Abstract: Interpretable Explanations of Black Box Classifiers Applied on Medical Images by Meaningful Perturbations Using Variational Autoencoders

Hristina Uzunova, Jan Ehrhardt, Timo Kepp, Heinz Handels

Institut für Medizinische Informatik, Universität zu Lübeck
uzunova@imi.uni-luebeck.de

The growing popularity of black box machine learning methods for medical image analysis makes their interpretability to a crucial task. To make a system, e.g. a trained neural network, trustworthy for a clinician, it needs to be able to explain its decisions and predictions. In our work we tackle the problem of explaining the predictions of medical image classifiers, trained to differentiate between different types of pathologies and healthy tissue [1].

There is a variety of neural network explanation methods, such as gradCAMs and guided backpropagation that directly use the learned network weights to deduct the most important image features. However, such methods are based on heuristics and depend on the network architecture. Another intuitive solution to determine which regions of an image influence the trained classifier is to find out whether the classifier changes its prediction when those regions are deleted. This idea is model-agnostic and can be formulated as an explicit minimization problem and thus efficiently implemented on the GPU. However, the meaning of "deletion"of image regions, in our case pathologies in medical images, is not defined. Usually, deleting image regions would be based on image perturbations, but intuitive solutions like replacing the values by zeros or blurring regions, may not have the desired effect for medical applications.

We contribute by defining the deletion of suspicious regions, as the replacement by their healthy looking equivalent generated using a variational autoencoder (VAE). We train the VAE on healthy images only and thus expect it to only be able to reconstruct healthy looking images in test phase even if the input contains pathologies. This healthy reconstruction is then used to perturb the pathological regions. In our tests on retinal OCTs with age-related macular degeneration and brain MRI images with lesions, we show that this perturbation method outperforms other perturbation techniques and shows more robust results compared to heuristic methods.

References

1. Uzunova H, Ehrhardt J, Kepp T, et al. Interpretable explanations of black box classifiers applied on medical images by meaningful perturbations using variational autoencoders. Proc SPIE. 2019;Accepted.