first, we proposed a metric that is readily available for most auto-segmentation applications without requiring additional training, changes to the network architecture or multiple runs at inference time; second, we focused on error detection based on clinically relevant performance metrics like the mean surface distance and the surface Dice. We acknowledge the caveats of our approach highlighted by Wahid *et al.* and appreciate the relevant suggestions of possible future research directions.

The first point raised by the authors concerns the use of softmax scores as uncertainty estimates. In particular, they point out that miscalibrated uncertainty estimations may compromise out-of-distribution detection and propose the use of Bayesian approaches like Monte Carlo [2] and deep ensembles [3] instead. In literature, softmax scores are indeed often reported to be miscalibrated [4]. In recent years, however, the perspective that large networks are highly miscalibrated has become less consensual, with recent work claiming that modern networks can be reasonably calibrated too [5,6]. Moreover, other evidence suggests that Bayesian approaches are also miscalibrated, and that applying techniques like temperature scaling can help reduce this problem [7,8]. We fully agree that the issue of calibration is a non-trivial one, worthy of an in-depth examination.

Notably, most published literature on deep learning network calibration focuses on classification networks, which have arguably different properties than auto-segmentation networks. State-of-the-art auto-segmentation networks are typically trained with Dice and/or binary cross entropy as loss functions. However, in clinical practice we would like to detect auto-segmentations that fail under clinically relevant criteria such as mean surface distance and surface Dice. These metrics are only indirectly related to the metrics used for training. The impact of network miscalibration on the ability to detect these clinically relevant errors is not straightforward and merits further investigation. In our work, the proposed metric was not explicitly calibrated and yet was effective in the detection of segmentations that required review, as demonstrated by the high AUC values obtained on independent test sets.

The second point raised by Wahid *et al.* concerns the entanglement of the different uncertainty types. We recognize that the softmax outputs intertwine epistemic and aleatoric uncertainties. To distinguish between the two categories of uncertainty, alternative methods of uncertainty estimation would be more appropriate. We would like to add, however, that the precise meanings of aleatoric and epistemic uncertainty are also subject of undergoing discussion in the medical image segmentation field. In particular, a remaining question is whether they can even be perfectly disentangled [9–12]. Finally, the degree to which various sources of uncertainty are correlated with the necessity of clinically relevant edits remains unclear and subject for further studies.

For future applied research work, we believe that model calibration should be assessed, better understood and if needed optimized in a heuristic manner, likely for each particular combination of dataset and model architecture. Similarly, the interpretation of the different uncertainty types should be carried out on a case-by-case basis, linking the mathematical model interpretation with the clinical interpretation of the auto-segmentation. Insights from both research directions are likely to benefit the entire field.

References

[1] Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, et al. A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations. Phys Imaging Radiat Oncol 2023; 28:100500. https://doi.org/10.1016/j.phro.2023.100500.

[2] Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep. arXiv:1506.02142. Learning 2015. https://doi.org/10.48550/arXiv.1506.02142.

[3] Lakshminarayanan B, Pritzel A, Blundell C. Simple and Scalable Predictive Uncertainty Estimation using. arXiv:1612.01474. Deep Ensembles 2016. https://doi.org/10.48550/arXiv.1612.01474.

[4] Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of Modern Neural Networks. In: Doina P, Yee Whye T, editors. Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research: PMLR; 2017. p. 1321–30.

[5] Minderer M, Djolonga J, Romijnders R, Hubis F, Zhai X, Houlsby N, et al. arXiv: 2106.07998 Neural Netw 2021. https://doi.org/10.48550/arXiv.2106.07998.

[6] Carrell AM, Mallinar N, Lucas J, Nakkiran P. p. arXiv:2210.01964 The Calibration Generalization Gap 2022. https://doi.org/10.48550/arXiv.2210.01964.

[7] Rahaman R, Thiery AH. arXiv:2007.08792. Uncertainty Quantification and Deep Ensembles 2020. https://doi.org/10.48550/arXiv.2007.08792.

[8] Laves M-H, Ihler S, Kortmann K-P, Ortmaier T. Well-calibrated Model Uncertainty with Temperature Scaling for Dropout Variational Inference. 2019. arXiv: 1909.13550. doi:10.48550/arXiv.1909.13550.

[9] van den Berg CAT, Meliadò EF. Uncertainty Assessment for Deep Learning Radiotherapy Applications. Sem in Radiat Oncol 2022;32:304–18. https://doi.org/10.1016/j.semradonc.2022.06.001.

[10] Kiureghian AD, Ditlevsen O. Aleatory or epistemic? Does it matter? Struct Saf 2009;31:105–12. https://doi.org/10.1016/j.strusafe.2008.06.020.

[11] Barragán-Montero A, Bibal A, Dastarac MH, Draguet C, Valdés G, Nguyen D, et al. Towards a safe and efficient clinical implementation of machine learning in radiation oncology by exploring model interpretability, explainability and data-model

https://doi.org/10.1016/j.phro.2023.100528
Received 12 December 2023; Accepted 18 December 2023
Available online 29 December 2023

dependency. Phys Med Biol 2022;67:11TR01. https://doi.org/10.1088/1361-6560/ac678a.

[12] Ståhl N, Falkman G, Karlsson A, Mathiason G. Evaluation of Uncertainty Quantification in Deep Learning. In: Lesot M-J, Vieira S, Reformat MZ, Carvalho JP, Wilbik A, Bouchon-Meunier B, Yager RR, editors. Information Processing and Management of Uncertainty in Knowledge-Based Systems. Cham: Springer International Publishing; 2020. p. 556-68. doi:https://doi.org/10.1007/978-3-030-50146-4_41.

Roque Rodríguez Outeiral, Nicole Ferreira Silvério, Patrick J. González,
Eva E. Schaake, Tomas Janssen, Uulke A. van der Heide, Rita Simões[*]
Department of Radiation Oncology, The Netherlands Cancer Institute,
Plesmanlaan 121, 1066 CX Amsterdam, the Netherlands

[*] Corresponding author.
E-mail address: r.simoes@nki.nl (R. Simões).