XAI for Analyzing and Unlearning Spurious Correlations in ImageNet

Christopher J. Anders [1] **David Neumann** [2] **Talmaj Marinc** [2] **Wojciech Samek** [2] **Klaus-Robert Müller** [1 3 4]
Sebastian Lapuschkin [2]

Abstract

Contemporary learning models for computer vision are typically trained on very large data sets with millions of samples. There may, however, be biases, artifacts, or errors in the data that have gone unnoticed and are exploitable by the model, which in turn becomes a biased 'Clever-Hans' predictor. In this paper, we contribute by providing a comprehensive analysis framework based on a scalable statistical analysis of attributions from explanation methods for large data corpora, here ImageNet. Based on Spectral Relevance Analysis we propose the following technical contributions and resulting findings: (a) a scalable quantification of artifactual classes where the ML models under study exhibit Clever-Hans behavior, (b) an approach denoted as Class-Artifact Compensation (ClArC) that allows to fine-tune an existing model to effectively eliminate its focus on artifacts and biases yielding significantly reduced Clever-Hans behavior.

1. Introduction

Throughout the last decade, Deep Neural Networks (DNN) have enabled impressive performance leaps on even the most complex tasks (LeCun et al., 2015; Mnih et al., 2015; Silver et al., 2016; Schütt et al., 2017; Krizhevsky et al., 2012). These models are typically (pre-)trained on very large datasets, e.g., ImageNet (Russakovsky et al., 2015), with millions of samples. Recently, it was discovered that biases, spurious correlations, as well as errors in the training dataset (Stock & Cissé, 2018) may have a detrimental effect on the training and/or result in "Clever-Hans" predictors (Pfungst, 1911; Lapuschkin et al., 2019), which only superficially solve the task they have been trained for[1]. Unfortunately, due to the immense size of today's datasets, a direct manual inspection and removal of artifactual samples can be regarded hopeless. However, analyzing the biases and artifacts in the *model* instead may provide insights about the training data indirectly. This however requires an inspection of the learning models beyond black box mode.

Only recently methods of explainable AI (XAI) (cf. (Samek et al., 2019) for an overview) were developed. They provide deeper insights into how an ML classifier arrives at its decisions and potentially help to unmask Clever-Hans predictors. XAI methods can be roughly categorized into two groups: methods providing *local* (e.g. (Bach et al., 2015; Selvaraju et al., 2017; Sundararajan et al., 2017; Shrikumar et al., 2017; Ribeiro et al., 2016; Zintgraf et al., 2017; Fong & Vedaldi, 2017)) explanations and those providing *global* (e.g. (Guyon & Elisseeff, 2003; Kim et al., 2018; Rajalingham et al., 2018)) explanations (Lundberg et al., 2019). Current approaches are of limited use when scaling the search for biases, spurious correlations, and errors in the training data set as that would require intense 'semantic' human labor. A recent technique, the Spectral Relevance Analysis (Lapuschkin et al., 2019) (SpRAy), aims to bridge the gap between local and global XAI approaches, by introducing automation into the analysis of large sets of local explanations, however still involves a considerable amount of manual analyses, especially in context of contemporary data sets with high numbers of classes and samples, such as ImageNet (Russakovsky et al., 2015).

In this paper, we propose (a) an extension to SpRAy, enabling large-scale analyses on data sets with hundreds of classes and millions of samples, for semi-automated dis-

[1]Machine Learning Group, Technische Universität Berlin, Germany [2]Video Coding & Analytics, Fraunhofer Heinrich Hertz Institut [3]Max-Planck-Institut für Informatik, Saarbrücken, Germany [4]Department of Brain and Cognitive Engineering, Korea University, Seoul, Korea. Correspondence to: Sebastian Lapuschkin <sebastian.lapuschkin@hhi.fraunhofer.de>, Klaus-Robert Müller <klaus-robert.mueller@tu-berlin.de>, Wojciech Samek <wojciech.samek@hhi.fraunhofer.de>.

Proceedings of the 37^{th} International Conference on Machine Learning, Vienna, Austria, PMLR 119, 2020.

[1]Clever Hans was a horse from Berlin that allegedly could do math – a media sensation from early 1900. Later in 1907 it was discovered that Hans would read the examinator's body language instead, and in this manner give the right answer but for the wrong reason, https://en.wikipedia.org/wiki/Clever_Hans. "Clever Hans strategies" for neural networks (Lapuschkin et al., 2019) are accordingly named as a homage to this infamous horse.

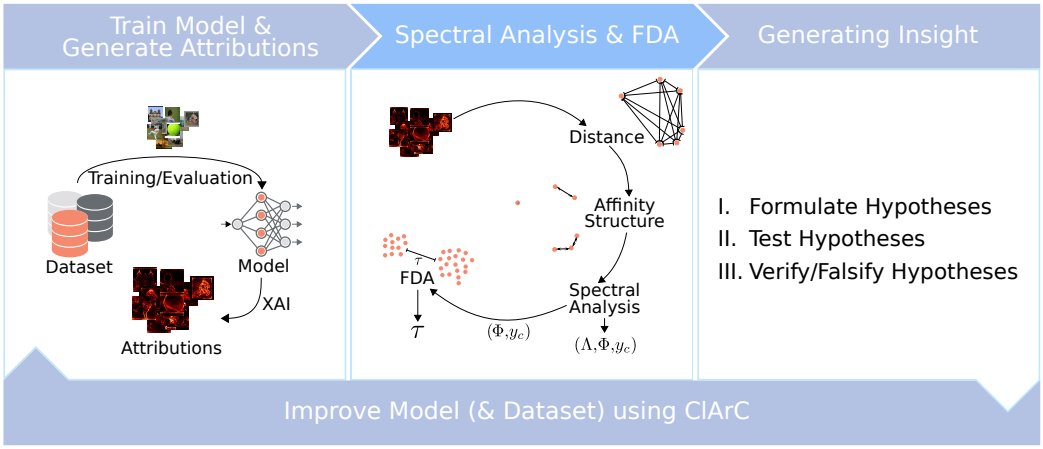

Figure 1. Overview of the SpRAy approach. *Left:* Large corpora of data can be used to train models for specific tasks. To gain insights into local model behavior, explanation methods can be employed. *Middle:* Using SpRAy, one can deduce global model behavior from a set of local explanations (see Algorithm 1 in the Supplement). *Right:* Based on this analysis, *striking* classification strategies can be identified and further investigated. Obtained insights can be used to improve the model and/or the dataset, e.g. using ClArC.

covery of undesirable Clever-Hans effects that are embedded into data and model. In addition, we provide (b) a novel framework denoted as Class-Artifact Compensation (ClArC), giving an intuition for Clever-Hans artifacts and their removal from a trained model. In this manner, we provide (c) a well-controlled quantitative strategy to detect, validate and remove such artifacts which we showcase for the ImageNet data corpus. These analyses allow interesting findings that are illuminating beyond our specific technical approach.

2. Methods

First we will discuss the ingredients for ClArC, namely, Spectral Relevance Analysis (SpRAy) (Lapuschkin et al., 2019), Fisher Discriminant Analysis (Fisher, 1936; Fukunaga, 1990), an intuition for Clever-Hans artifacts and based on that and a procedure to *remove* the influence of Clever-Hans artifacts from the respective classes.

2.1. Spectral Relevance Analysis

The SpRAy (Lapuschkin et al., 2019) is a meta-analysis tool for finding patterns in model behavior, given sets of instance-based explanatory attribution maps. The SpRAy algorithm has its core in Spectral Clustering (SC) (Meila & Shi, 2001; Ng et al., 2002) and — via the use of attribution maps as input — enables the analysis of the input data from the model's perspective for finding (hidden) characteristics of specific classes, as exploited by the model. As output, SpRAy yields a spectral embedding Φ of the input data and the spectrum of (eigen)values $\Lambda = \{\lambda_i\}_{i=1\ldots q}$, which is used to analyze the structure of clusters (i.e. cluster number and nesting) discovered in the data, via the eigen- or spectral gap (von Luxburg, 2007), or to rank a set of analyzed classes w.r.t. to their *potential* for exhibiting Clever-Hans phenomena (Lapuschkin et al., 2019). Due to the direct correspondence of the given inputs to (the colums of) Φ, we use the embedding for computing visualizations in R^2, e.g. via t-SNE (Maaten & Hinton, 2008) or UMAP (McInnes & Healy, 2018). An algorithmic summary can be found in Algorithm 1 in the Supplement.

2.2. Fisher Discriminant Analysis for Clever-Hans Identification

A critical decision in clustering approaches is the number of desired clusters. While for small datasets like Pascal VOC (Everingham et al., 2007) it suffices to analyze the per-class eigen-spectrum (Lapuschkin et al., 2019); datasets with a large number of classes cannot be feasibly analyzed by manual comparison and ranking of the eigen-spectra of all classes to identify those exhibiting spurious model behavior. In order to automate this process, we propose Fisher Discriminant Analysis (FDA) to rank all class-wise clusterings by their respective (linear) separability. FDA (Fisher, 1936; Fukunaga, 1990) is a widely popular method for classification as well as class- (or cluster-) structure preserving dimensionality reduction. FDA finds an embedding space by maximizing between-class scatter $S^{(b)}$ and minimizing within-class scatter $S^{(w)}$, given by

$$S^{(w)} = \sum_{k=1}^{K} \sum_{x_i \in \mathbf{c}_k^K} (x_i - \mu_k)(x_i - \mu_k)^\top \tag{1}$$

$$S^{(b)} = \sum_{k=1}^{K} (\mu_k - \mu)(\mu_k - \mu)^\top. \tag{2}$$

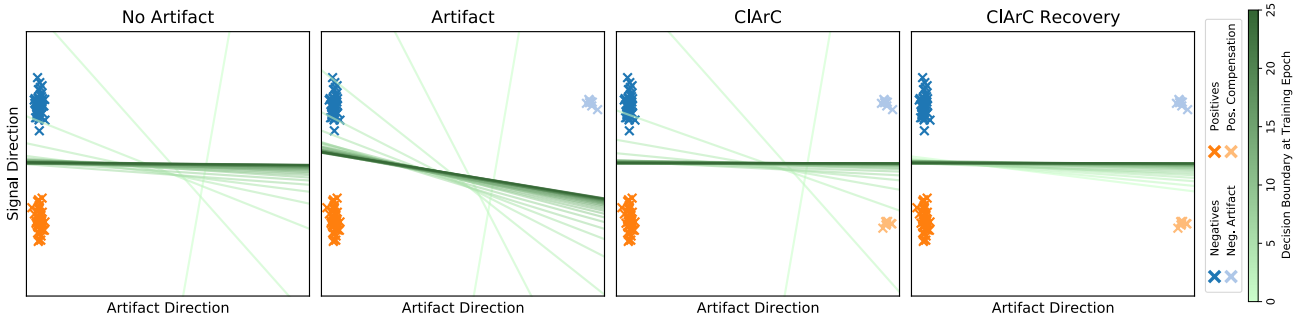

Figure 2. Logistic regression on data with, among possibly others, a discriminative *signal* direction and an *artifact* direction which is only represented in one of the two classes. The decision-hyperplane is shown over the SGD-based training-process of 25 epochs in shades of green, with: **No Artifact**: no artifact in the data; **Artifact**: a Clever-Hans artifact in the negative class (blue); **ClArC**: the previous and the artifact-direction added to some samples of the positive class (orange); **ClArC Recovery**: the previous, but continuing the training after **Artifact**. The introduction of an artifact to positive samples changes the decision boundary. By introducing the same artifact direction to negatives samples, this effect can be reduced significantly.

Here, $\mathbf{C}^K$ is a clustering with K clusters $\mathbf{c}_k^K$ with $k \in \{1, \dots, K\}$, μ_k the sample mean of cluster k and μ the mean over the whole set of samples. The solution of FDA can be understood as directions of maximal separability between clusterings, and, when normalized and plugged into the original objective, gives scores of separability $R(\mathbf{C}^K)$. In our specific use-case, for each class, we compute separability scores $R(\mathbf{C}^K)$ on the spectral embedding Φ and each clustering $\mathbf{C}^K$ in a set of clusterings $\{\mathbf{C}^K\}$. We then compute the *class*-separability score τ as

$$\tau = \frac{1}{|\{\mathbf{C}^K\}|} \sum_{\mathbf{C}^K} R(\mathbf{C}^K). \tag{3}$$

2.3. Class-Artifact Compensation (ClArC)

Let us consider a toy model based on logistic regression to better grasp the influence Clever-Hans artifacts on models trained with stochastic gradient descent. Intuitively, Clever-Hans artifacts can be described as *directions* in the data space, e.g. a watermark on some pixels of an image, which manifest as *shifts* along artifact directions in latent space. Figure 2 shows SGD-training over 25 epochs with, possibly among others, a signal direction and an artifact direction in the data.

The *No Artifact* setting shows the case where there are no artifacts in the data, such that the two classes, positive and negative, are classified only by using the signal direction, e.g. the decision boundary is approximately perpendicular to the signal direction. We now introduce artifactual features to some negative examples, as shown in the *Artifact* setting in Figure 2. This visibly rotates the decision boundary, such that the model now also uses information along the artifact direction. This means that, even though the artifact direction is (by design) not intended to correlate with the label, data points with the positive label and an artifact may be falsely predicted as negatives by the model. To balance this effect out, we can isolate the artifact direction, and add it to some of the positive samples, which is shown in Figure 2 *ClArC*. We call this approach *Class-Artifact Compensation* (ClArC), and can see that the decision boundary rotates back to a direction orthogonal to the signal, i.e. it returns to ignoring the artifact direction. The *ClArC Recovery* setting shows that this training modification can be used for fine-tuning models which were previously trained on data containing class-limited artifacts. We use this approach in Section 3.3 to unlearn artifacts identified using SpRAy.

3. Experiments and Evaluations

In this section, we apply our ClArC framework to the ImageNet dataset. While here results based on LRP (Bach et al., 2015) and our extended SpRAy are shown, the procedure is also readily applicable to all members of the XAI zoo (cf. (Samek et al., 2019)); results based on SmoothGrad (Smilkov et al., 2017) can be found in the Supplement.

3.1. Identifying Clever-Hans Candidates with FDA

As SpRAy in its original form still contained manual processing steps, our proposed algorithm allows to scale to many classes and samples. For this we apply FDA on each class' spectral embedding Φ. We report the respective cluster separability scores τ (Eq. (3)). While clearly algorithmic alternatives to FDA are conceivable, τ quantifies simply and intuitively how much different class specific classification strategies are. Large τ denotes outlierness in problem solving — solid indicators for Clever-Hans candidates (Lapuschkin et al., 2019) — whereas low τ does not indicate any strikingly "irregular" prediction behavior. Figure 4 lists a ranking of the ImageNet classes with the highest and

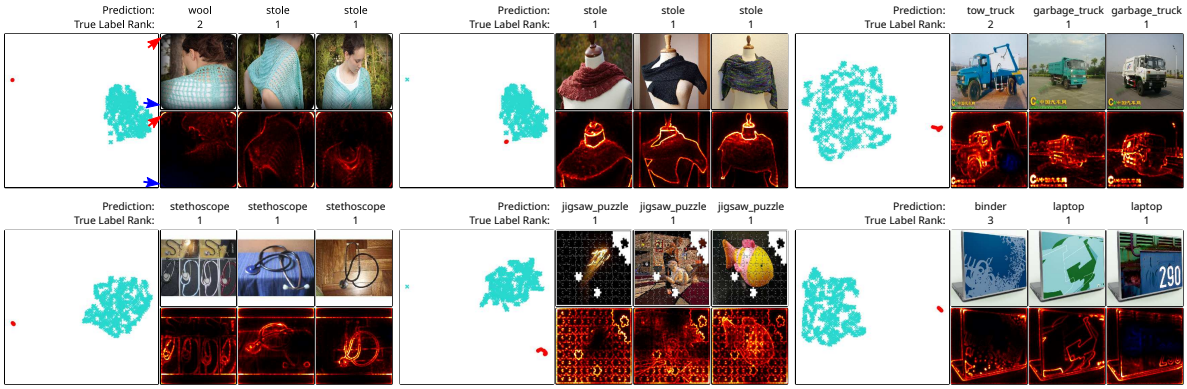

Figure 3. Each panel shows the UMAP (left) with samples and heatmaps (right) of significant clusters, highly separated from the rest of the samples. For each class, some images and their respective attributions from the identified cluster are shown. Red dots in the UMAPs identify the clusters the samples to the right have been grouped into. Relevance maps to the right are color coded to identify relevant image regions *supporting* the classifier decision in hot colors (red to yellow) irrelevant regions in black color and relevant regions *contradicting* the final prediction in cold (blue to cyan) hues. The text above the sample images shows the classifier's top-1 predicted class, and the prediction rank of the true label.

lowest τ values with a striking result for class laptop, due to a large cluster with copies of almost the same image (see UMAP of its spectral embedding with a significant cluster is depicted in Figure 3 (bottom right)).

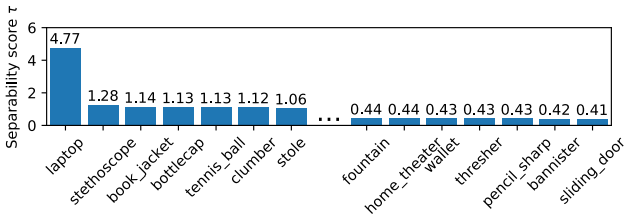

Figure 4. Mean separability score τ of spectral embedding of attributions based on Fisher Discriminant Analysis. A high τ means there are significantly different decision strategies being used, potentially of Clever-Hans type.

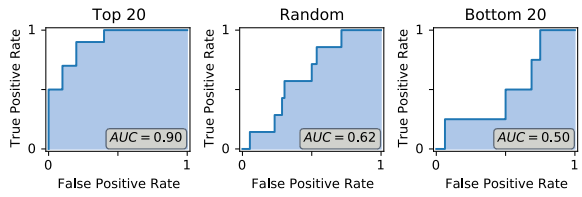

Figure 5. ROC-Curves for artifact-existence versus FDA-Ranking. Left: Top 20 classes with highest values of τ. Mid: 63 random classes with any values of τ. Right: Bottom 20 classes with lowest values of τ.

We inspect the validity of the class ranking for Clever Hans candidates generated by FDA in a small experiment, by screening a subset of all 1000 ImageNet classes, namely (1) those with the 20 highest τ scores, (2) those with the 20 lowest τ scores and (3) 63 randomly picked classes. In all three cases, we assume a positive Clever-Hans "prediction" per class due to a large value of τ. We then produce "ground truth" labels via manual assessment of the existence of a Clever-Hans candidate. Using this information we produce receiver operator characteristic (ROC) curves and corresponding area under the curve (AUC) values.

The results show a clear picture validating that a high τ score is indeed a strong indicator for the presence of Clever-Hans phenomena (Figure 5 (left), high AUC). Both randomly selected or bottom 20 classes (Figure 5 (mid, right)) yield essentially random AUC scores due to only sporadically encountered Clever-Hanses. However, the AUC $\gg 0$ here also show that even a τ rating in the lowest 2-percentile does not guarantee a class to be free of Clever-Hans behavior. Summarizing, large τ is an excellent indicator for Clever-Hans behavior, but small τ is no ultimate guarantee for their absence, so further research will be needed here to ideally bring forward indicators that can provide a theoretical bound for absence of Clever-Hans behavior.

3.2. Inspecting and Isolating Clever-Hans Candidates

Based on the ordering by FDA and τ established in the previous section, we will now manually investigate whether the Clever-Hans candidate classes show indeed the prominent Clever-Hans artifacts to be expected. The SpRAy framework provides as a side effect (through its spectral embedding space Φ) also a basis for visualizing clusters of heatmaps, here we use UMAP. Promising clusters are often located far away from the rest of datapoints in the UMAP embedding, see e.g. Figure 3 top right the UMAP scatter-plot of class "garbage truck", where, the red cluster-members all show examples of images of the same water-

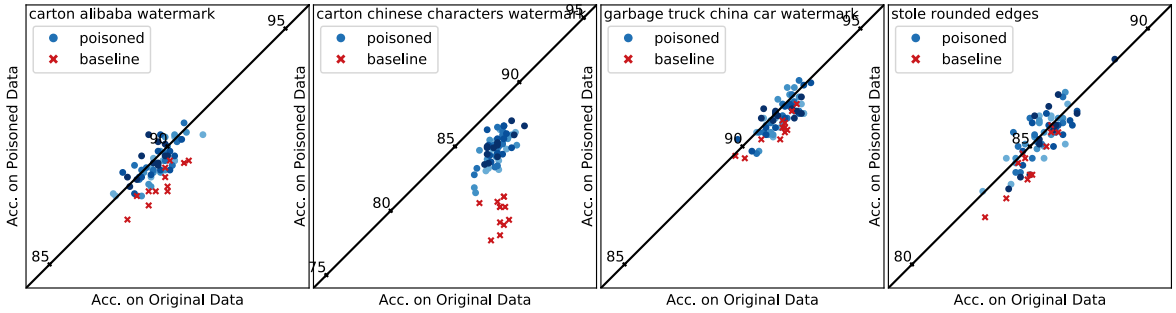

Figure 6. Original (x-axis) vs. poisoned (y-axis) validation set accuracy of baseline model (red) and ClArC'ed models (blue) over training poison rate (color shade, darker is higher). All models show a tendency towards the bottom-right, meaning the addition of artifact candidates degrades the model performance, thus validating their Clever-Hans property. ClArC'ed models (blue) show better performance on the poisoned validation set, implying increased robustness against Clever-Hans artifacts. A slight tendency towards the right for ClArC'ed models can be observed, suggesting an overall better generalization.

mark with high attribution in LRP. Another intriguing example is the top middle UMAP plot of class "stole", where, while not as separated as for other examples, we find a cluster of mannequins wearing stoles, with high attribution scores on the mannequin's head. For further Clever-Hans candidates, using other models and attribution methods, see Supplement.

To test whether Clever-Hans candidate clusters indeed influence the model, we first isolate the artifact v and then blend the artifact onto the image to stay within pixel-domain and achieve better results. Specifically, we compute the pixel-wise mean over all affected samples c_i as RGB information $v = \frac{1}{C}\sum_i^C c_i$ and use an alpha channel inversely proportional to the pixel-wise standard deviation $\alpha \propto \sqrt{\frac{1}{C}\sum_i^C (c - c_i)^2}^{-1}$. For more elaborate artifacts like the mannequin head, it is also possible to use a manual approach to extract v and α. The isolated artifact v can then be applied freely to other samples p_i with

$$\bar{c}_i = (1 - \alpha) \odot p_i + \alpha \odot v \tag{4}$$

where $\odot$ is the element-wise product, and the Clever-Hans effect can be compensated accordingly; see next section for a comprehensive validation using ClArC.

3.3. Validating and Un-Hans'ing the Model by Class-Artifact Compensation

Setup We will now apply ClArC for each detected Clever-Hans candidate artifact for a certain class: In addition to the artifact candidate class, we choose 19 other random classes from the ImageNet training set. Then we fine-tune a pre-trained model, here VGG16 (Simonyan & Zisserman, 2014), for 60 training epochs during which we add the artifact candidate following ClArC (see Eq.(4)), to each sample of the non-candidate classes with probability p, where $p \in \{0, 0.1, \ldots, 0.5\}$. The setup $p = 0$ serves as a baseline, where no artifact is added during training. Subsequently we prepare two validation sets, each only of the involved 20 classes. One is based on the original ImageNet validation set, the other is a *poisoned* version of it, where we add the artifact to 100% of the samples. The underlying *hypotheses* is the following: cleaning the data with ClArC and subsequent fine tuning of the VGG16 model will make the model disregard the Clever Hans strategy. We therefore expect the CLArC'ed models to exhibit stable generalization performance even if the artifact is not present. Conversely, the unchanged model is expected to show (due to its Clever Hans strategy) a drop in performance from the original to the poisoned version of the ImageNet validation set.

Validation Figure 6 shows scatter plots per class examples of all model accuracies, with original validation set results (x-axis) versus results on the *poisoned* validation set (y-axis). For all training setups, including the baseline, models show a significant tendency towards the bottom-right of the plot. This indicates that the Clever-Hans strategy was correctly identified as we observe a better performance on the original validation set (giving the correct answer for the 'wrong' Clever- Hans artifact-driven reason). If this artifactual clue is removed from the data (by using the poisoned validation set), then the prediction error of the CLArC'ed model remains virtually unchanged whereas the baseline model using a Clever Hans strategy shows significantly increased errors. This validates the Clever-Hans candidates as actual Clever-Hans strategies (ab)used by the model and confirms the above hypothesis.

Un'hansing By further inspection of the validation accuracies in Figure 6, we can observe a clear general trend of the ClArC'ed model's poisoned validation accuracies (blue) above the baseline model's poisoned validation accuracies (red). Furthermore, there is no indication of an

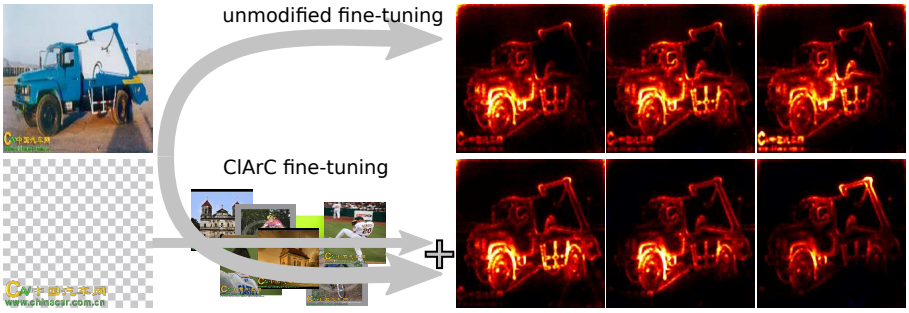

Figure 7. Left: Input-sample affected by an identified Clever-Hans artifact, with programmatically isolated artifact below. *Right:* Attribution maps illustrating the model's use of input features after 1, 5, and 10 (left to right) epochs of unmodified (top) and *ClArC* (bottom) fine-tuning in direct comparison.

inferior performance of ClArC'ed models compared to the baseline models on the original validation set as all point lie below the diagonal. Instead, we can even see a slightly increased performance. This signifies not only an increased robustness against respective Clever-Hans artifacts, but interestingly suggests an overall better generalization of the ClArC'ed model.

By inspecting and comparing model attributions over the training iterations given in Figure 7, we can observe a clear focus on the artifact (watermark) which is being fully eliminated in the ClArC'ed model, as opposed to the essentially unchanged baseline model. We can even observe the ClArC'ed model in the same instance increasing its focus on the cargo-container of the garbage truck, suggesting that the model could achieve a better understanding of the object itself. Additional experiments and observations with various artifacts on different models with analogous outcomes can be found in the supplement.

4. Conclusion

Deep Learning models have gained high practical usability by pre-training on large corpora and then reusing the learned representation for transferring to novel related data. A prerequisite for this practice is the availability of large corpora of rather standardized and, most importantly, representative data. If artifacts or biases are present in data corpora, then the representations formed are prone to inherit these flaws. This is clearly to be avoided, however, it requires either clean data or detection and subsequent removal of the influence of artifacts, biases etc. of data bases that would cause dysfunctional representation learning. In this paper we have used explanation methods (e.g. LRP (Bach et al., 2015) and SmoothGrad (Smilkov et al., 2017), for an overview see (Samek et al., 2019; Montavon et al., 2018)) and introduced the ClArC framework to scalably and automatically detect, validate and alleviate Clever Hans behaviour in the ImageNet corpus. While we mainly used LRP and SmoothGrad (see Supplement), the proposed ClArC framework is independent of the particular XAI method. ClArC encompasses a first simple intuitive model of how artifacts may harm generalization. As this intuitive model is based on logistic regression, it is rather crude, but it already shows the main effects caused by artifacts: deterioration of generalization ability. For neural networks it may, however, still serve as a reasonable guideline and indeed our large-scale experiments on ImageNet show analogous effects, that can exhibit a dramatic drop of generalization for some classes (see Fig. 6). Interestingly un-Hansing is shown to provide uniformly better generalization ability.

Let us reiterate that without removing, or at least considering such data artifacts, learning models are prone to adopt Clever-Hans strategies (Lapuschkin et al., 2019), thus, giving the correct prediction for an artifactual/wrong reason. Once these artifacts are absent in the wild such Clever-Hans models will experience significant loss in generalization (see Fig. 6). This makes them especially vulnerable to adversarial attacks that can harvest all such artifactual issues in a data corpus (Carlini & Wagner, 2017).

Future work will therefore focus on the important intersection between security and functional cleaning of data corpora, e.g., to lower the attack risk when building on top of pre-trained models. In addition we will explore improvements in detecting potentially compromised classes beyond FDA.

Acknowledgement

We thank the reviewer for their constructive feedback. We acknowledge Pan Kessel for invaluable discussions. This work was supported in part by the German Ministry for Education and Research (BMBF) under Grants 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A and 01IS18037A. This work is also supported by the Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (No. 2017-0-001779),

as well as by the Research Training Group "Differential Equation- and Data-driven Models in Life Sciences and Fluid Dynamics (DAEDALUS)" (GRK 2433) and Grant Math+, EXC 2046/1, Project ID 390685689 both funded by the German Research Foundation (DFG).

References

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE*, 10(7):e0130140, 2015. doi: 10.1371/journal.pone.0130140. URL http://dx.doi.org/10.1371/journal.pone.0130140.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. IEEE, 2017.

Cuturi, M. and Doucet, A. Fast computation of wasserstein barycenters. In *International Conference on Machine Learning (ICML)*, pp. 685–693, 2014.

Everingham, M., Gool, L., Williams, C., Winn, J., and Zisserman, A. The pascal visual object classes challenge results. *URL: http://host.robots.ox.ac.uk/pascal/VOC/voc2007/workshop/everingham_cls.pdf*, 2007.

Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

Fong, R. C. and Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 3449–3457, 2017. doi: 10.1109/ICCV.2017.371. URL https://doi.org/10.1109/ICCV.2017.371.

Fukunaga, K. Chapter 1 - introduction. In *Introduction to statistical pattern recognition*. Academic Press Professional, Inc., Boston, 1990. ISBN 978-0-08-047865-4.

Guyon, I. and Elisseeff, A. An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182, 2003.

Huang, G., Liu, Z., van der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243. URL https://doi.org/10.1109/CVPR.2017.243.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viégas, F. B., and Sayres, R. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proc. of International Conference on Machine Learning (ICML)*, pp. 2673–2682, 2018. URL http://proceedings.mlr.press/v80/kim18d.html.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagennet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1097–1105, 2012.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096, 2019. doi: 10.1038/s41467-019-08987-4. URL http://dx.doi.org/10.1038/s41467-019-08987-4.

LeCun, Y. The mnist database of handwritten digits. http://yann.lecun.com/exdb/mnist/, 1998.

LeCun, Y., Bengio, Y., and Hinton, G. E. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539. URL https://doi.org/10.1038/nature14539.

Lundberg, S. M., Erion, G. G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S. Explainable AI for trees: From local explanations to global understanding. *CoRR*, abs/1905.04610, 2019. URL http://arxiv.org/abs/1905.04610.

Maaten, L. v. d. and Hinton, G. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov): 2579–2605, 2008.

McInnes, L. and Healy, J. UMAP: uniform manifold approximation and projection for dimension reduction. *CoRR*, abs/1802.03426, 2018. URL http://arxiv.org/abs/1802.03426.

Meila, M. and Shi, J. A random walks view of spectral segmentation. In *Proceedings of the Eighth International Workshop on Artificial Intelligence and Statistics, AISTATS 2001, Key West, Florida, US, January 4-7, 2001*, 2001. URL http://www.gatsby.ucl.ac.uk/aistats/aistats2001/files/meila177.ps.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., et al. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

Montavon, G., Samek, W., and Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, 2018.

Ng, A. Y., Jordan, M. I., and Weiss, Y. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pp. 849–856, 2002.

Peyré, G., Cuturi, M., and Solomon, J. Gromov-wasserstein averaging of kernel and distance matrices. In *Proc. of International Conference on Machine Learning (ICML)*, pp. 2664–2672, 2016. URL http://proceedings.mlr.press/v48/peyre16.html.

Pfungst, O. *Clever Hans:(the horse of Mr. Von Osten.) a contribution to experimental animal and human psychology*. Holt, Rinehart and Winston, 1911.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., and DiCarlo, J. J. Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, 38(33):7255–7269, 2018.

Ribeiro, M. T., Singh, S., and Guestrin, C. 'why should I trust you?': Explaining the predictions of any classifier. In *Proc. of ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 1135–1144, 2016. doi: 10.1145/2939672.2939778. URL http://doi.acm.org/10.1145/2939672.2939778.

Rubner, Y., Tomasi, C., and Guibas, L. J. A metric for distributions with applications to image databases. In *Procedings of the Sixth International Conference on Computer Vision (ICCV-98), Bombay, India, January 4-7, 1998*, pp. 59–66, 1998. doi: 10.1109/ICCV.1998.710701. URL https://doi.org/10.1109/ICCV.1998.710701.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3): 211–252, 2015.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller (Eds.), K.-R. Explainable AI: Interpreting, explaining and visualizing deep learning. *Springer LNCS 11700*, 2019.

Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K.-R., and Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nature Communications*, 8:13890, 2017.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proc. of IEEE International Conference on Computer Vision (ICCV)*, pp. 618–626, 2017. doi: 10.1109/ICCV.2017.74. URL https://doi.org/10.1109/ICCV.2017.74.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *Proc. of International Conference on Machine Learning (ICML)*, pp. 3145–3153, 2017. URL http://proceedings.mlr.press/v70/shrikumar17a.html.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529 (7587):484–489, 2016.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL http://arxiv.org/abs/1409.1556.

Smilkov, D., Thorat, N., Kim, B., Viégas, F. B., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017. URL http://arxiv.org/abs/1706.03825.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. J. Convolutional wasserstein distances: efficient optimal transportation on geometric domains. *ACM Trans. Graph.*, 34(4):66:1–66:11, 2015. doi: 10.1145/2766963. URL https://doi.org/10.1145/2766963.

Stock, P. and Cissé, M. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, pp. 504–519, 2018. doi: 10.1007/978-3-030-01231-1_31. URL https://doi.org/10.1007/978-3-030-01231-1_31.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *Proc. International Conference on Machine Learning (ICML)*, pp. 3319–3328. JMLR.org, 2017.

von Luxburg, U. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

Zintgraf, L. M., Cohen, T. S., Adel, T., and Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. In *Proc. of International Conference on Learning Representations (ICLR)*, 2017.

5. Supplement

We state the pipeline of our experiments using pseudo-code in Section 5.1. Furthermore, we complement our findings from our manuscript by analyzing artifacts on different architectures in Section 5.2, showing additional results for ClArC'ed model performance in Section 5.3 with qualitative attribution analysis in 5.4, more Clever-Hans candidates with LRP in Section 5.5 and with SmoothGrad in Section 5.6, analyze qualitatively and quantitatively the effects of adding and removing artifacts to samples in Section 5.7 and present the observation that the same artifacts can (obviously) also be found in the original validation set in Section 5.8. Finally, we explore using different distance metrics, i.e. Wasserstein Distance, with SpRAy in Section 5.9.

5.1. Algorithms for Experiments in Section 3

Algorithm 1: Extended Spectral Relevance Analysis

Data: Input class y,
Training data set $X_y = \{x_1, x_2, ..., x_i\}$ with samples from class y,
Model f operating on X_y
Result: Eigenvalues $\Lambda = \{\lambda\}$,
Spectral embeddings $\Phi \in \mathrm{R}^{n \times q}$,
Clusterings $\mathbb{K}$,
Mean separability score τ,
Visualization embeddings $V \in \mathrm{R}^2$

/* compute attributions for $x \in X_y$, e.g. LRP */
$R = \{\}$;
for $x \in X_y$ **do**
  $R_x = \text{attribution}(f, x)$;
  $R.\text{append}(R_x)$;
end
/* Spectral Relevance Analysis */
$\Phi, \Lambda, \mathbb{K} = \text{SpRAy}(R)$;
/* Compute separability scores given by Fisher Discriminant Analysis */
for $\mathbf{C} \in \mathbb{K}$ **do**
  $S_{\mathbf{C}} = \text{FDA}(\Phi, \mathbf{C})$;
end
/* Compute mean separability score [Eq. (3)] */
$\tau = \frac{1}{|\mathbb{K}|} \sum_{\mathbf{C} \in \mathbb{K}} S_{\mathbf{C}}$;
/* Compute visualizations for the embedding, e.g. t-SNE, UMAP, etc. */
$V = \text{visualize\_embedding}(\Phi)$;
return $\Lambda, \Phi, \mathbb{K}, \tau, V$

Algorithm 2: ClArC Un-hans'ing of a model.

Data: Training data set
$X = \{(x_1, y_1), (x_2, y_2), ..., (x_i, y_i)\}$,
Model f operating on X,
Number of epochs n_e,
Learning rate η,
Artifacts $A = \{a_1, a_2, ..., a_j\}$,
Poison rates p.
Result: Un-hans'ed model f'

/* Un-hans the model f */
$f' \leftarrow f$;
for $e \in 1..n_e$ **do**
  for $(x, y) \in X$ **do**
    /* Apply [Eq. (4)] all artifacts a with poison rate p */
    $x' \leftarrow x$;
    for $a \in A$ **do**
      $x' \leftarrow \text{apply\_artifact}(x', a, p)$;
    end
    /* Perform one round of gradient descent using Adam */
    $f' \leftarrow \text{train}(f', \eta, x', y)$;
  end
end
return f'

Algorithm 1 shows the full pipeline of our extended version of SpRAy to ImageNet. It takes a subset X_y of the full training data set X that only contains samples of class y, as well as a model f as input. It first computes the attributions for all samples in X_y, e.g. using LRP. Then, it performs the original SpRAy algorithm on it, which generates the spectral embeddings Φ, the eigenvalues Λ, and the clusterings $\mathbb{K}$.

5.2. ImageNet Artifacts across Different DNN-Architectures

In Section 3.2 we describe a series of systematic prediction biases discovered using the SpRAy technique for several affected classes. In all these cases, the downloaded VGG-16 model has overfit on input features which are characteristic for certain object classes in context of the ImageNet dataset. We thus assume that other neural network architectures sharing the same data source for training may also share certain Clever-Hans strategies with the investigated VGG-16 classifier.

Figure S1 exemplarily shows LRP heatmaps computed for the VGG-19 (Simonyan & Zisserman, 2014) and the DenseNet-121 (Huang et al., 2017) model — which have also been downloaded as pre-trained predictors optimized

on the ImageNet data corpus — for samples which exhibit data artifacts as discovered for the VGG-16 model. We notice that both the architecturally very similar VGG-16 and VGG-19 architectures heatmaps are very concentrated on shape features such as edges and color-gradient rich image areas. The heatmaps computed for the DenseNet-121 model on the other hand are much more focused on class- and object-specific textures and colors. For all investigated samples, we however notice that all three models tend to use the same w.r.t. the true class semantically unrelated yet correlated features for prediction. That is, for class "carton", all three models support their predictions with a set of barely visible and centered watermark consisting of chinese characters for prediction, as well as a second orange and small watermark appearing in the bottom right corner of "carton" images with high frequency. Similarly for classes "garbage truck", "jigsaw puzzle" and "stole" shown in Figure S1 all three models support their prediction based on the discovered yellow watermark, the cut-outs of the digitally added puzzle pattern, the rounded image corners and the wooden mannequin head.

Considering the systematicity of use of these data artifacts by all three models, we strongly recommend a thorough categorization of Clever-Hans behavior of machine learning models and their data sources essential components of future dataset creation efforts.

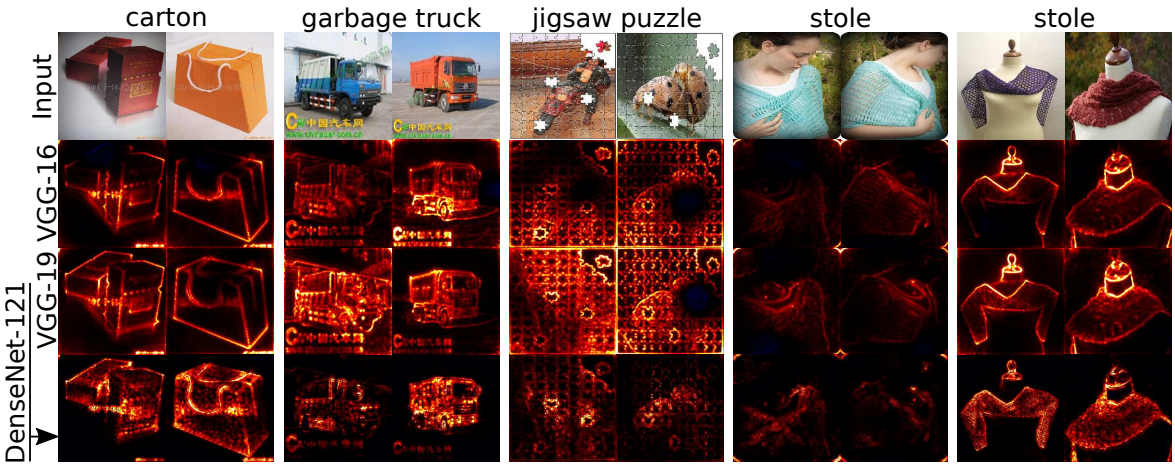

Figure S1. Heatmaps for classes and samples with on ImageNet and VGG-16 discovered data and prediction artifacts for the VGG-19 and DenseNet-121 models. *Top to bottom:* Input samples, heatmaps for VGG-16, VGG-19 and DenseNet-121. *Left to right:* Colums show (in pairs) artifacts for classes "carton", "garbage truck", "jigsaw puzzle", "stole" (rounded corners) and "stole" (prediction supported by mannequin head), which have been discoverd from a VGG-16 classifier, but apply to all three models.

5.3. Additional ClArC Experiments

Figure S2 shows an extreme drop in performance for the baseline model when poisoning all models with the "jigsaw puzzle" artefact. This is to be expected, as this artefact is the only discriminative feature for its class for all samples that show this artefact.

In Figure S3, the experiments of Section 3.3 are repeated with 1 artefact class + 9 randomly chosen other classes. The results are somewhat less expressive than for 20 classes. This may be caused by the fact that with more classes we poison a higher absolute number of samples.

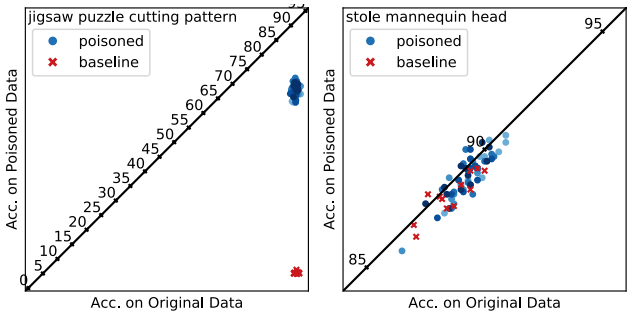

Figure S2. Original (x-axis) vs. poisoned (y-axis) validation set accuracy of baseline model (red) and ClArC'ed models (blue) over poisoning (shade). Additional results for Figure 6.

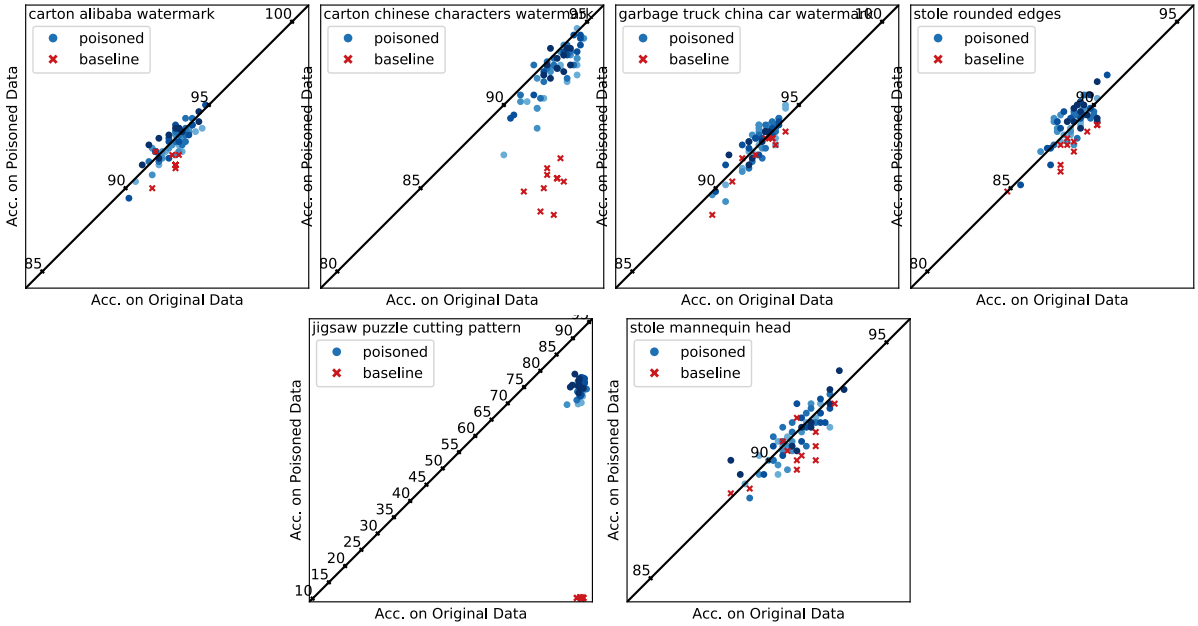

Figure S3. Original (x-axis) vs. poisoned (y-axis) validation set accuracy of baseline model (red) and ClArC'ed models (blue) over poisoning (shade). Additional results for the experiment in Section 3.3 with 10 classes instead of 20 (Figure 6).

5.4. Additional ClArC Training-Attribution Validation

Figure S4 demonstrates a setting highly similar to the one for class "garbage truck" discussed in Section 3.3: The discovered data artifact — here a digitally rounded image corners with white background — exhibits extremely high regional consistency and only covers very limited parts of the image area. Once the isolated corner feature has been added to *all* samples during our experiment, the model quickly has disassociated the artifact from the label "stole". After continued re-trainng, LRP begins to attribute negative relevance to rounded image corners, indicating that the process of un-Hansing went beyond mere forgetting by creating a negative association between corner artifact and class label.

The second data artifact discovered for class "stole" is a frequently shown wooden "mannequin head" co-appearing with the woven stoles themselves. Since here, the expression of the artifact was much more diverse in pose and position and has shown almost no regional consistency, we manually isolated a (very) limited amount of prototypical "mannequin heads" from the data and randomly (within

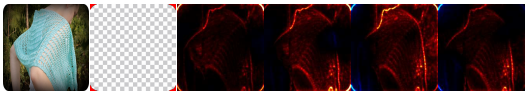

Figure S4. Un-Hans'ing Experiment for class "stole" and the "rounded corners" artifact. *Left to right:* Example input, the artifact (with transparent background, and the white corner pattern here shown in read for visibility reasons), heatmap expressions computed during the un-Hans'ing process.

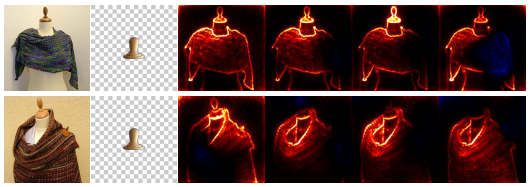

Figure S5. Un-Hans'ing experiment for class "stole" and the "mannequin head" artifact. *Left to right:* Example inputs, the artifact (a manually isolated wooden manneqin head), heatmap expressions computed during the un-Hans'ing process.

reason) added wooden stump as an image element to each sample of each batch during re-training. Figure S5 shows the progression of un-Hansing at hand of two different input sample. While for the sample shown at the top of the figure the model has not disassociated between this particular expression of the "mannequin head" feature (at times, the feature's accumulated positive relevance even increased), the model has ceased to support its prediction for class "stole" with the artifactual feature for the bottom image.

Lastly, we investigate the "digital jigsaw puzzle pattern" artifact discoverd for class "jigsaw puzzle", which appears in multiple variants. Each variant, however, is expressed with almost complete and pixel-identical consistency. We therefore select one variant of the artifact and add it as a mask to *all* training samples of the un-Hansing training subset B extracted from the ImageNet corpus. Here again, we can observe that the model *forgets* the association between this particular pattern and the class label "jigsaw pattern": In Figure S6, positive relevance completely disappears from the digital jigsaw pattern during un-Hansing, such that the feature is not used anymore for predicing "jigsaw". What prevails, however, is a strongly negative relevance map on the fornicating ladybug pair of ladybugs, indicating the model's reasoning that the insects' presence speaks *against* class "jigsaw" (and rather for a competing network output). The effect of forgetting this consistently expressed yet very large artifactual feature has an understandably catastrophic effect to the model's capability to predict the original ImageNet label for affected samples (cf. Table S2).

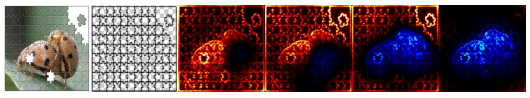

Figure S6. Un-Hans'ing Experiment for class "jigsaw puzzle" and the "digital puzzle pattern" artifact. *Left to right:* Example input, the artifact (semi-automatically isolated jigsaw pattern), heatmap expressions computed during the un-Hans'ing process.

Our experiments show that unlearning patterns from the trained neural networks is possible, but might be non-trivial for more abstract patterns than the our discussed ones, which are mostly spatially fixed in pixel space.

5.5. More Clever-Hans Candidates

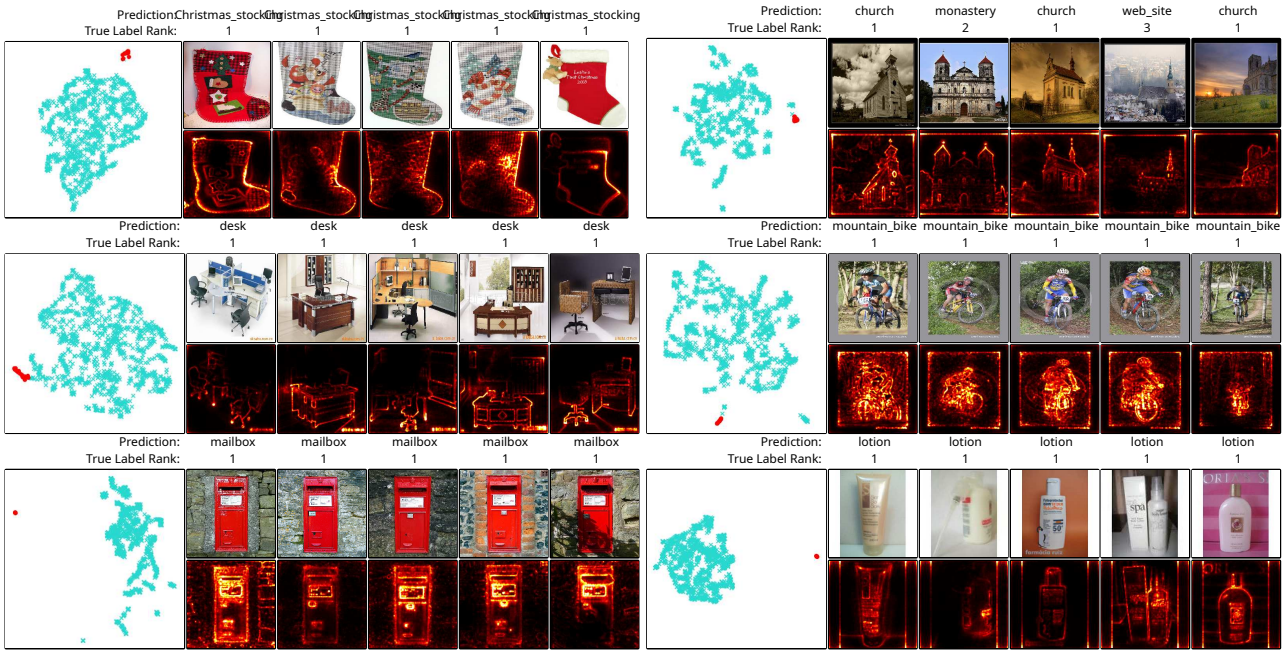

Figure S7. UMAP (left) with samples and heatmaps (right) of significant clusters, highly separated from the rest of the samples. For each class, some images and their respective attributions from the identified cluster are shown. Red dots in the UMAPs identify the clusters the samples to the right have been grouped into. Relevance maps to the right are color coded to identify relevant image regions *supporting* the classifier decision in hot colors (red to yellow) irrelevant regions in black color and relevant regions *contradicting* the final prediction in cold (blue to cyan) hues. The text above the sample images shows the classifier's top-1 predicted class, and the prediction rank of the true label.

5.6. Clever-Hans Candidates with SmoothGrad

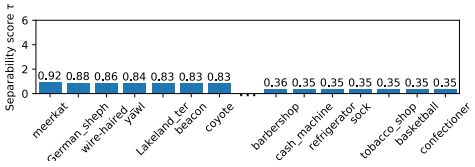

Figure S8. Mean separability score τ of spectral embedding of SmoothGrad attributions based on Fisher Discriminant Analysis. A high τ means there are significantly different decision strategies being used, potentially of Clever-Hans type.

Figure S9 shows the UMAP visualization and some highlighted clusters with SmoothGrad (Smilkov et al., 2017) attribution heatmaps. We sample 50 times with a noise-level (as described in (Smilkov et al., 2017)) of 10%. Overall, we can observe some similar clusters to our LRP experiments (e.g. "laptop", "mailbox"), but finding significant examples with somewhat more of a challenge than with LRP. The FDA ranking, shown in Figure S8 also hints at less easily detectable artifacts with low scores even for the most separable classes.

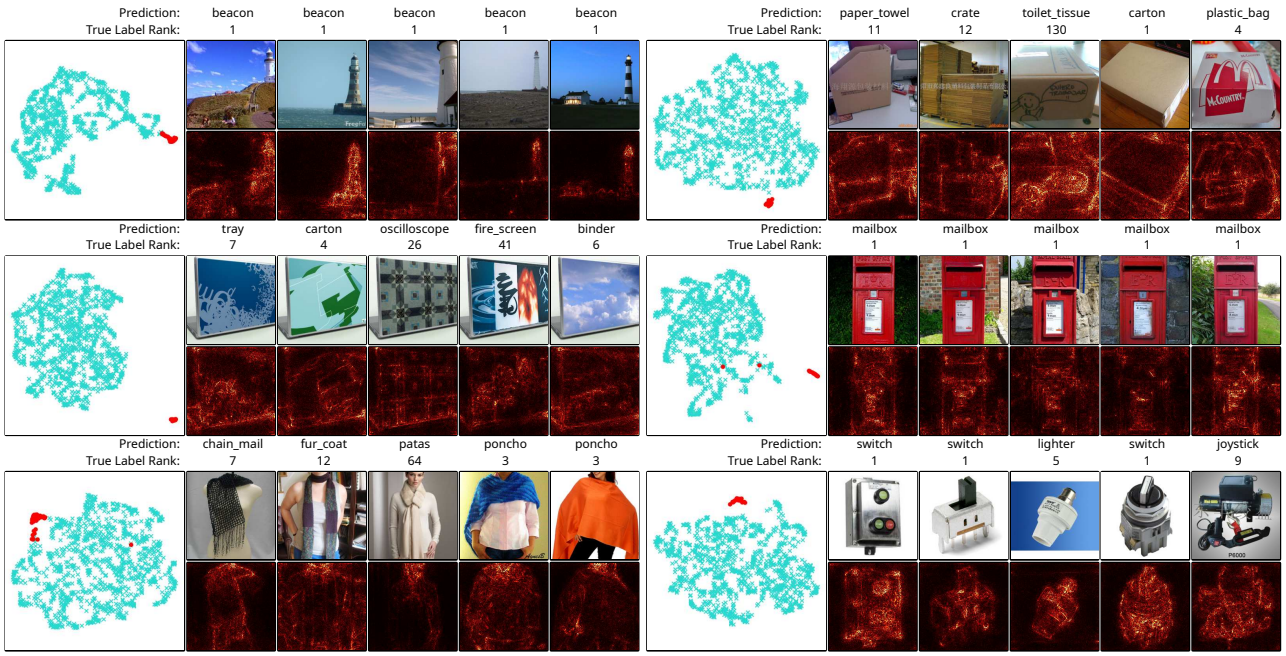

Figure S9. UMAP (left) with samples and heatmaps (right) of significant clusters, highly separated from the rest of the samples. For each class, some images and their respective attributions from the identified cluster are shown. Red dots in the UMAPs identify the clusters the samples to the right have been grouped into. Relevance maps to the right are color coded to identify relevant image regions *supporting* the classifier decision in hot colors (red to yellow). The text above the sample images shows the classifier's top-1 predicted class, and the prediction rank of the true label.

5.7. Model Behavior of Addition and Removal of Artifacts

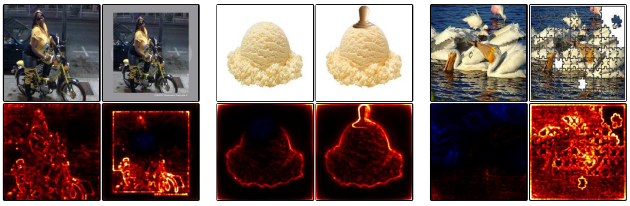

Figure S10. Addition of discovered artifacts to samples of other classes. Relevance maps are computed w.r.t. the class of origin of the artifact. *Left:* Addition of a border which *transforms* a "moped" into "mountain bike". *Mid:* The addition of the "mannequin head" increases the classifier output for class "stole". Note how the model interprets the lack of a "mannequin head" on top of the ball of ice cream in the left heatmap as contradictory feature. *Right:* Further note how the model considers the white color in the image corners as features for "stole". Adding a digital puzzle pattern pattern to any image forces a high probability "jigsaw puzzle" prediction.

We summarized the results for a quantitative verification of selected hypotheses in Table S1, with mean prediction rank difference $\mu(\Delta(\text{rk}))$ and mean prediction difference $\mu(\Delta f(x))$.

By *removing* an artifact, we can estimate to what degree a model has learned to (solely) base its decision on the artifactual feature. If the model reacts strongly to the removal of the artifactual image feature, it has (with high probability) resorted to the artifact as a main source of information for the respective target class. If the model does only show a weak reaction or none at all, it may have learned (several) backup strategies for detecting the concept of the target label.

We measure the model's sensitivity to the artifacts discussed in this section, by using digital inpainting techniques on the affected samples in the validation set. Table S2 compiles measurements $\mu(\Delta(\text{rk}))$ and $\mu(\Delta f(x))$ for artifact removals on classes "stole", "jigsaw puzzle" and "mountain bike". While the prediction for class "mountain bike" is almost completely unaffected again, and the classifier seems to have developed backup plans for predicting class "stole" in the absence of rounded image corners and wooden mannequin heads, the "jigsaw puzzle" classifier catastrophically fails in two out of three cases when a discovered digitally pasted jigsaw puzzle pattern is removed from the affected samples. The model has thus, for the class "jigsaw puzzle", strongly overfitted to the discovered dataset bias.

Table S1. The effect of *adding* a class-related artifact to samples of other classes, towards the prediction of the artifact's class of origin. For all artifacts except the freely placable "mannequin head", we randomly selected 2000 samples from other classes and measured the effect of the artifact addition. The $\mu(\Delta(\text{rk}))$ measures the mean change in prediction *ranking* due to the artifact addition and $\mu(\Delta f(x))$ measure the mean change in the artifact's *class probability*. High(er) values mean that the model is (strongly) affected by the artifact in its decision for the artifact's class of origin

class	bias	samples	$\mu(\Delta(\text{rk}))$	$\mu(\Delta f(x))$
stole	rounded corners	2000	58.14	0.0004
stole	mannequin "head"	10	106.10	0.0081
jigsaw puzzle	jigsaw pattern 1	2000	220.98	0.0160
jigsaw puzzle	jigsaw pattern 2	2000	355.60	0.8415
jigsaw puzzle	jigsaw pattern 3	2000	356.42	0.9540
mountain bike	watermark	2000	-101.02	0.0001

Table S2. The effect of *removing* a class-related artifact from image samples, towards the prediction of the artifact's class of origin. The $\mu(\Delta(\text{rk}))$ measures the mean change in prediction *ranking* due to the artifact addition and $\mu(\Delta f(x))$ measure the mean change in the artifact's *class probability*. Low(er) values mean that the model is (strongly) affected by the artifact removal in its decision for the artifact's class of origin

class	bias	samples	$\mu(\Delta(\text{rk}))$	$\mu(\Delta f(x))$
stole	rounded corners	10	-0.70	-0.1756
stole	mannequin "head"	13	-0.62	-0.3713
jigsaw puzzle	jigsaw pattern 1	44	-0.11	-0.0146
jigsaw puzzle	jigsaw pattern 2	44	-112.52	-0.9160
jigsaw puzzle	jigsaw pattern 3	44	-208.41	-0.9305
mountain bike	watermark	17	0.00	0.0206

5.8. Validation Set Artifacts

As an additional interesting observation, we have also found classes with examples in the validation set of ImageNet that show the same type of artifacts as used in some of the discovered Clever-Hans prediction strategies (e.g. see Figure S11), putting the model's performance on the validation set for any of the affected classes in question.

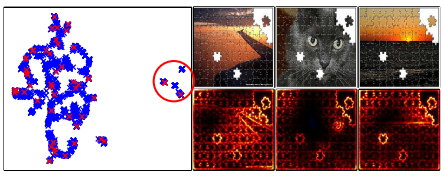

Figure S11. Left: UMAP of Spectral Embedding on union of training *(red)* and validation set *(blue)* for class "jigsaw puzzle". *Right:* Images of the validation set in the previously identified "jigsaw puzzle" bias (top) with attributions (bottom).

5.9. Alternative distance measures

SpRAy has originally only been shown with euclidean distance to compute the neighborhood graph (Lapuschkin et al., 2019). In the application of images, this means that image similarity is identified by spatial properties, i.e. having the same attribution intensities at the same pixel renders high similarity. This is a reasonable approach, especially if one would like to focus on spatial properties such as watermarks or padding. However, when the domain of interest are spatially unrelated shapes or color distributions, other measures of similarity may be needed. A recently very popular distance metric is the Optimal Transport, or *Wasserstein-Distance*. In the context of computer vision, it is also known as the *Earth-Mover's Distance* (Rubner et al., 1998). Its benefit is that it "feels" like a very natural distance metric (Solomon et al., 2015).

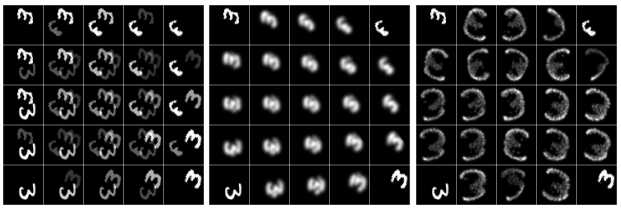

Figure S12. Barycenters of four rotated and translated MNIST digits. The original images are in the four corners. Metrics are euclidean (*left*), Wasserstein distance (*middle*) and Gromov-Wasserstein distance (*right*).

Wasserstein distances use distances between spatially fixed points over the same identical image grid. The Gromov-Wasserstein (Peyré et al., 2016) distance matches points

Figure S13. Top: Significant Gromov-Wasserstein distance based SpRAy cluster of class "great grey owl" with the corresponding attribution maps below the samples. *Bottom:* Same for class "ring-neck snake".

by their pairwise distances, instead of using a fixed image grid with a fixed amount of points. This means that however points are spatially distributed, if in both sets there are points whose pairwise relations are similar, then their Gromov-Wasserstein distance will be small. A somewhat intuitive visualization of euclidean distance, Wasserstein distance, and Gromov-Wasserstein distance is shown in Figure S12. We show 4 samples of hand-written digits (LeCun, 1998) in 4 corners, translated and rotated. All images that lie on the line between the corners are barycenters (Cuturi & Doucet, 2014) of the corner images, weighted by the Chebyshev distance to all samples. The metrics used to compute the barycenters are the 3 previously mentioned metrics. Wasserstein barycenters are computed as in (Solomon et al., 2015). For the Gromov-Wasserstein distance, we need to compute pairwise distances between points in the image. Points are extracted from the images by choosing each pixel one after another, starting with the largest pixel value, until 99 percent of the total sum of all pixel values is reached. We can nicely see that the Wasserstein distance seems translation invariant, but fails with different rotations. Gromov-Wasserstein distance shows to be invariant to rotation, translation, and mirroring, since all the information is contained in only the pairwise relations.

We can recognize groupings of complicated shapes, invariant of scale, location or translation on clusters found with Gromov-Wasserstein distance at the base of SpRAy. Examples for two distilled clusters from classes "ring-neck snake" — where the snake's head and its brightly colored neck appear to be the relevant features — and "great grey owl" — where the patterns highlighting the face (eyes and beak) and shape of the head seem to be the common denominator — can be seen in Figure S13. However, despite the favorable invariance properties, deducting distinct (and automatedly testable) hypotheses for these strategies turns out to be a non-trivial task, since clusters are semantically much harder to interpret compared to groupings found with a euclidean distance at the root.