Localized Shortcut Removal

Nicolas M. Müller *
Fraunhofer AISEC, Germany
nicolas.mueller@aisec.fraunhofer.de

Jochen Jacobs *
TU Munich, Germany
jochen.jacobs@tum.de

Jennifer Williams
University of Southampton, UK
J.Williams@soton.ac.uk

Konstantin Böttinger
Fraunhofer AISEC, Germany
konstantin.boettinger@aisec.fraunhofer.de

Abstract

Machine learning is a data-driven field, and the quality of the underlying datasets plays a crucial role in learning success. However, high performance on held-out test data does not necessarily indicate that a model generalizes or learns anything meaningful. This is often due to the existence of machine learning shortcuts - features in the data that are predictive but unrelated to the problem at hand. To address this issue for datasets where the shortcuts are smaller and more localized than true features, we propose a novel approach to detect and remove them. We use an adversarially trained lens to detect and eliminate highly predictive but semantically unconnected clues in images. In our experiments on both synthetic and real-world data, we show that our proposed approach reliably identifies and neutralizes such shortcuts without causing degradation of model performance on clean data. We believe that our approach can lead to more meaningful and generalizable machine learning models, especially in scenarios where the quality of the underlying datasets is crucial.

1. Introduction

Shortcuts in machine learning data refer to false features that are strongly correlated with the target class but are not expected to be present in real-world applications. These features are easy for neural networks to learn, but they may not generalize beyond the training data. Shortcuts can arise from various factors, such as the data collection process, data collection techniques, or the type of data being collected. Often, these shortcuts are highly localized and spatially much smaller than true features [7, 13, 26]. For instance, a neural network trained on an image dataset where all images of class k exclusively contain watermarks has been shown to rely solely on the presence of the watermark to predict the class [1, 13]. Indeed, identifying shortcuts during data collection or preprocessing can be a challenging task. This is evidenced by the fact that there are many datasets released to the public that contain shortcuts [1, 7, 9, 15].

Training a model on data with shortcuts can lead to an over-reliance on irrelevant features. This results in seemingly high performance on held-out data if the shortcut is present, which may be the case if the test data is sampled via the same process as the training data. Such models may not generalize well to out-of-distribution (OOD) data, which is a common issue in machine learning known as domain generalization [27].

In this paper, we introduce a supervised neural network that can learn the essential features of a dataset, even if there are localized shortcuts present (known or unknown). To accomplish this, we use an adversarially trained "neural lens" that can remove shortcut features and provide a visual representation of the avoided shortcuts. Our model is successful in identifying and in-painting shortcuts in various datasets, such as chest x-rays from the COVID QU-Ex dataset [9]. Importantly, this process doesn't harm the model's performance when no shortcuts are present.

2. Related Work

In machine learning, shortcuts come in varying degrees of spatiality, ranging from small and localized to global. Local examples include logos and watermarks in image datasets, such as the Pascal VOC 2007 dataset's watermark on horse photos [1,13], or hospital- or device-specific marks in chest x-ray images [7, 26]. Meanwhile, global shortcuts include the presence of pastures as an easy indicator for the class "Cow" [4], or artefacts in pooled medical databases, where patient positioning, imaging device type, and image size are utilized by the model to infer the target class [18]. These shortcuts are problematic not only in supervised com-

*equal contribution

puter vision but also in self-supervised learning [10] and when using pretext tasks to design feature extractors [8, 14]. Additionally, shortcuts are not limited to image datasets; they can also be observed in audio datasets. For instance, the amount of leading silence in the ASVspoof Challenge Dataset on audio deepfake detection can be utilized to predict the target class [15, 24].

2.1. Automatic shortcut removal

One potential solution to address the presence of shortcuts in a dataset is to remove them. For instance, in the context of self-supervised representation learning, Minderer et al. [14] suggest incorporating a U-Net [19], referred to as a "lens," in front of the classification network. The lens is trained adversarially and enables the elimination of local shortcuts, such as logos, through in-painting. However, this approach is restricted to self-supervised learning. In the supervised domain, adversarial autoencoders have been proposed by Baluja et al. [3] and Poursaeed et al. [16]. In this approach, an autoencoder is added at the beginning of a classification network and trained adversarially to generate images that appear similar to the original input, but can mislead the classifier into producing incorrect output. Similarly, Xiao et al. [25] introduce AdvGAN, which incorporates a GAN-discriminator as an additional loss for the autoencoder, leading to less noticeable perturbations. While these methods share similarities with the architecture proposed in this work, none utilize the generated adversarial images to robustly train the classifier.

2.2. Improving model robustness

An alternative approach for addressing shortcuts is to enhance the robustness of models against them. Wang et al. [23] propose the use of gradient-reversal to deceive helper networks that consider only small local patches, while the global network is encouraged to classify the overall input correctly. A similar idea is explored in [6]. To prevent a network from focusing excessively on shortcuts that exist only in a subset of the dataset, Dagaev et al. [5] suggest weighted training, which involves assigning lower weights to images that can be accurately classified by a low-capacity network, assuming that those contain shortcuts. However, this approach may not be effective when a significant number of images in the dataset contain shortcuts, unlike our proposed method, c.f. Sec. 5.1. Lastly, for known shortcuts, one can artificially introduce them into the dataset and encourage the model to disregard them [2]. The drawback of this approach is that the shortcuts must be identified beforehand.

3. Architecture

To remove shortcuts in supervised problems, we adopt an unsupervised learning architecture [14]. A low-capacity Image-to-Image network (called "Lens Network") is placed in front of the classification network. This lens is then trained jointly, but adversarially, with the classifier to decrease its performance. The idea is that the lens is trained to isolate features of the image that the classification network is paying attention to. Since the capacity of the lens is limited, only simple features (i.e. shortcuts) can be removed by the lens. To further enforce this, we extend the training loss with an additional reproduction loss L_{repr}. This ensures that the lens modifies the original image only slightly.

Inspired by [17], we propose using two U-Net-based networks, an attention network A and a replacement network R, as shown in Fig. 1. Network A determines the location of the shortcut in the original image, while network R computes a suitable replacement for the shortcut.

Given an input image I, we obtain a shortcut-removed image I' as follows:

$$I' = A \cdot R + (1 - A) \cdot I. \tag{1}$$

The capacity of the attention network corresponds to the complexity of the shortcuts identified and should be chosen accordingly. Since the task of the replacement network is more complex than that of the attention network, therefore we accord a larger model capacity (i.e. more up- and downsampling steps) to R than to A. Lens and classification model are trained jointly via $L = \lambda L_{repr} + L_{CE}$ where L_{CE} is the cross entropy loss of the classification network C and λ is a hyperparameter controlling how much the lens is allowed to modify the input image:

$$L_{repr} = \max\left(\rho, \frac{1}{wh}\sum_{ij} A_{ij}\right) - \rho. \tag{2}$$

$\rho \in [0, 100\%]$ is a hinge hyperparameter that controls the percentage of the image that can be modified without penalty. Note that while gradients from the cross-entropy loss flow into both the lens and classifier, the reproduction loss only affects the lens. In our experiments, we use the ResNet18 [11] architecture as classifier C.

We have noticed oscillations during training, where the classifier stops paying attention to the shortcuts once they are removed, leading the lens to stop removing them. To counteract this, we pass a copy of the image directly to the classifier. This ensures consistent focus on the shortcuts and attenuates oscillations during training.

4. Data and synthetic shortcuts

We assess the performance of our proposed architecture on both synthetic and real-world datasets. Initially, we examine our model's efficacy by introducing artificial shortcuts on CIFAR10 [12] and ImageNet [20]. Specifically, for CIFAR10, we create "Color Dot" and "Location Dot" shortcuts by in-painting a circle in which the color or location

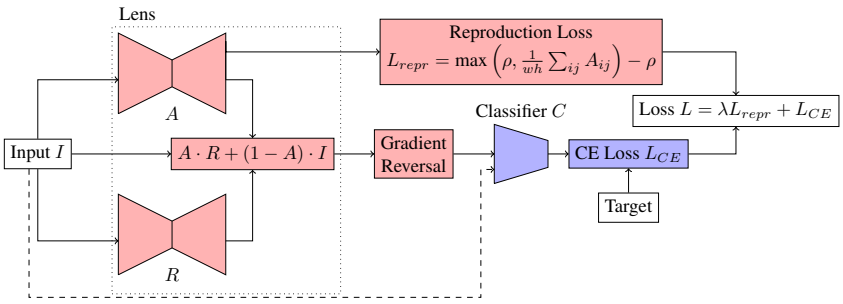


Figure 1. Architecture of our proposed *attention lens* model. The lens (red) consists of an attention module A and a reconstruction module R, both of which are U-Nets. Its output is passed to the original classifier C (blue), trained via cross-entropy loss. Optionally, input images are also passed to the original classifier. The lens is trained via the classifier's inverted gradients and a reproduction penalty loss L_{repr}.

corresponds to the target class, as shown in Fig. 2. In addition, we use a subset of the visually similar target classes, "goose" and "pelican," from the ImageNet dataset [20] to simulate real-world scenarios where classes have overlapping visual features, such as in medical image analysis. To enhance visual similarity, we convert these images to grayscale and introduce shortcuts by overlaying a single logo or a textual watermark across the entire image.

Furthermore, we conduct an evaluation on real-world data, specifically, the covid-qu-Ex dataset [9], which comprises x-ray images of the human chest labelled as either "healthy", "COVID-19", or "pneumonia". Chest X-ray images have been previously found to contain shortcuts [26], especially when obtained from multiple sources, such as different hospitals. Upon visually examining the dataset, we observe a significant amount of text, markers, and medical equipment in the corners of the images that may serve as shortcuts, provided they are correlated to the target class. Such shortcuts can severely impede the practical applicability of machine learning models in real-world scenarios [7].

5. Experiments and Results

5.1. Synthetic Data

This section presents the results of our proposed model when training on shortcut-perturbed data, and evaluating on clean test data (CIFAR and ImageNet).

Experimental Setup. For the attention network A, we chose 3 downsampling steps and 5 downsampling steps for the replacement network R. For the CIFAR-based experiments, we use $\rho = 2.5\%$, while for the ImageNet experiments, we use $\rho = 5.0\%$ (logo shortcut) or $\rho = 10.0\%$ (watermark shortcut). Classifier and Lense have different learning rate ($1.5 \cdot 10^{-6}$ and $1 \cdot 10^{-4}$, respectively). We use $\lambda = 15$ and train the model for 30 epochs on CIFAR10, and 50 epochs on ImageNet.

Results. Based on the results presented in Table 1, we make the following observations. Firstly, the absence of shortcuts does not impair the test accuracy, indicating that our proposed solution is effective without any drawbacks. Secondly, our proposed shortcuts prove to be highly effective, leading to a substantial decrease in test performance (first row). For instance, the Color Dot shortcut lowers the accuracy from 75% to 28.5%, reflecting the model's over-reliance on the simplistic shortcut features. However, with the lens activated, the adverse impact of the shortcuts is almost entirely mitigated. The performance of the "Color Dot" shortcut on CIFAR10 is restored from 28.5% to 70.5% of the original 75%, for example.

	Shortcut	W/o Lens	With Lens
CIFAR10	None	$\mathbf{75.1 \pm 2.4}$	$\mathbf{76.7 \pm 2.3}$
	Color Dot	28.5 ± 0.9	$\mathbf{70.5 \pm 2.1}$
	Location Dot	41.9 ± 7.0	$\mathbf{69.0 \pm 3.2}$
ImageNet	None	$\mathbf{78.9 \pm 1.1}$	$\mathbf{76.1 \pm 2.8}$
	Logo	51.9 ± 2.0	$\mathbf{74.1 \pm 9.0}$
	Watermark	52.4 ± 1.4	$\mathbf{61.0 \pm 5.2}$

Table 1. The effect of the lens network, measured in test accuracy. We train a ResNet18 architecture on datasets with and without shortcuts and subsequently assess the model's performance on clean validation data. The experiment is repeated three times, and the mean test accuracy and a 95% confidence interval are reported.

Visualization: Figure 2 presents example outputs of the attention lens when training on the CIFAR10 Color Dot shortcut. We make the following observations based on the visualization: Firstly, the attention lens successfully removes the shortcuts from the image. Secondly, for the Color Dot shortcut, recoloring the dots is sufficient to eliminate the shortcut as only the color of the dot is deterministic of the class. Additionally, we perform similar experiments for the Location Dot shortcut. The model correctly learns that

Figure 2. Examples of shortcuts and lens output on CIFAR10 training data. **Row 1** shows the input image with the color dot shortcut added: for example, all cars have a blue dot. **Row 2** shows the output of the lens, where the shortcut is mitigated by recoloring.

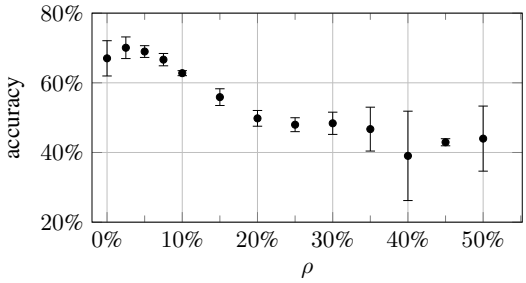


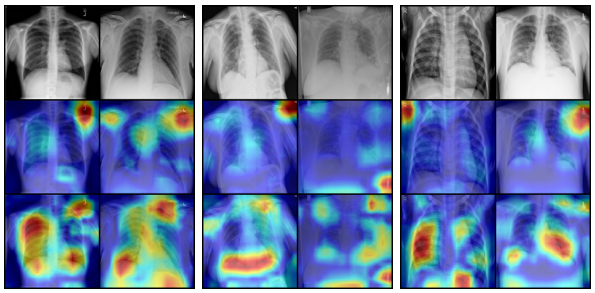
Figure 3. Accuracy on the *clean* validation set when training our model on the CIFAR10 dataset, with varying degrees of ρ (Location Dot shortcut).

the Location Dot shortcut cannot be removed by recoloring the dots. Instead, the lens fills the dots by in-painting a best-effort background.

In order to determine the optimal value of ρ, we conduct a CIFAR10 Location Dot experiment with varying values of ρ. Specifically, we evaluate each of the candidate values for ρ over three independent runs, and reported the mean accuracy and 95% confidence interval in Fig. 3. Our findings suggest that the optimal value of ρ for this particular shortcut is around $\rho = 2.5\%$, which approximately corresponds to the percentage of the image occupied by the shortcut. A significantly higher value of ρ leads to the lens over-manipulating the image, resulting in a poor classifier performance on the original images.

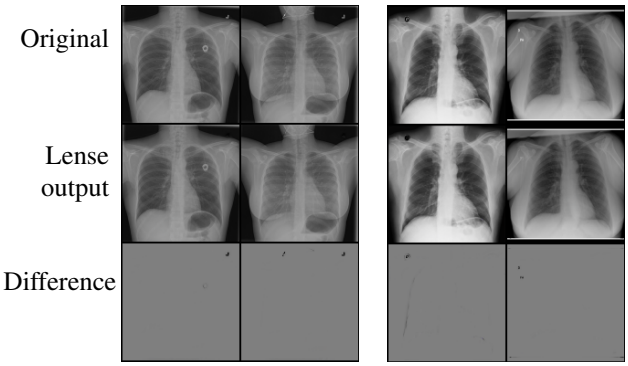
5.2. Real-World Data

For the covid-qu-Ex dataset, we trained the network with hyperparameters $\lambda = 5$, $\rho = 0.25\%$, 2 downsampling steps in the attention network, and 5 downsampling steps in the replacement network. We used a learning rate of $2 \cdot 10^{-4}$ for both the lens and classifier. As there is no validation set without shortcuts for covid-qu-Ex, we evaluated the effectiveness of the lens in identifying shortcuts using GradCAM [22]. Figure 4 shows the GradCAM images for all three classes and both trained networks. From these experiments, we made several observations. First, without the lens, the network predominantly focused on areas in the corners of the images, mostly in areas with text. Second,



(a) Normal (b) COVID (c) Pneumonia

Figure 4. GradCAM images showing network attention when training on the covid-qu-Ex dataset. **Row 1** is the input image from the validation set. **Row 2** is the classifier attention of a network trained without, and **Row 3** with our proposed model.



(a) Normal (b) COVID

Figure 5. Lens output and attention on x-ray images from the covid-qu-Ex dataset for the classes classes COVID and Normal. **Row 1** shows original images. **Row 2** shows the output of the lens. **Row 3** shows the difference between rows 1 and 2.

with the attention lens, the network focused on more relevant sections of the image, including the lungs. Our proposed approach not only explains shortcuts but also corrects them, as shown in Fig. 5, where highly localized shortcuts such as markers and text are removed.

6. Conclusion

In this paper, we propose a method for detecting and eliminating small but highly influential shortcuts in machine learning datasets. Our approach is built upon the hypothesis that genuine features are typically more global in nature, whereas shortcuts are localized but highly predictive. However, we acknowledge that there may be datasets containing global shortcuts such as image background [21] or ambient lighting, but leave this for future work. To validate our proposed approach for localized shortcut detection, we conduct experiments on both synthetic and real-world datasets and demonstrate our model's effectiveness.

References

[1] The pascal visual object classes challenge 2007 (voc2007). http://host.robots.ox.ac.uk/pascal/VOC/voc2007/index.html. (Accessed on 08/03/2022). 1

[2] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 2

[3] Shumeet Baluja and Ian Fischer. Adversarial Transformation Networks: Learning to Generate Adversarial Examples. *arXiv:1703.09387 [cs]*, Mar. 2017. 2

[4] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in terra incognita. In *European Conference on Computer Vision (ECCV)*, 9 2018. 1

[5] Nikolay Dagaev, Brett D. Roads, Xiaoliang Luo, Daniel N. Barry, Kaustubh R. Patil, and Bradley C. Love. A Too-Good-to-be-True Prior to Reduce Shortcut Reliance. *arXiv:2102.06406 [cs]*, Oct. 2021. arXiv: 2102.06406. 2

[6] Nikolay Dagaev, Brett D Roads, Xiaoliang Luo, Daniel N Barry, Kaustubh R Patil, and Bradley C Love. A too-good-to-be-true prior to reduce shortcut reliance. *Pattern Recognition Letters*, 166:164–171, 2023. 2

[7] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. 1, 3

[8] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. In *IEEE International Conference on Computer Vision (ICCV)*, 12 2015. 2

[9] Tahir et Al. COVID-19 infection localization and severity grading from chest X-ray images. *Computers in Biology and Medicine*, 139:105002, Dec. 2021. 1, 3

[10] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised Representation Learning by Predicting Image Rotations. *arXiv:1803.07728 [cs]*, Mar. 2018. arXiv: 1803.07728. 2

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[12] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[13] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019. 1

[14] Matthias Minderer, Olivier Bachem, Neil Houlsby, and Michael Tschannen. Automatic Shortcut Removal for Self-Supervised Representation Learning. pages 6927–6937. PMLR, Nov. 2020. 2

[15] Nicolas M. Müller, Franziska Dieckmann, Pavel Czempin, Roman Canals, Konstantin Böttinger, and Jennifer Williams. Speech is Silver, Silence is Golden: What do ASVspoof-trained Models Really Learn? *ASVspoof 2021*, Sept. 2021. 1, 2

[16] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[17] Albert Pumarola, Antonio Agudo, Aleix M. Martinez, Alberto Sanfeliu, and Francesc Moreno-Noguer. GANimation: Anatomically-aware Facial Animation from a Single Image. 2

[18] Caleb Robinson, Anusua Trivedi, Marian Blazes, Anthony Ortiz, Jocelyn Desbiens, Sunil Gupta, Rahul Dodhia, Pavan K Bhatraju, W Conrad Liles, Aaron Lee, et al. Deep learning models for covid-19 chest x-ray classification: Preventing shortcut learning using feature disentanglement. *medRxiv*, 2021. 1

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 2, 3

[21] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv:1911.08731*, 2019. 4

[22] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, Oct. 2017. 4

[23] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[24] Xin Wang, Junichi Yamagishi, Massimiliano Todisco, Héctor Delgado, Andreas Nautsch, Nicholas Evans, Md Sahidullah, Ville Vestman, Tomi Kinnunen, Kong Aik Lee, et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020. 2

[25] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating Adversarial Examples with Adversarial Networks. *arXiv:1801.02610 [cs, stat]*, Feb. 2019. arXiv: 1801.02610. 2

[26] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, Nov. 2018. 1, 3

[27] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2022. 1