Precise Feature Localization with Explainable AI

Deep neural networks (DNNs) are powerful tools for accurate predictions in various applications and have even shown to be superior to human experts in some domains, for instance for Melanoma detection. However, they are vulnerable to data artifacts, such as band-aids, rulers or skin markers in the Melanoma detection task. In our previous blog post, we demonstrated the application of various model correction approaches to unlearn undesired model behavior and ultimately increase the security of these models.

Automated artifact localization

A challenge for model correction methods requiring prior explanations, e.g., Right for the Right Reasons (RRR) [1] and Contextual Decomposition Explanation Penalization (CDEP) [2], is their dependence on binary masks localizing the source of the undesired behavior, i.e., the data artifact, in input space. The collection of these masks is very labor-intensive and often considered as infeasible. To address that problem, we propose a method for precise feature localization based on Explainable AI. Specifically, once the artifact is discovered via SpRAy or CRP, both yield clusters of samples, which can be used to fit a Concept Activation Vector (CAV) h_l to model the artifact in latent space of a layer l. This represents the direction from artifactual to non-artifactual samples obtained from a linear classifier. This vector can be considered as superposition of concepts encoded in the layer. The location of the concept modeled by the CAV can now be estimated by computing the attribution heatmap with relevances $R_l(x)$ at layer l initialized as:

$$R_l(x) = a_l(x) \circ h_l$$

The resulting attribution scores in input space can be turned into binary masks, precisely localizing the artifacts, as shown in Figure 1. These binary masks, in turn, can be used as prior explanations for model correction methods, such as RRR and CDEP.

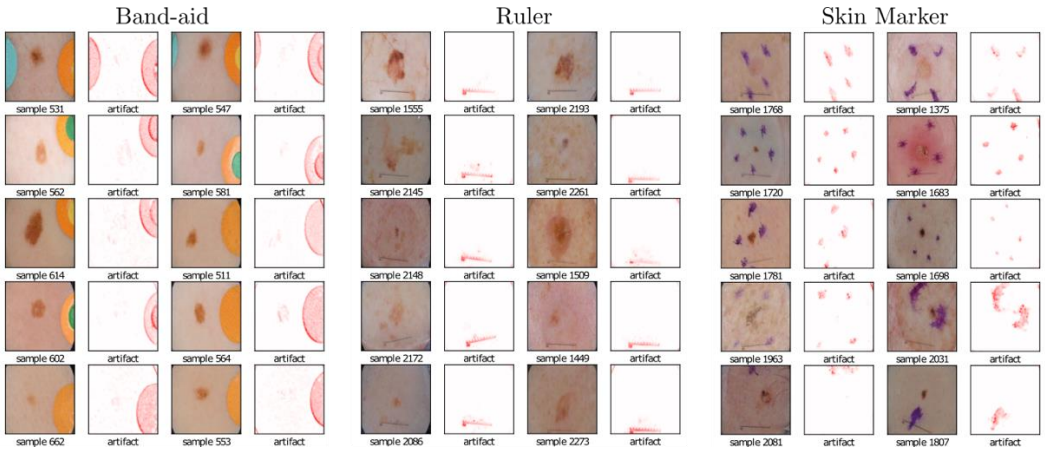

Figure 1: Artifact Heatmaps for localization of band-aid, ruler, and skin marker.

Conclusions

To reduce the manual effort required to apply model correction methods depending on prior explanations, such as RRR and CDEP, we introduced an XAI-based approach for precise artifact localization. This increases the applicability of the approaches, as the lack of ground truth data required for the correction has been a major obstacle. The results are obtained from our recent paper on iterative XAI-based model improvement [3].

Relevance to iToBoS

In iToBoS, many different AI systems will be trained for specific tasks, which in combination will culminate in an “AI Cognitive Assistant”. All those systems will need to be explained with suitable XAI approaches to elucidate all possible and required aspects of the systems’ decision making. Throughout the iToBoS project, we must detect and avoid the usage of data artifacts for model predictions.

Authors

Frederik Pahde, Fraunhofer Heinrich-Hertz-Institute

Maximilian Dreyer, Fraunhofer Heinrich-Hertz-Institute

Sebastian Lapuschkin, Fraunhofer Heinrich-Hertz-Institute

References

[1] Ross, A. S., Hughes, M. C., & Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717.

[2] Rieger, L., Singh, C., Murdoch, W., & Yu, B. (2020, November). Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In International conference on machine learning (pp. 8116-8126). PMLR.

[3] Pahde, F., Dreyer, M., Samek, W., & Lapuschkin, S. (2023). Reveal to Revise: An Explainable AI Life Cycle for Iterative Bias Correction of Deep Models. arXiv preprint arXiv:2303.12641.