Shortcut Detection with Variational Autoencoders

Nicolas M. Müller [*1] **Simon Roschmann** [*1 2] **Shahbaz Khan** [*1 2] **Philip Sperl** [1] **Konstantin Böttinger** [1]

Abstract

For real-world applications of machine learning (ML), it is essential that models make predictions based on well-generalizing features rather than spurious correlations in the data. The identification of such spurious correlations, also known as shortcuts, is a challenging problem and has so far been scarcely addressed. In this work, we present a novel approach to detect shortcuts in image and audio datasets by leveraging variational autoencoders (VAEs). The disentanglement of features in the latent space of VAEs allows us to discover feature-target correlations in datasets and semi-automatically evaluate them for ML shortcuts. We demonstrate the applicability of our method on several real-world datasets and identify shortcuts that have not been discovered before. The code is available at github.com/Fraunhofer-AISEC/shortcut-detection-vae.

1. Introduction

Machine learning (ML) addresses a wide range of real-world problems such as quality control (Yang et al., 2020), medical diagnosis (Rajpurkar et al., 2022) and facial recognition (Adjabi et al., 2020). However, transferring new ML technology from the lab to the real world is often difficult due to the limited capacity of ML models to generalize. Among the reasons for this limitation are shortcuts: features in data X that correlate only statistically with the target Y, but are inconsequential for the specific ML task. Geirhos et al. (2020) define shortcuts as a certain group of decision rules learned by neural networks. Shortcuts perform well on training data and on independent and identically distributed (i.i.d.) test data but fail on out-of-distribution (o.o.d.) data.

Shortcut learning has been particularly evident in the medical field. A recent MIT technology report (Heaven, 2021) reveals that hundreds of tools were developed during the COVID-19 pandemic to diagnose the disease from chest X-rays, but none of them was found reliable enough for clinical use. The predictions of these models were often not based on the appearance of the lungs in the X-ray images. As the datasets were acquired from different sources for positive and negative cases, most of the models ended up learning the systematic differences in the data, e.g. the pose of the patient being scanned. Shortcut learning can also be a cause of ethical concerns towards ML. For example, due to the existence of spurious correlations in the training data, ML models have been found to reinforce gender stereotypes (Bolukbasi et al., 2016; Dastin, 2018).

Detecting shortcuts is a major challenge to ensure the reliability and fairness of artificial intelligence (AI). Existing approaches (Zech et al., 2018; Singla & Feizi, 2021) often rely on heatmaps and are thus limited to the identification of spatial shortcuts. We propose a novel method that is capable of identifying a variety of spurious features in images including background, color, object zoom level, and human facial characteristics. To learn representations robust to spurious correlations, Zhang et al. (2022) propose a contrastive approach which ensures that the hidden representations for samples of the same class are close to each other. Kirichenko et al. (2022) suggest to retrain the last layer of a model on a small dataset without spurious correlations. Our approach can serve as a preliminary step by identifying the shortcuts to be addressed.

Concurrent to our work, Yang et al. (2022) introduced Chroma-VAE. The authors partition the latent space of a VAE into two subspaces where one subspace is initially trained with an appended classifier to isolate shortcuts. Subsequently, the final classifier is trained on the second, shortcut-free subspace. While this method is appealing, we believe that a human-in-the-loop approach is required to ultimately distinguish between shortcuts and valid features. In contrast to Yang et al. (2022) we leverage latent space traversal in VAEs and focus on the identification of shortcuts in datasets rather than the creation of a robust classifier.

We utilize a VAE to identify correlations between features of input X and target Y in a dataset $D = (X, Y)$. We provide tools for visualization and statistical analysis on the latent space of the VAE which allow a human judge

[*]Equal contribution [1]Fraunhofer AISEC, Germany [2]Technical University of Munich, Germany. Correspondence to: Nicolas M. Müller <nicolas.mueller@aisec.fraunhofer.de>, Simon Roschmann <simon.roschmann@tum.de>, Shahbaz Khan <shahbaz.khan@tum.de>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

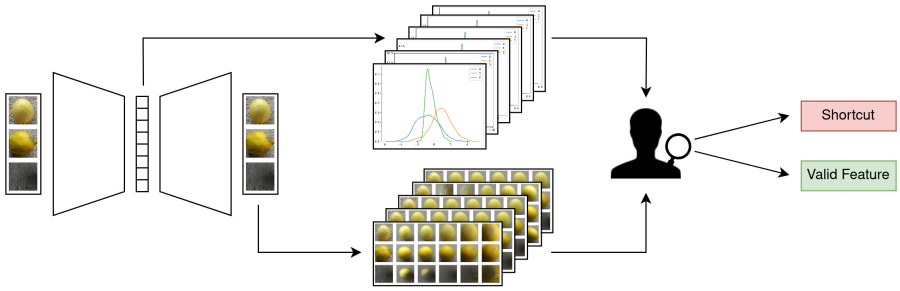


Figure 1. Shortcut detection with a VAE. We first train a Beta-VAE on a dataset containing potential shortcuts. The model discovers independent latent factors in the data and encodes them in its latent dimensions. Evaluating the distribution in the latent space (top) and the weights of a VAE classifier yields a set of candidate dimensions with high predictiveness. The visualization of corresponding latent space traversals (bottom) enables a human judge to identify the meaning of these candidate dimensions and evaluate them for shortcuts/valid features.

to reliably detect ML shortcuts as those correlations that are not meaningful to the task at hand. We demonstrate the applicability of our approach by finding real-world shortcuts in publicly available image and audio datasets.

2. Methodology

Let $D = (X, Y)$ be a dataset, where $X = \{x^{(i)} | x^{(i)} \in \mathbb{R}^{h \times w \times 3}\}_{i=1}^{N}$ are the images, $Y = \{y^{(i)} | y^{(i)} \in \{1, ..., C\}\}_{i=1}^{N}$ are the corresponding targets and C is the number of classes.

We train a Beta-VAE (Higgins et al., 2016) on such a dataset with potential shortcuts. The model discovers independent latent factors in a dataset and encodes them in its latent dimensions. Hence, each latent variable z_j of a trained Beta-VAE is likely to represent a descriptive property of the data, e.g. brightness, color, orientation, or shape (Higgins et al., 2016). To establish feature-target correlations in the data, we measure how predictive each latent variable z_j is for each value of the target variable y (see Section 2.1). We perform a statistical analysis on the latent space of a VAE and assess the utility of its latent dimensions for linear classification. For latent variables z_j that show a strong correlation with the target variable y, we create visualizations that allow a human judge to easily decide whether the property encoded by z_j is a valid feature or a shortcut (see Section 2.2).

2.1. Identification of Feature-Target Correlations

Forwarding an image $x^{(i)}$ through the encoder of a trained Beta-VAE yields the parameters $\mu^{(i)}$ and $\sigma^{(i)}$ of the posterior $q_\phi(z|x^{(i)}) = \mathcal{N}(z; \mu^{(i)}, (\sigma^{(i)})^2 I)$. The mean $\mu^{(i)}$ can be considered as the latent representation for $x^{(i)}$. We forward the entire dataset through the trained VAE, obtaining $\mu^{(i)}$ for all $i \in \{1, \ldots, N\}$. We utilize these representations to determine the feature-target correlations in the data. We propose two different methods for this analysis.

Statistical Analysis on the Latent Space. For each latent dimension j and all target classes $c \in \{1, ..., C\}$, we analyze the distributions $p(z_j | y = c)$. We are interested in finding those dimensions j where two different classes result in two highly disparate distribution estimates. This indicates that feature z_j is highly correlated with the target classes (a necessary, but not sufficient requirement for a shortcut). The separation between any two distributions is quantified in terms of the Wasserstein distance (Vaserstein, 1969) between them. We hypothesize that the maximum pairwise Wasserstein distance (MPWD) for a particular dimension represents its capability to separate classes.

Classification with VAE Encoder. Another approach to understanding the correlation between z_j and y is to employ a classifier to predict y given z. We pick the encoder of our trained Beta-VAE and append a fully connected layer. After freezing the encoder backbone, we optimize its dense classification head. The weights in the last layer of this model denote the correlation between the latent dimensions and the target classes. For a dense classification head $h(z) \rightarrow y$, we define the predictiveness of a feature z_j as

$$\text{pred}(z_j) = \sum_c |\theta_{jc}| \tag{1}$$

where θ_{jc} is the weight of the neuron in h which maps input z_j to output $h(z_j)_c$.

2.2. Evaluation of Feature-Target Correlations for Shortcuts

After identifying the most predictive features in the latent space, we can now evaluate them for shortcuts. To facilitate the distinction between valid and spurious correlations, we provide a human judge with visualizations that convey the meaning of the features.

Visualizing z_j via Latent Space Traversal. The first approach to understanding the high-level features encoded in

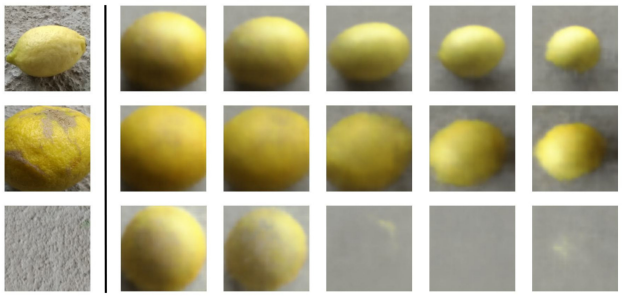


Figure 2. Latent space traversal. The first column shows examples for the three classes ('good quality', 'bad quality', 'empty background') from the *Lemon Quality* dataset. We vary the value of dimension 2 in the latent space from $z_2^{min} = -3.75$ (second column) to $z_2^{max} = 3.64$ (rightmost column) while keeping the values of the other dimensions fixed. Decoding the new latent representation reveals that dimension 2 encodes the zoom level of the images. While close-up shots correlate with class 'bad quality', distant shots correlate with class 'good quality'.

a latent dimension z_j is to visualize the effect of traversal in that dimension on the decoder output using the trained Beta-VAE. Given an instance $x^{(i)}$, we obtain $z^{(i)} = \mu^{(i)}$ from the encoder and then visualize the decoder output $Dec(z^{(i)} + \lambda)$. The linear interpolation in the latent space is specified by λ where $\lambda_j \in [z_j^{min}, z_j^{max}]$, $z_j^{min} = \min(\{z_j^{(i)}\}_{i=1}^N)$, $z_j^{max} = \max(\{z_j^{(i)}\}_{i=1}^N)$ and $\lambda_k = 0$ for $k \neq j$. This helps us to understand whether the latent variable z_j encodes a useful feature or a shortcut. Figure 2 illustrates the latent traversal for the *Lemon Quality* dataset.

Visualizing Images for Extreme z_j. To validate the meaning attributed to z_j, we propose to additionally compute the embeddings $z_j^{(i)}$ for all the instances $x^{(i)}$ in the dataset, and identify those instances which minimize or maximize z_j. This step is outlined in the Appendix.

2.3. On the Necessity of Human Judges

Our method requires a human judge to evaluate if the information encoded by the latent variable z_j constitutes a feature or a shortcut. We argue that human interventions are inevitable for ML-based shortcut detection (Zech et al., 2018; Geirhos et al., 2020; Singla & Feizi, 2021). ML models learn relations between input and output based on statistics alone. Unlike humans, they are hardly equipped with prior knowledge of the real world beyond the given dataset and the specified task. However, this prior knowledge may be necessary to evaluate whether a correlation is valid or spurious. Therefore, to assess the candidate feature-target correlations identified by our model, we need a human judge in the last step of our approach. This human supervision is limited to inferring the meaning of a latent dimension from the visualization of its latent traversal.

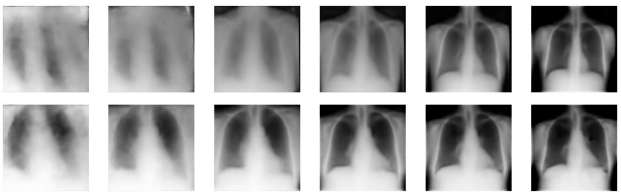


Figure 3. Evaluation of our method on the *COVID-19* dataset. The position of the patient's chest in the X-ray images is revealed as a spurious attribute through latent space traversal. Patients with 'covid' appear closer to the scanner (left) while patients with 'no covid' appear further away such that the black background on the top and bottom sides is visible (right).

3. Evaluation

3.1. Experimental Setup

To obtain a low-dimensional latent representation, we use a Beta-VAE (see Appendix B.1) with $dim(z) \in \{10, 32\}$ depending on the data. For our encoder, we employ a ResNet (He et al., 2016) backbone pretrained on the ImageNet (Russakovsky et al., 2015) dataset. We append two separate linear layers for predicting the parameters μ and σ of the posterior distribution. The decoder of our model consists of 5 hidden layers, each applying transposed convolutions with padding 1, stride 2, and kernel size 3. The channel dimensions of the successive hidden layers in the decoder are chosen as follows: 512, 256, 128, 64, and 32. We apply ReLU activation and batch normalization in the hidden layers and Sigmoid activation in the output layer. While our model can operate on varying input sizes, we use images of size 128×128 in our experiments. We do not perform any data augmentation so as to retain the original characteristics of the input data including any shortcut.

Our model is trained using the Adam (Kingma & Ba, 2014) optimizer with a learning rate of 0.001 and a batch size of 32. Since β determines the tradeoff between reconstruction and sampling quality, we perform hyperparameter tuning on β for each dataset. To estimate the predictiveness of latent features, we append a linear layer to the frozen encoder of a trained Beta-VAE. The resulting classifier with cross-entropy loss is trained with the Adam optimizer and a learning rate of 0.001 on batches of size 32. We ran all experiments on an Nvidia Titan X 12GB GPU.

3.2. Datasets

Our approach is evaluated on six datasets: *Waterbirds* (Sagawa et al., 2019), *Colored MNIST* (Arjovsky et al., 2019), *CelebA* (Liu et al., 2015), *Lemon Quality* (Köroğlu, 2020), *COVID-19* (DeGrave et al., 2021), *ASVspoof* (Wang et al., 2020). Details on these datasets are provided in the Appendix.

Table 1. Shortcut detection results. For every dataset, we denote the latent dimension of a trained VAE which encodes a spuriously correlated attribute. They are picked from a set of candidate dimensions shortlisted based on how predictive (see Section 2.1) they are in relation to all dimensions. The semantic meaning encoded by these candidate dimensions is revealed with a traversal in the latent space (see Section 2.2).

Dataset	Spurious Dimension	Spurious Attribute	MPWD	Predictiveness
ASVspoof	30	Leading silence	2/32	2/32
CelebA	26	Gender	3/32	3/32
Colored MNIST	14	Color	1/32	1/32
COVID-19	1	Patient's position	1/32	2/32
Lemon Quality	2	Zoom level	1/10	2/10
Waterbirds	1	Background	2/32	3/32

3.3. Results

The empirical results are summarized in Table 1. While the particular dimension representing a spurious feature can vary between different training runs, any VAE trained with well-chosen hyperparameters should be able to encode the feature in one of its latent dimensions. The MPWD and the predictiveness of a spurious dimension are reported relative to all latent dimensions. Across all experiments, we conclude that a human judge has to review only the top $k = 3$ latent dimensions with the highest MPWD and predictiveness to identify a shortcut if present in the data.

Our approach correctly identifies the known shortcut in the *COVID-19* dataset. The statistical analysis on the latent space of a trained VAE shows that its latent variable z_1 has the highest MPWD and a high predictiveness $pred(z_1)$ (among top 2). The latent traversal illustrated in Figure 3 reveals that the latent variable z_1 encodes the position of a patient. Similarly, we detect the color shortcut in the *Colored MNIST* dataset, the background shortcut in the *Waterbirds* dataset and the gender shortcut in the *CelebA* dataset. To the best of our knowledge, we are the first to identify the existence of a shortcut in the *Lemon Quality* dataset. As depicted in Figure 2, our method reveals the correlation between lemon quality and zoom level. We make a step towards the generalization of our method to other domains by identifying shortcuts in spectrograms from the *ASVspoof* dataset.

To compare our results to heatmap-based approaches (Zech et al., 2018; Singla & Feizi, 2021), we generate heatmaps (see Appendix) of maximally activating features of the penultimate layer for a standard CNN classifier. The heatmaps for the *Lemon Quality* dataset focus on brown patches for class 'bad quality' and on the background for class 'good quality'. This spatial shortcut is also identified by our method. For *CelebA*, we obtain heatmaps that focus on meaningful facial features, including hair, forehead and eyes. In contrast to our approach, none of these heatmaps make the gender shortcut in the dataset obvious.

3.4. Limitations

Dai and Wipf (2019) have argued that latent representations with independent dimensions do not necessarily correspond to any semantically-meaningful form of disentanglement. Locatello et al. (2019) have theoretically proven that the unsupervised learning of disentangled representations is impossible without inductive biases on the learning approaches and the datasets. The human supervision in our pipeline addresses the concern of Dai and Wipf (2019). Once we quantitatively discover a latent variable that shows a high correlation with a particular class label, a human judge can inspect its latent traversal. The human-in-the-loop approach allows to determine whether the recovered visual pattern in a particular latent dimension matches a real-world concept.

A Beta-VAE can encode a disentangled representation only if the underlying factors of a dataset are independent. The existence of shortcuts in a dataset makes it particularly difficult to fully separate the factors of variation. However, we demonstrate that our approach even works if two or more real-world concepts are entangled in one latent dimension. Figure 3 illustrates how the change of a spurious attribute (e.g. position) can coincide with the change of the main distinguishing factor (e.g. lung appearance). Finally, we note that VAEs are sensitive to the choice of hyperparameters, including the number of latent dimensions and the weights in the loss function.

4. Conclusion

In this paper, we introduce a novel approach to detect spurious correlations in machine learning datasets. Our method utilizes a VAE to discover meaningful features in a given dataset and enables a human judge to identify shortcuts effortlessly. We evaluate our approach on image and audio datasets and successfully reveal the inherent shortcuts. We hope that our work inspires researchers to explore the potential of VAEs in the field of shortcut learning. It would be interesting to see if our method can be extended to other domains.

References

Adjabi, I., Ouahabi, A., Benzaoui, A., and Taleb-Ahmed, A. Past, present, and future of face recognition: A review. *Electronics*, 9(8):1188, 2020.

Alcorn, M. A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4845–4854, 2019.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., and Kalai, A. T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

Brunese, L., Mercaldo, F., Reginelli, A., and Santone, A. Explainable deep learning for pulmonary disease and coronavirus covid-19 detection from x-rays. *Computer Methods and Programs in Biomedicine*, 196:105608, 2020.

Cohen, J. P., Morrison, P., and Dao, L. Covid-19 image data collection. *arXiv preprint 2003.11597*, 2020. URL https://github.com/ieee8023/covid-chestxray-dataset.

Dai, B. and Wipf, D. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.

Dastin, J. Amazon scraps secret ai recruiting tool that showed bias against women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/, 11 2018. (Accessed on 12/26/2022).

DeGrave, A. J., Janizek, J. D., and Lee, S.-I. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ghoshal, B. and Tucker, A. Estimating uncertainty and interpretability in deep learning for coronavirus (covid-19) detection. *arXiv preprint arXiv:2003.10769*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Heaven, W. D. Hundreds of ai tools have been built to catch covid. none of them helped. mit technology review. https://www.technologyreview.com/2021/07/30/1030329/machine-learning-ai-failed-covid-hospital-diagnosis-pandemic/, 07 2021. (Accessed on 11/13/2022).

Hemdan, E. E.-D., Shouman, M. A., and Karar, M. E. Covidx-net: A framework of deep learning classifiers to diagnose covid-19 in x-ray images. *arXiv preprint arXiv:2003.11055*, 2020.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. *Openreview*, 2016.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Kirichenko, P., Izmailov, P., and Wilson, A. G. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

Kolesnikov, A. and Lampert, C. H. Improving weakly-supervised object localization by micro-annotation. *arXiv preprint arXiv:1605.05538*, 2016.

Köroğlu, Y. E. Lemon quality dataset. https://www.kaggle.com/datasets/yusufemir/lemon-quality-dataset, 2020. (Accessed on 11/13/2022).

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(1):1–8, 2019.

LeCun, Y., Cortes, C., and Burges, C. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010.

Liu, Z., Luo, P., Wang, X., and Tang, X. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pp. 4114–4124. PMLR, 2019.

Malhotra, G. and Bowers, J. What a difference a pixel makes: an empirical examination of features used by cnns for categorisation. *Openreview*, 2018.

Müller, N. M., Dieckmann, F., Czempin, P., Canals, R., Böttinger, K., and Williams, J. Speech is silver, silence is golden: What do asvspoof-trained models really learn? *arXiv preprint arXiv:2106.12914*, 2021.

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., and Acharya, U. R. Automated detection of covid-19 cases using deep neural networks with x-ray images. *Computers in biology and medicine*, 121:103792, 2020.

Rajpurkar, P., Chen, E., Banerjee, O., and Topol, E. J. Ai in health and medicine. *Nature Medicine*, 28(1):31–38, 2022.

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.

Sagawa, S., Koh, P. W., Hashimoto, T. B., and Liang, P. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Schörkhuber, C. and Klapuri, A. Constant-q transform toolbox for music processing. In *7th sound and music computing conference, Barcelona, Spain*, pp. 3–64, 2010.

Shetty, R., Schiele, B., and Fritz, M. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8218–8226, 2019.

Singla, S. and Feizi, S. Causal imagenet: How to discover spurious features in deep learning? *arXiv preprint arXiv:2110.04301*, 2021.

Vaserstein, L. N. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Viviano, J. D., Simpson, B., Dutil, F., Bengio, Y., and Cohen, J. P. Saliency is a possible red herring when diagnosing poor generalization. *arXiv preprint arXiv:1910.00199*, 2019.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. *Caltech Online Library*, 2011.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

Wang, X., Yamagishi, J., Todisco, M., Delgado, H., Nautsch, A., Evans, N., Sahidullah, M., Vestman, V., Kinnunen, T., Lee, K. A., et al. Asvspoof 2019: A large-scale public database of synthesized, converted and replayed speech. *Computer Speech & Language*, 64:101114, 2020.

Xiao, K., Engstrom, L., Ilyas, A., and Madry, A. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020.

Yang, J., Li, S., Wang, Z., Dong, H., Wang, J., and Tang, S. Using deep learning to detect defects in manufacturing: a comprehensive survey and current challenges. *Materials*, 13(24):5755, 2020.

Yang, W., Kirichenko, P., Goldblum, M., and Wilson, A. G. Chroma-vae: Mitigating shortcut learning with generative classifiers. *Advances in Neural Information Processing Systems*, 35:20351–20365, 2022.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., and Oermann, E. K. Confounding variables can degrade generalization performance of radiological deep learning models. *arXiv preprint arXiv:1807.00431*, 2018.

Zhang, M., Sohoni, N. S., Zhang, H. R., Finn, C., and Ré, C. Correct-n-contrast: A contrastive approach for improving robustness to spurious correlations. *arXiv preprint arXiv:2203.01517*, 2022.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

A. Related Work

Machine Learning Shortcuts. Geirhos et al. (2020) locate the origin of shortcuts in the data and the learning process of ML models. The inherent contextual bias in datasets provides opportunities for shortcuts. Natural image datasets contain spurious correlations between the target variable and the background (Xiao et al., 2020), the object poses (Alcorn et al., 2019) or other co-occurring distracting features (Kolesnikov & Lampert, 2016; Shetty et al., 2019). In discriminative learning, a model uses a combination of features to make a prediction. Following the "principle of least effort", models tend to rely only on the most obvious features, which often correspond to shortcuts (Geirhos et al., 2020). For instance, convolutional neural networks (CNNs) trained on ImageNet were found to be biased towards the texture of the objects instead of their shapes (Geirhos et al., 2018). In another example, it was discovered that CNNs solely used the location of a single pixel to distinguish between object categories (Malhotra & Bowers, 2018).

Identification of Shortcuts. The identification of shortcuts in supervised machine learning is still in its infancy. Zech et al. (2018) use activation heatmaps to reveal the spurious features learned by CNNs trained on X-ray images. As outlined by Viviano et al. (2019), saliency maps can only explain spatial shortcuts, e.g. source tags on images (Lapuschkin et al., 2019), but fail to identify more complicated ones, e.g. people's gender (Sagawa et al., 2019). Singla and Feizi (2021) select the highest activations of neurons in the penultimate layer of a CNN classifier and back-project them onto the input images. The resulting heatmaps highlight the features in the image that maximize neural activations. Under human supervision, the highlighted regions, for a subset of images, are labelled as 'core' (part of the object definition) or 'spurious' (only co-occurring with the object). Using this labelled dataset, they train a classifier to automatically identify the core and spurious visual features for a larger dataset. To diagnose shortcut learning, Geirhos et al. (2020) suggest performing o.o.d. generalization tests. Evaluating the model on o.o.d. real-world data in addition to the i.i.d. test set reveals whether a model is actually generalizing on the intended features or simply learning shortcuts from the training data.

Robustness against Shortcuts. To learn representations robust to spurious correlations, Zhang et al. (2022) propose a two-stage contrastive approach. The method first identifies training samples from the same class with different model predictions. Contrastive learning then ensures that the hidden representations for samples of the same class with initially different predictions become close to each other. Kirichenko et al. (2022) suggest to retrain the last layer of a classifier on a small dataset without spurious correlations. While the reweighting of the last layer reduces the model's reliance on background and texture information, the requirement of a shortcut-free subset remains a limitation of this approach. Our method could serve as a preliminary step by identifying the shortcuts to be addressed.

B. Methodology

B.1. Variational Autoencoder

Our approach for shortcut detection is based on variational autoencoders (Kingma & Welling, 2013) (VAEs), which are probabilistic generative models that learn the underlying data distribution in an unsupervised manner. A VAE attempts to model the marginal likelihood of an observed variable x:

$$p_\theta(x) = \int p_\theta(z) p_\theta(x|z)\, dz \tag{2}$$

The unobserved variable z lies in a latent space of dimensionality $d = \dim(z)$, $d \ll \dim(x)$. Each instance $x^{(i)}$ has a corresponding latent representation $z^{(i)}$.

The model assumes the prior over the latent variables to be a multivariate normal distribution $p_\theta(z) = \mathcal{N}(z; 0, I)$ resulting in independent latent factors $\{z_j\}_{j=1}^{d}$. The likelihood $p_\theta(x|z)$ is modelled as a multivariate Gaussian whose parameters are conditioned on $z \sim p_\theta(z)$ and computed using the VAE decoder. As outlined by Kingma and Welling (2013), the true posterior $p_\theta(z|x)$ is intractable and hence approximated with a variational distribution $q_\phi(z|x)$. This variational posterior is chosen to be a multivariate Gaussian

$$q_\phi(z|x^{(i)}) = \mathcal{N}(z; \mu^{(i)}, (\sigma^{(i)})^2 I) \tag{3}$$

whose mean $\mu^{(i)} \in \mathbb{R}^d$ and standard deviation $\sigma^{(i)} \in \mathbb{R}^d$ are obtained by forwarding $x^{(i)}$ through the VAE encoder. The objective function is composed of two terms. A Kullback-Leibler (KL) divergence term ensures that the variational posterior

distribution remains close to the assumed prior distribution.

$$D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) = -\frac{1}{2}\sum_{j=1}^{d}\left(1 + \log((\sigma_j^{(i)})^2) - (\sigma_j^{(i)})^2 - (\mu_j^{(i)})^2\right) \tag{4}$$

A log-likelihood loss, on the other hand, helps to accurately reconstruct an input image $x^{(i)}$ from a sampled latent variable $z^{(i)} = \mu^{(i)} + \sigma^{(i)} \odot \epsilon$ where $\epsilon \sim \mathcal{N}(0, I)$. Thus the encoder and decoder are trained to maximize the following objective function:

$$\mathcal{L}(\theta, \phi; x^{(i)}) = -D_{KL}(q_\phi(z|x^{(i)})||p_\theta(z)) + log\ p_\phi(x^{(i)}|z^{(i)}) \tag{5}$$

For modelling images, both the encoder and decoder of a VAE consist of CNNs.

Higgins et al. (2016) introduce the Beta-VAE to better learn independent latent factors. The authors propose to augment the vanilla VAE loss in Equation (5) by weighing the KL term with a hyperparameter β. Choosing $\beta > 1$ enables the model to learn a more efficient latent representation of the data with better disentangled dimensions.

B.2. Hyperparameter Tuning of Beta-VAE

Following Higgins et al. (2016), we tune the hyperparameters $dim(z)$ and β of the Beta-VAE for every dataset. To achieve maximum disentanglement, the number of latent dimensions should match the number of factors of variation in a dataset. This number is usually not known a priori. Choosing a latent space of too high dimensionality leads to a lot of uninformative dimensions with low variance. A latent space of too few latent dimensions, on the other hand, leads to entangled representations of features in the latent space. Therefore, it is necessary to identify the optimal number of latent dimensions to achieve maximally disentangled factors, ideally one in each dimension.

To obtain the best combination of β and $dim(z)$ for a given dataset, we first train a Beta-VAE with $\beta = 1$ and $dim(z) = 32$ on all datasets. We then compare the variance of the distribution in each dimension to that of a Gaussian prior. In the presence of dimensions with relatively low variance, we reduce the number of latent dimensions. Once we find a Beta-VAE with consistently informative dimensions, i.e. with the variance comparable to the prior, we fix $dim(z)$ and focus on fine-tuning β. As outlined by Higgings et al. (2016), for relatively low values of β, the VAE learns an entangled latent representation since the capacity in the latent space is too high. On the other hand, for relatively high values of β, the capacity in the latent space becomes too low. The VAE performs a low-rank projection of the true data generative factors and again learns an entangled latent representation. This renders some of the latent dimensions uninformative.

We find the highest possible β for which all dimensions of the VAE remain informative with a variance close to the prior. In line with the findings of Higgings et al. (2016), $\beta > 1$ is required for all datasets to achieve good disentanglement (see Table 2).

Table 2. Beta-VAE hyperparameters

Dataset	$dim(z)$	β
ASVspoof	32	1.25
CelebA	32	10.0
Colored MNIST	32	2.5
COVID-19	32	1.5
Lemon Quality	10	3.0
Waterbirds	32	1.75

B.3. Visualizing Images for Extreme z_j

To validate the meaning attributed to z_j, we propose to additionally compute the embeddings $z_j^{(i)}$ for all the instances $x^{(i)}$ in the dataset, and identify those instances which minimize or maximize z_j. Particularly, we perform $\arg\text{sort}_i(\{z_j^{(i)}\}_{i=1}^N)$ for every dimension j in the latent space and display the input images $x^{(i)}$ corresponding to the first l and the last l indices of the sorted values. Figure 4 depicts $l = 27$ images of the *Lemon Quality* dataset with minimum and maximum values in latent dimension 2.

(a) Inputs from the *Lemon Quality* dataset which minimize z_2.

(b) Inputs from the *Lemon Quality* dataset which maximize z_2.

Figure 4. Real-world images from the training dataset that correspond to the minimum and maximum encoded values in latent dimension 2. While bad-quality lemons are mostly captured as close-up shots, good-quality lemons are photographed from a distance.

C. Evaluation

C.1. Datasets

We evaluate our approach on datasets with artificially introduced shortcuts to demonstrate its ability to identify spurious correlations. Our method can also be applied to real-world datasets to reveal previously unknown shortcuts. We perform a train-val-test split in the ratio 80:10:10 on each dataset unless explicitly specified.

Waterbirds. Sagawa et al. (2019) extract birds from the Caltech-UCSD Birds-200-2011 dataset (Wah et al., 2011) and combine them with background images from the Places dataset (Zhou et al., 2017). As a result, 95% of the water birds appear on a water background while 95% of the land birds appear on a land background. Since this spurious correlation is only introduced in the train set of the *Waterbirds* dataset consisting of 4,795 samples, we use the same to create the train-val-test splits for our experiments.

Colored MNIST. Arjovsky et al. (2019) inject color as a spurious attribute into the MNIST (LeCun et al., 2010) dataset. Following this idea, Zhang et al. (2022) create a colored MNIST dataset consisting of five subsets with five associated colors. A fraction p_{corr} of the training samples are assigned colors based on the group they belong to. The remaining samples are assigned a random color. We follow the color assignment in (Zhang et al., 2022) and choose $p_{corr} = 0.995$ for coloring the 70,000 MNIST samples.

CelebA. The CelebFaces Attributes Dataset (Liu et al., 2015) contains 202,599 images of celebrities, each annotated with 40 facial attributes. Sagawa et al. (2019) train a classifier to identify the hair color of the celebrities and discover that the target classes (blond, dark) are spuriously correlated with the gender (male, female). We stick to the setup specified in (Sagawa et al., 2019) with the official train-val-test splits of the *CelebA* dataset.

Lemon Quality. To demonstrate the detection of shortcuts in quality control, we apply our method on the *Lemon Quality* dataset (Köroğlu, 2020). The dataset consists of 2,533 images labelled with one of three classes, namely 'good quality', 'bad quality', and 'empty background'.

COVID-19. Following (Brunese et al., 2020; Ghoshal & Tucker, 2020; Hemdan et al., 2020; Ozturk et al., 2020; DeGrave et al., 2021), we obtain a dataset with 112,528 samples for COVID-19 detection by combining COVID-19-positive radiographs from the GitHub-COVID repository (Cohen et al., 2020) and COVID-19-negative radiographs from the ChestX-ray14 repository (Wang et al., 2017).

ASVspoof. The ASVspoof 2019 Challenge Dataset (Wang et al., 2020) is used to train and benchmark systems for the detection of spoofed audio and audio deepfakes. Müller at al. (2021) observe that the length of the silence at the beginning of an audio sample differs significantly between benign and malicious data. Previous deepfake detection models have exploited this shortcut. We chose a subset of the training data (benign audio and attack $A01$), which results in a binary classification dataset comprising 6,380 samples. We transform the audio samples to CQT spectrograms (Schörkhuber & Klapuri, 2010), and obtain a frequency-domain representation with 257 logarithmically spaced frequency bins, capturing up to 8 kHz (Nyquist frequency given the input is 16 kHz).

C.2. Results

We provide illustrations of the latent space traversal for the *Colored MNIST*, *CelebA* and *ASVspoof* dataset.

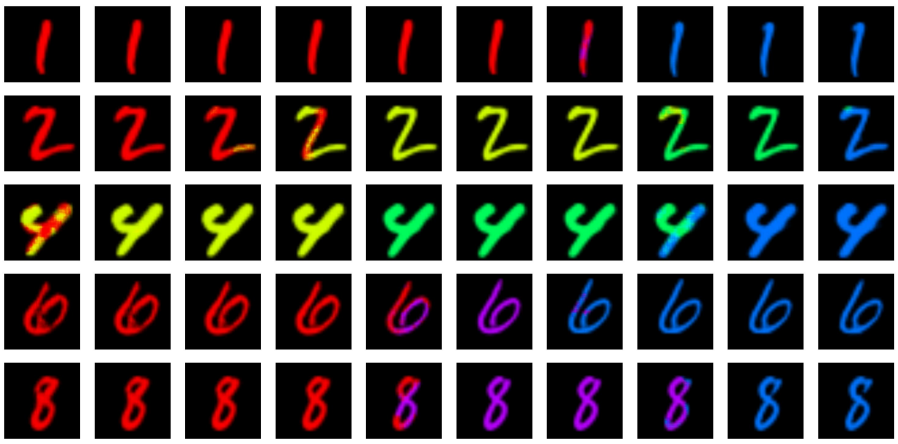


Figure 5. Evaluation of our method on the *Colored MNIST* dataset. The illustrated dimension represents the attribute 'color'. The minimum values encode the colors red (left) while the maximum values encode the color blue (right).

Figure 6. Evaluation of our method on the *CelebA* dataset. The latent traversal reveals the meaning of dimension 26 as 'gender'. The high correlation (as measured in terms of MPWD and predictiveness) of this attribute with the target variable 'hair color' indicates the existence of a spurious correlation.

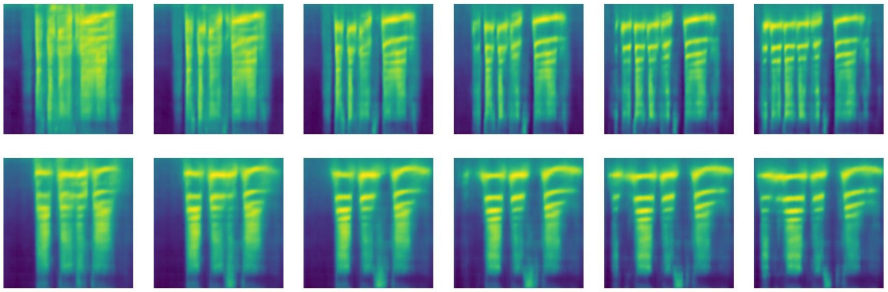


Figure 7. Evaluation of our method on the *ASVspoof* dataset. Latent traversal reaffirms the known spurious correlation between the leading silence in the audio and the target class. Leading silence (left) in the spectrogram is an indicator of benign audio samples while no leading silence (right) is common in spoofs.

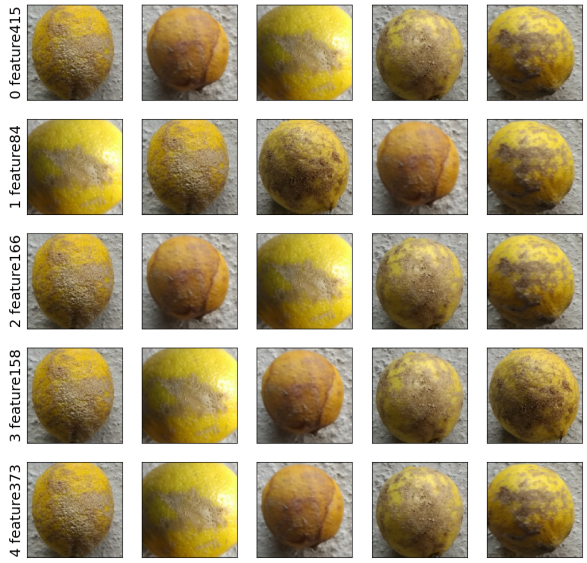
C.3. Comparison

To further evaluate our approach, we compare it to one of the most established Explainable AI techniques, namely heatmaps (Zech et al., 2018; Singla & Feizi, 2021). We train a CNN $C(x) = L(F(x))$, which consists of a convolutional feature extractor F and a linear model L, where L computes a linear combination of the features $F(x)$. We design F to include five convolutional layers with 32, 64, 128, 256, and 512 filters, using a kernel of size 3×3. Each of these layers is followed by ReLU activation and max pooling.

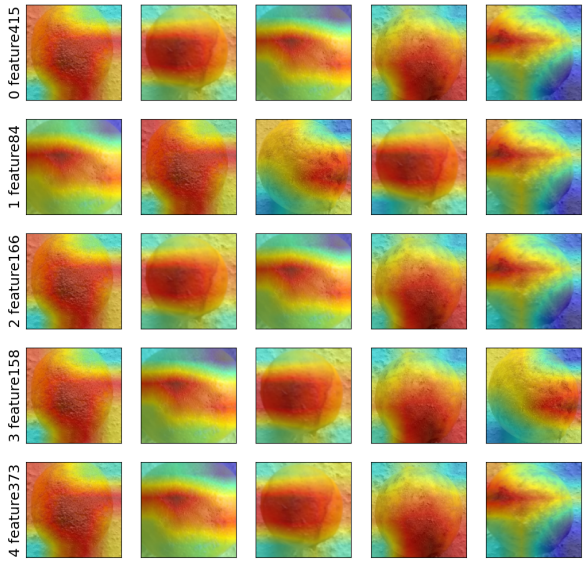
For each image x, we obtain features $F(x)$ in the penultimate layer. We compute the average contribution of each feature $F(x)_i$ towards the activation $L(F(x)_i)_c$ of a particular class c by multiplying it with the associated weight. The average contribution of each feature to the prediction of a certain class is computed over all images of that class. We pick the top 5 most predictive features and then find the top 5 images that yield the highest activation. We compute the heatmaps by projecting these top features onto the original images.

The heatmaps for the *Lemon Quality* dataset (see Figure 8) indicate that the maximally activated features for class 'bad quality' are brown patches on the fruit, which is a meaningful feature. However, the heatmaps for class 'good quality' (see Figure 9) largely focus on the background. Therefore, a heatmap-based approach is able to hint at the spatial shortcut in the *Lemon Quality* dataset. This shortcut is also successfully identified by our method as described in Section 3.3.

Additionally, we compute the heatmaps on the *CelebA* dataset. The results are illustrated in Figure 10 and Figure 11. We can see that the model focuses on the hair in order to classify the 'hair color', as well as other facial features such as the forehead and eyes. However, this does not enable to identify the gender shortcut. In contrast, our approach could successfully reveal this spurious correlation (see Figure 6).

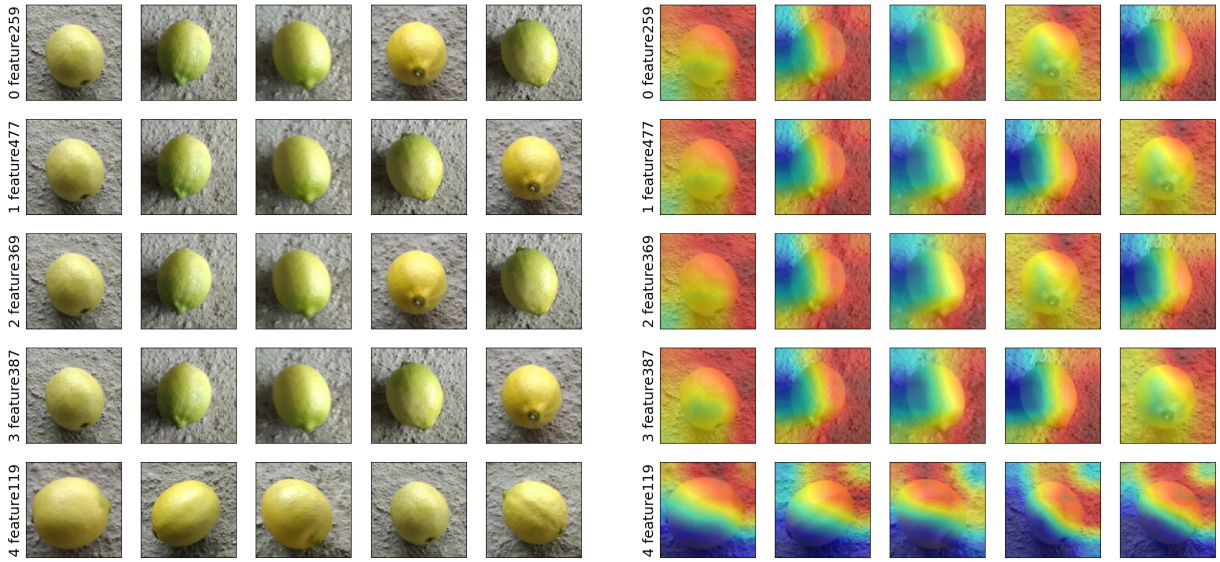


(a) For the 5 most predictive features (from top to bottom) we display the images which yield the highest activations (from left to right).



(b) For the 5 most predictive features (from top to bottom) we display the heatmaps on images which yield the highest activations (from left to right).

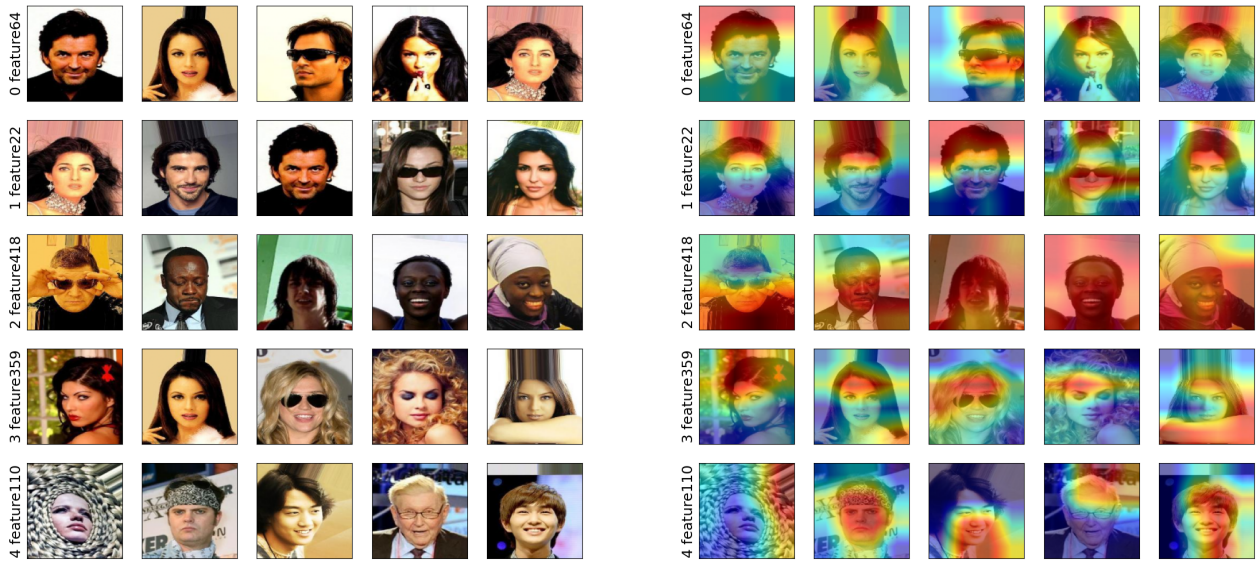
Figure 8. Original images along with heatmaps for class 'bad quality' of the *Lemon Quality* dataset. A standard CNN classifier identifies the brown patches as a meaningful feature for classification.



(a) For the 5 most predictive features (from top to bottom) we display the images which yield the highest activations (from left to right).

(b) For the 5 most predictive features (from top to bottom) we display the heatmaps on images which yield the highest activations (from left to right).

Figure 9. Original images along with heatmaps for class 'good quality' of the *Lemon Quality* dataset. A standard CNN classifier leverages the background as a spatial shortcut for classification.

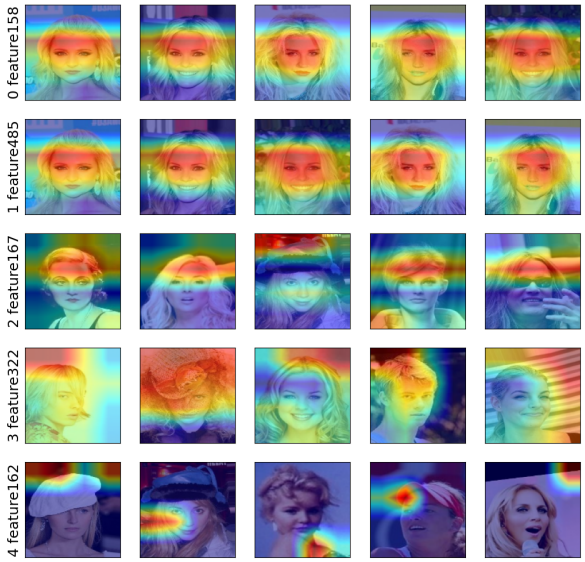


(a) For the 5 most predictive features (from top to bottom) we display the images which yield the highest activations (from left to right).

(b) For the 5 most predictive features (from top to bottom) we display the heatmaps on images which yield the highest activations (from left to right).

Figure 10. Original images along with heatmaps for class 'dark hair' of the *CelebA* dataset. While a few heatmaps focus on the valid attribute hair, none of them reveal the gender shortcut.

(a) For the 5 most predictive features (from top to bottom) we display the images which yield the highest activations (from left to right).



(b) For the 5 most predictive features (from top to bottom) we display the heatmaps on images which yield the highest activations (from left to right).

Figure 11. Original images along with heatmaps for class 'blond hair' of the *CelebA* dataset. The heatmaps highlight meaningful facial features like forehead and eyes but fail to reveal the gender shortcut.