Improving AI models with XAI

In the past years, Deep Neural Networks (DNNs) have shown exceptional results across multiple domains, including general image classification tasks, winning Atari games or even detecting skin cancer by lesion classification. These networks are commonly trained on large datasets, such as ImageNet (millions of images of several classes) or ISIC 2019 (images of skin lesions). Unfortunately, these datasets often contain unwanted artifacts (e.g. copyright tag) that have remained undetected during dataset creation. If a given artifact only occurs in one class, this artifact is called Clever Hans (CH) artifact and can cause the model to learn a correlation between the artifact and the class label. This leads to seemingly good model performance in the test lab, but due to right predictions for the wrong reasons . For example, about one-fifth of the images of class “horse” in the Pascal VOC 2007 image categorization dataset (Everingham et al., 2007) contain a copyright tag in the bottom of the image, because they are from the same photographer (Lapuschkin et al. 2016, Lapuschkin et al. 2019). DNNs learn a relationship between the existence of the copyright tag and the class “horse”, which, in fact, is only a CH artifact. Another example for a CH artifact was detected in the ISIC 2019 dataset: The largest class, i.e. *melanocytic nevus*, contains several images with colorful band-aids next to the lesion. Because this artifact — which is unrelated to the classification task — only occurs in one single class, DNNs might learn a relationship upon that spurious correlation and the artifact can be considered as CH artifact.

Clever Hans Artifact Detection with XAI

Explainable AI (XAI) technologies have proven to be a powerful tool to detect cases in which AI models make predictions for the wrong reason. Specifically, attribution methods such as layer-wise relevance propagation (LRP) can be used to compute the attribution of each input value (e.g. pixel value in an image) to the final prediction (Bach et al., 2015, Montavon et al., 2019). These attributions can be visualized for each prediction using heatmaps. Spectral Relevance Analysis (SpRAy) can be used to detect characteristics in model behavior, of which some might be failures based on spurious correlations, such as the two examples described above (Lapuschkin et al., 2019, Anders et al., 2022).

Model Improvement with XAI

Going even further, XAI can also be used to improve models based on artifacts that have been detected. This can be achieved by training an artifact model, which aims to model the expression of a data artifact direction either in the original input space, or in any latent feature representation learned by the model. Recently, Anders et al., 2022 suggested a family of methods called Class Artifact Compensation (ClArC), which, after having identified the artifact, aim to model the artifact direction based on exemplary reference data, and then are using the artifact model to improve neural networks. The full ClArC workflow is visualized in Fig 1.

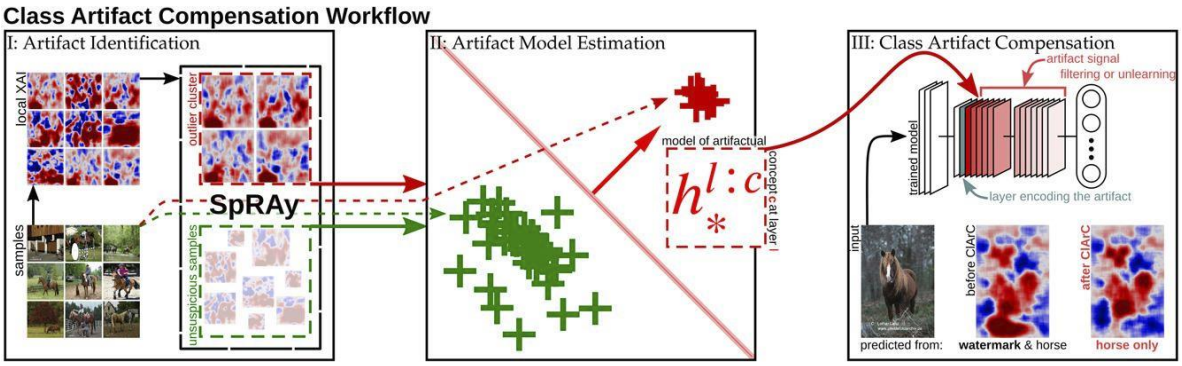

Fig. 1: Visualization of full ClArC workflow, consisting of (I) artifact identification, (II) artifact model estimation and (III) class artifact compensation

Specifically, two methods were developed: (1) Augmentive ClArC (A-ClArC) uses the artifact model to augment a random subset of training samples with the learnt artifact in order to desensitize the model to the artifact's expression. For example, in the context of the ISIC 2019 case, A-ClArC would add the colorful band-aid to some random training samples. Finetuning the model on the augmented training set forces the model to learn a representation that is invariant to the artifact. However, due to the size of today's AI models and potentially limited access to training samples (e.g. due to data privacy), finetuning can be too expensive or even impossible. Therefore, (2) Projective ClArC (P-ClArC) was developed, which uses the artifact model at test time to remove all information related to the artifact from test samples. This can be implemented as a simple linear layer in a DNN and does not require any additional training. While it does not allow the model to learn a new representation, P-ClArC implicitly "cleans" the sample from any features related to the artifact. The intuition behind A-ClArC and P-ClArC using a toy example is shown in Fig. 2.

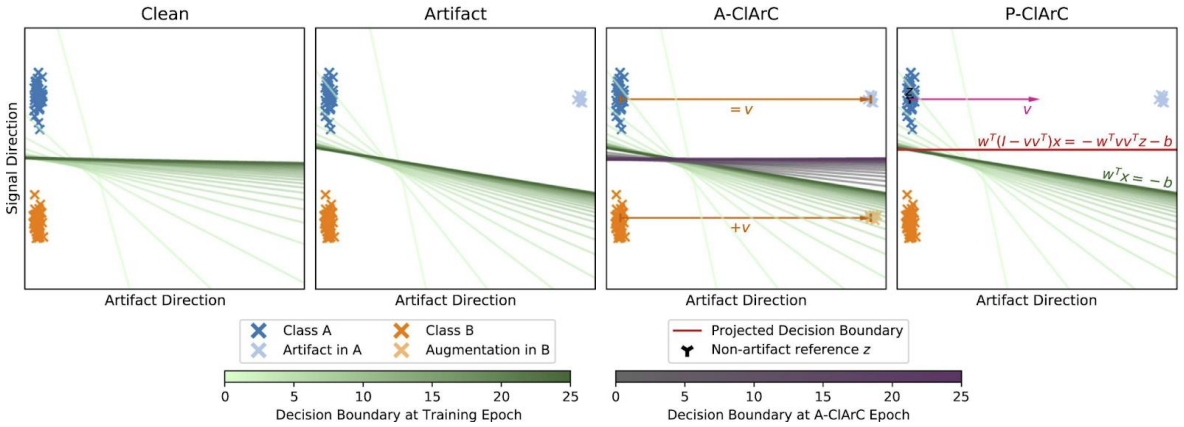

Fig 2: ClArC on a toy example with 2 classes, where an artifact was added to one class. From left to right: (1) Decision boundary for classification task with 2 classes on "clean" data. (2) Decision boundary for data with CH artifact (light blue) in one class. (3) A-ClArC: the artifact direction ***v*** is applied to some samples from the orange class. The decision boundary for a finetuned model is invariant to artifact direction. (4) P-ClArC: without finetuning, the decision boundary is updated by cleaning the test samples based on artifact direction.

The paper has shown that ClArC is able to "unlearn" information based on CH artifacts that were present in the training data. This leads to models that are less sensitive to changes in the test data,

as they depend on true features, rather than on dataset-specific CH artifacts and are expected to perform better in the real world. At Fraunhofer Heinrich Hertz Institute (FHHI), we are currently working on more precise artifact modeling, which is key in order to successfully apply ClArC methods.

Relevance to iToBoS

In iToBoS we are using XAI techniques to ensure the AI models are making decisions (e.g. cancer prediction) for the right reason. Specifically, we want to use XAI methods in conjunction with the doctor's judgment to (1) detect CH artifacts in the training data (e.g. colorful band-aids as in ISIC 2019) and (2) to improve the performance of AI models with ClArC methods.

Literature

Anders, Christopher J., et al. "Finding and removing clever hans: Using explanation methods to debug and improve deep models." *Information Fusion* 77 (2022): 261-295.

Bach, Sebastian, et al. "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation." *PloS one* 10.7 (2015): e0130140

Everingham, Mark, et al. "The pascal visual object classes (voc) challenge." *International journal of computer vision* 88.2 (2010): 303-338.

Lapuschkin, Sebastian, et al. "Analyzing classifiers: Fisher vectors and deep neural networks." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

Lapuschkin, Sebastian, et al. "Unmasking Clever Hans predictors and assessing what machines really learn." *Nature communications* 10.1 (2019): 1-8.

Montavon, Grégoire, et al. "Layer-wise relevance propagation: an overview." *Explainable AI: interpreting, explaining and visualizing deep learning* (2019): 193-209

Authors

Frederik Pahde, FHHI
Sebastian Lapuschkin, FFHI