Identifying Spurious Correlations using Counterfactual Alignment

Joseph Paul Cohen[*†] Stanford University
Louis Blankemeier Stanford University
Akshay Chaudhari Stanford University

1. Introduction

Challenges related to neural network generalization and fairness often arise due spurious correlations (aka shortcut learning [4]). Spurious correlations can lead to models making decisions based on factors not aligned with expectations of the model creator, causing poor generalization.

Our objective is to understand the features used in the predictions of black-box classifiers, including pre-trained models, as this is a common use case encountered by practitioners. In this analysis, we utilize counterfactual (CF) images, synthetic images simulating a change in the class label of an image [9]. These synthetic images have features modified such that the prediction of the classifier changes. We can then view the synthetic images to understand the reasons that a prediction was made.

Specifically, we are interested in CF images that are directly generated using the gradients of a classifier [1, 2, 5]. Generating CF images with respect to a classifier is rooted in similar logic to adversarial examples. However, the distinction lies in the constraint that CF images remain within the data manifold of plausible images. The latent space of an autoencoder provides such a data manifold. This approach enables us to study the specific features used by a given classifier on a particular input. By keeping the classifier and autoencoder independent, different classifiers can be analysed using a singular fixed autoencoder as a reference point.

The task of interpreting these CF images presents new challenges of scale. Spurious correlations may only exist in a handful of samples where rare features occur together. Locating these samples requires investigating the features used for each prediction, which can be automated using the approaches we present in this study.

2. CF Alignment Methodology

This work proposes the CF alignment approach as described in Fig. 1. For this approach, we generate counterfactuals using Latent Shift [2]. This approach

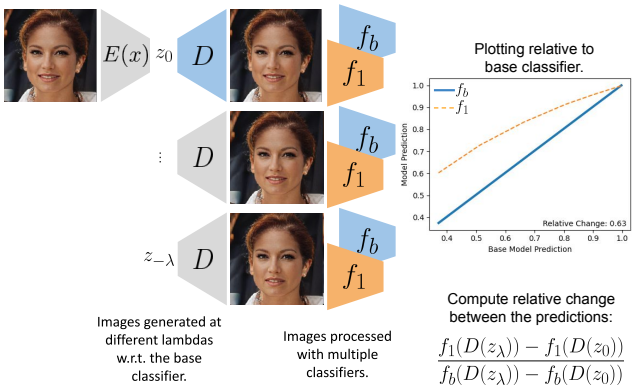


Figure 1. Overview of the alignment methodology. An image is encoded, reconstructed, and then processed by a classifier. The counterfactual is generated by subtracting the gradient of the classifier output w.r.t. the latent representation. The resulting representation is reconstructed back into an image. The reconstructed images are processed with multiple classifiers and the classifier outputs can be plotted side by side to study their alignment. The base model value can be used as the x-axis to more easily compare it to the predictions of another classifier. The output changes can be quantified and compared using relative change.

requires an encoder/decoder model $D(E(x))$ where E is the encoder and D is the decoder, and a classifier f which predicts a target y as follows: $y = f(x)$. The autoencoder and the classifier are trained independently without specific requirements except for being differentiable and operating on the same data domain. To compute a counterfactual, an input image x is encoded using $E(x)$, to produce a latent representation z. Perturbations of the latent space are computed for a base classifier f_b in Eq. 1 which is then used to produce λ-shifted samples as shown in Eq. 2. λ is determined using an iterative search where the value is decreased in steps until the classifier's prediction is reduced by 0.6 or begins to increase, as this was found this sufficiently change the prediction.

$$z_\lambda = z_0 - \lambda \frac{\partial f_b(D(z_0))}{\partial z} \quad (1) \qquad x'_\lambda = D(z_\lambda) \quad (2)$$

These counterfactual samples can then be input into downstream classifiers that predict various attributes (e.g. f_1, f_2) and observe how their predictions change.

[*]joseph@josephpcohen.com
[†]Work not related to position at Amazon.

Relative Change To algorithmically quantify the relationship between the base classifier and each downstream classifier, we employ the Relative Change metric (Eq. 3):

$$R(f_1, f_b, z_0) = \frac{f_1(D(z_\lambda)) - f_1(D(z_0))}{f_b(D(z_\lambda)) - f_b(D(z_0))} \quad (3)$$

Here, $D(z_\lambda)$ and $D(z_0)$ represent the reconstructions of the latent representations z_λ and z_0 (perturbed and original, respectively) by the decoder D. In our experiments, we opted for Relative Change over the traditional correlation measure, motivated by the observation that correlation occasionally yielded false positives when only a slight change in the prediction of f_1 occurred compared to the base classifier f_b. Relative Change captures the direction as well as the magnitude of the change, offering a more pragmatic assessment of the impact of counterfactual perturbations on downstream classifiers.

The Relative Change metric is a key contribution that allows us to make sense of the CF alignment results at scale over large datasets.

3. Experiments

The experiments in this work are performed on the CelebA dataset [7]. The pre-trained face classifiers used in this work are sourced from [8]. They were trained on the CelebA dataset [7] to predict 40 different facial attributes. We leverage the VQ-GAN autoencoder from [3] trained on the FacesHQ dataset, which combines the CelebA dataset [7] and the Flickr-Faces-HQ (FFHQ) dataset [6]. The resolution of this model is 256x256.

Studying specific examples In order to gain an intuition for this approach we present an example showing an image x and a CF version $x_{-pointy\_nose}$ of that image for the classifier *pointy_nose* shown in Fig. 2. The CF alignment plots are shown below with their relative change score indicating positive or negative relative change.

Validation by inducing spurious correlations In order to further verify the CF alignment approach, we construct a classifier with a known spurious correlation and then demonstrate that this bias is observable by CF alignment. A spurious correlation can be induced in the classifier by composing classifiers:

$$f_{biased}(x) = f_{smiling}(x) + 0.3 f_{arched\_eyebrows}(x) \quad (4)$$

The CF alignment plot for the base *smiling* classifier predictions in Fig. 3a show that there is a negligible relationship with *arched_eyebrows*, having a relative change of 0.01.

The resulting biased classifier can be observed using eyebrow features in Fig. 3b. The CF alignment plot shows that the prediction of *arched_eyebrows* now changes and is aligned with *smiling* with a high relative change of 0.97.

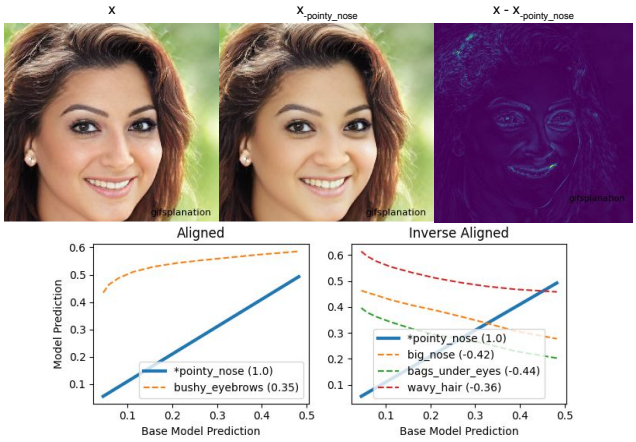
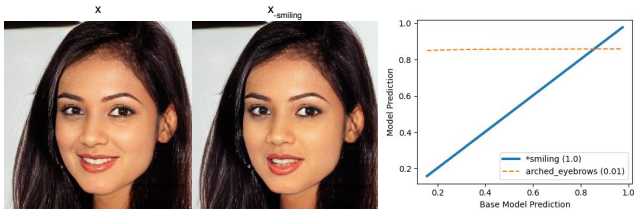
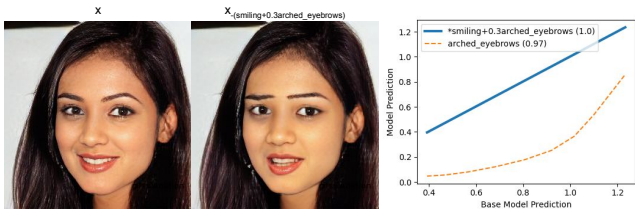


Figure 2. CF alignment for *pointy_nose* showing an inverse alignment with *big_nose* and potential spurious relationships with eyebrows eyes and hair. The relative change is shown next to each classifier name.



(a) CF for *smiling* showing eyebrows are unchanged. The horizontal line indicates the prediction of *arched_eyebrows* is not influenced by the features used for *smiling* in this image.



(b) CF for the modified *smiling* classifier with an arched eyebrow spurious correlation showing that eyebrows are now changed.

Figure 3. Example of detecting a spurious correlation in a biased classifier. The classifier is biased with arched eyebrows and this is observed in the alignment plot as well as in the counterfactual image. In order to decrease the prediction of the (now biased) classifier, the CF image can remove the arched eyebrows as well.

4. Conclusion

In this work we propose counterfactual (CF) alignment along with the relative change metric. These methods enable us to reason about the feature relationships between classifiers in aggregate and to locate specific examples with spurious correlations. Overall, the proposed approach may serve as an end-to-end or human-in-the-loop system to automatically detect and quantify spurious correlations for image classification tasks.

References

[1] Rachana Balasubramanian, Samuel Sharpe, Brian Barr, Jason Wittenbach, and C. Bayan Bruss. Latent-CF: A Simple Baseline for Reverse Counterfactual Explanations. In *Neural Information Processing Systems (NeurIPS) Fair AI in Finance Workshop*, 2020. 1

[2] Joseph Paul Cohen, Rupert Brooks, Sovann En, Evan Zucker, Anuj Pareek, Matthew P. Lungren, and Akshay Chaudhari. Gifsplanation via Latent Shift: A Simple Autoencoder Approach to Counterfactual Generation for Chest X-rays. *Medical Imaging with Deep Learning*, 2021. 1

[3] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *Computer Vision and Pattern Recognition*, 2021. 2

[4] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. 1

[5] Shalmali Joshi, Oluwasanmi Koyejo, Been Kim, and Joydeep Ghosh. xGEMs: Generating Examplars to Explain Black-Box Models, 2018. 1

[6] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015. 2

[8] Simon Vandenhende, Stamatios Georgoulis, Bert De Brabandere, and Luc Van Gool. Branched Multi-Task Networks: Deciding What Layers To Share. *British Machine Vision Conference*, 2020. 2

[9] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review. In *Neural Information Processing Systems (NeurIPS) Retrospectives Workshop*, 2020. 1