Unifying Image Counterfactuals and Feature Attributions with Latent-Space Adversarial Attacks

Jeremy Goldwasser[1]; Giles Hooker[2]
[1]University of California, Berkeley, [2]University of Pennsylvania

Abstract

Counterfactuals are a popular framework for interpreting ML predictions. Drawing connections to adversarial attacks, we propose an easy-to-implement method, Counterfactual Attacks, that generates high-quality image counterfactuals. In addition, given an auxiliary dataset of image descriptors, we show how to accompany counterfactuals with feature importance scores. These can be aggregated into global counterfactual explanations that highlight the overall features driving model predictions.

Background

Counterfactual explanations modify an input (e.g. image) in a concise, coherent way that causes the prediction to flip. Their *what-if* nature renders them useful to understand models' causal behavior, at least in a local sense.

Typical approaches, designed for tabular data, find counterfactuals via gradient-based optimization. For computer vision models, however, such techniques may yield *adversarial attacks* – indistinguishable images with wildly different predictions. To circumvent this issue, prior works operate in a meaningful latent space learned by generative models. While effective, existing methods are cumbersome to implement.

Moreover, they only offer qualitative explanations of individual images. No prior works ground image changes with objective metrics, or aggregate them to explain models' global reasoning.

Counterfactual Attacks

Our algorithm enables SOTA generative models to be trained off-the-shelf, unlike previous methods. Its counterfactuals are identified via adversarial attacks on the natural image manifold.

```
Require: Input $x$, predictor $f$, target $y' > f(x)$, encoder $\mathcal{E}$, generator $\mathcal{G}$, learning rate $\eta$
Ensure: Counterfactual $x'$ such that $f(x') \geq y'$
$z \leftarrow \mathcal{E}(x)$
while $f(\mathcal{G}(z)) \leq y'$ do
    $z \leftarrow z + \eta \nabla_z f(\mathcal{G}(z))$        {Latent gradient ascent}
end while
$x' \leftarrow \mathcal{G}(z)$        {Generate counterfactual image}
return $x'$
```

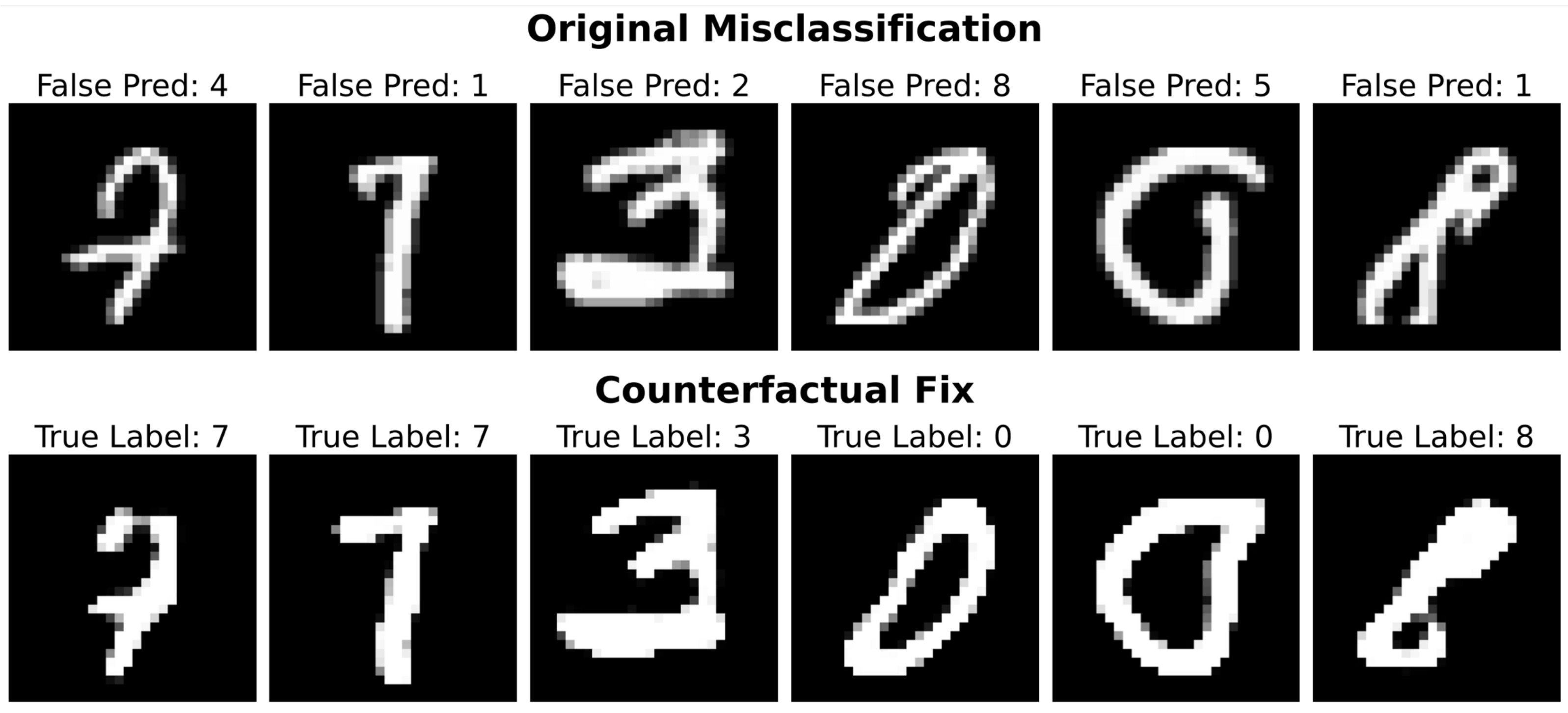

Figure 1. Counterfactuals highlight the reasons that MNIST images were misclassified. Generative model is a VAE.

Local Feature Attributions

Many image datasets contain labeled attributes like facial features. We demonstrate how to use these labels to quantify the content of counterfactual explanations. First, fit logistic regression models $g_a(\cdot)$ predicting attributes a from images' low-dimensional latent representations. Then, the importance of each attribute may be scored as

$$\phi_a(x, x') = g_a(x') - g_a(x).$$

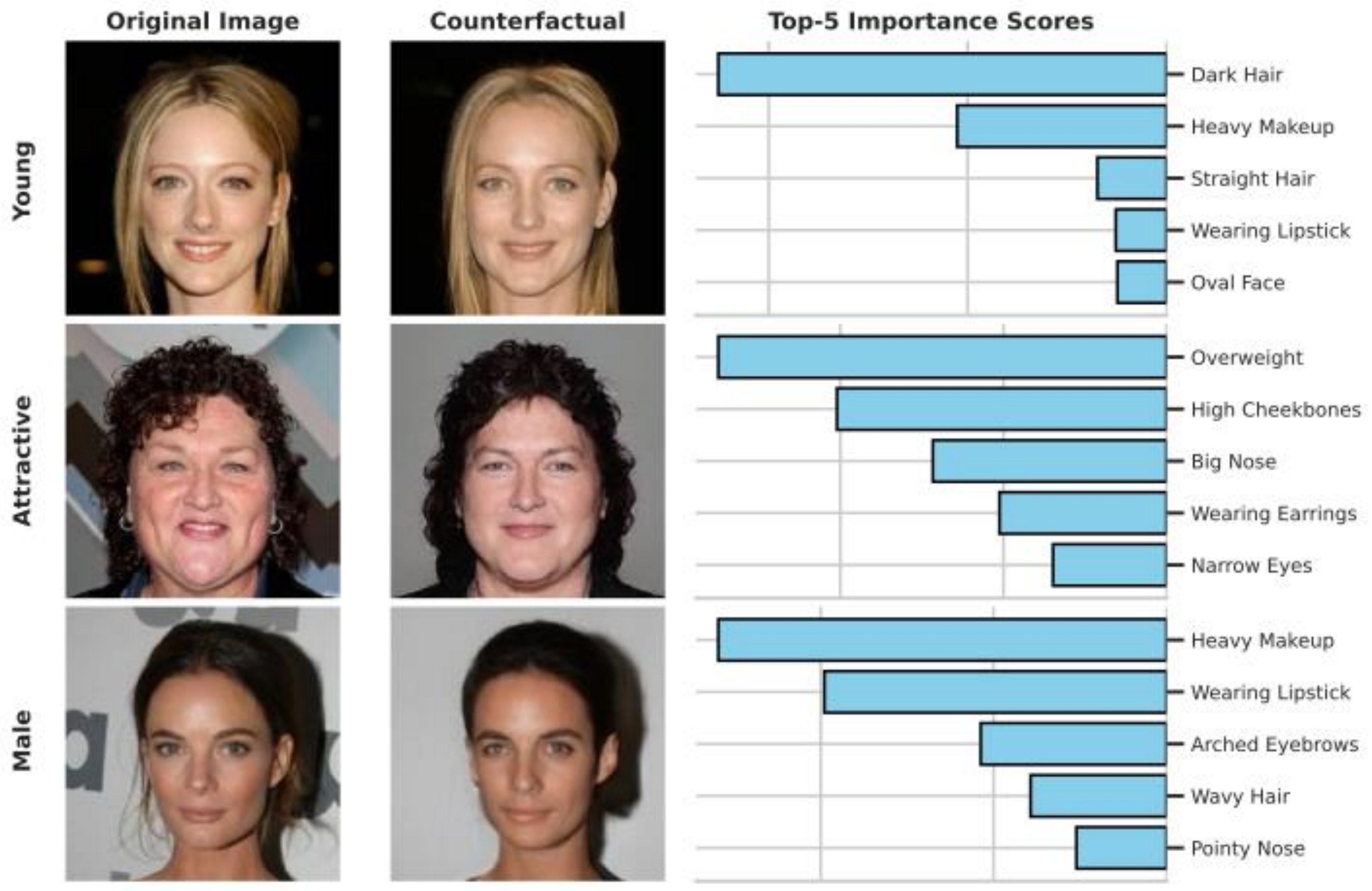

Figure 2. Counterfactual Attacks run on CelebA age, attractiveness, and gender classifiers, using StyleGAN3. Labels identify facial features driving prediction.

Global Feature Attributions

This objective measure of variable importance may be extended beyond local explanations. Averaging over many samples, the global score for a binary classifier is

$$\psi_a = \sum_{i=1}^{n} \phi_a(x_i, x_i') s_i, \text{ where } s_i = \begin{cases} 1, & \text{if } f(x_i) < 0.5 \\ -1, & \text{otherwise} \end{cases}$$

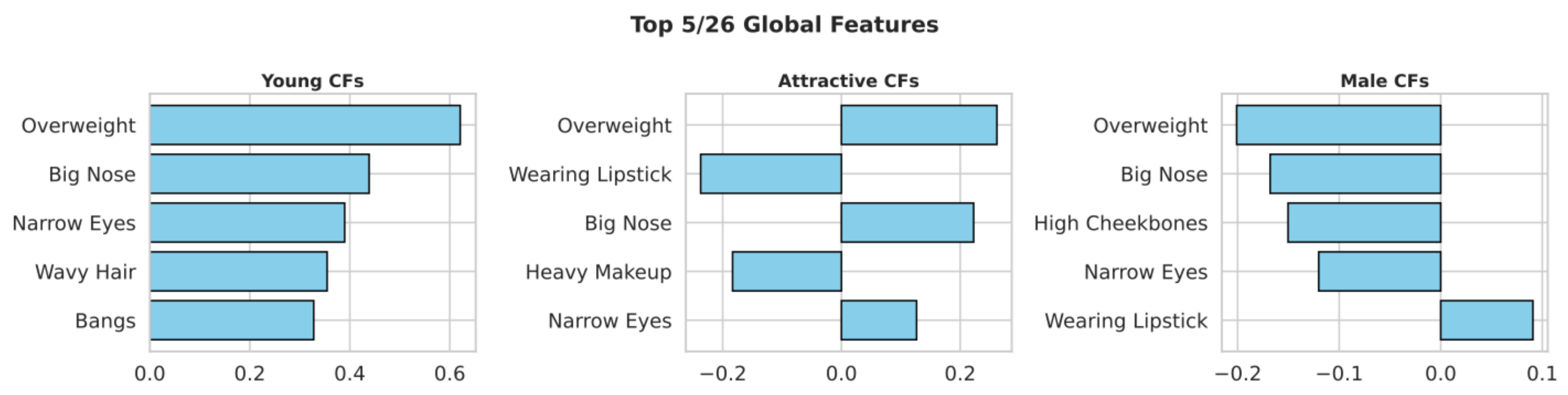

Figure 3. Global scores indicate the most important facial features for the 3 classifiers.

Discussion

Counterfactual methods like ours deliver actionable insights into model behavior. They may inform individuals on how to produce a desired result – for example, to have more legible handwriting, or present themselves more youthfully. Developers may also benefit from understanding the patterns a model dwells on, as facilitated by our global importance scores. Exposing unwanted behaviors may guide efforts to improve a model, or prevent it from being deployed. Otherwise, a demonstrably safe model may be pushed for adoption. These explanations may even enable new patterns to be discovered from image data.