Abstract: Adversarial Examples as Benchmark for Medical Imaging Neural Networks

Magdalini Paschali[1], Sailesh Conjeti[2], Fernando Navarro[1], Nassir Navab[1,3]

[1]Computer Aided Medical Procedures, Technische Universität München, Germany
[2] Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE), Bonn, Germany
[3]Computer Aided Medical Procedures, Johns Hopkins University, USA
magda.paschali@tum.de

Deep learning has been widely adopted as the solution of choice for a plethora of medical imaging applications, due to its state-of-the-art performance and fast deployment. Traditionally, the performance of a deep learning model is evaluated on a test dataset, originating from the same distribution as the training set. This evaluation method provides insight regarding the generalization ability of a model. However, in medical imaging scenarios, especially in cases when a deep learning framework is utilized by a physician for a real-world application, the samples forwarded into the model might belong to a distribution different from the one of the training dataset, or might suffer from noise which cannot usually be modelled by a known distribution, thus raising the need for an evaluation scheme that investigates the robustness of a model, i.e. its performance on data originating from a manifold different from the training one.

To this end, we recently proposed [1] to utilize adversarial examples [2], images that look imperceptibly different from the originals but are consistently missclassified by deep neural networks, as surrogates for extreme test case scenarios, like the ones mentioned above. Extensive evaluation was performed on state-of-the-art classification and segmentation deep neural networks, for the challenging tasks of fine-grained skin lesion classification and whole brain segmentation, leveraging a variety of methods to generate adversarial examples. The results showcased the significant difference in the performance of the utilized networks on clean and on adversarial images. Specifically, networks that performed equally well regarding their generalizability had an astounding 20% difference in robustness, highlighting the need for the proposed, more thorough evaluation technique to uncover which neural network was able to grasp a deeper understanding of the training data and when deployed in real-world applications can showcase a higher robustness to out-of-distribution test samples.

References

1. Paschali M, Conjeti S, Navarro F, et al. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. Proc MICCAI. 2018; p. 493–501.
2. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. Int Conf Learn Representations. 2014;Available from: http://arxiv.org/abs/1312.6199.