Detecting Bias with Generative Counterfactual Face Attribute Augmentation

Emily Denton
Google Research
dentone@google.com

Ben Hutchinson
Google Research
benhutch@google.com

Margaret Mitchell
Google Research
mmitchellai@google.com

Timnit Gebru
Google Research
tgebru@google.com

Abstract

We introduce a simple framework for identifying biases in a smiling attribute classifier. Our method poses counterfactual questions of the form: how would the prediction change if this face characteristic had been different? We leverage recent advances in generative adversarial networks to build a realistic generative model of face images that affords controlled manipulation of specific image characteristics. We introduce a set of metrics that measure the effect of manipulating a specific property of an image on the output of a trained classifier. Empirically, we identify several different factors of variation that affect the predictions of a smiling classifier trained on CelebA.

1. Introduction

Recent studies have uncovered systematic biases in a variety of computer vision systems. For example, an analysis of commercial gender classification systems demonstrated disparities with respect to perceived gender and skin type, with darker-skinned females exhibiting the highest error rates [1, 20]. An analysis of state-of-the-art object detection systems found decreased pedestrian detection accuracy on darker skin tones [27]. Gender-based bias has also been observed in image captioning [10] and image classification [28].

In light of this research, there is a growing need for methods that facilitate auditing and explaining undesirable skews/stereotypes in computer vision models. We propose a novel framework for identifying biases of a face attribute classifier. Specifically, we consider a classifier trained to predict the presence or absence of a smile on a face image. However, our approach generalizes to a broad range of face-centric classification problems. Our method, which we refer to as *counterfactual face attribute manipulation*, tests how the prediction of a trained classifier would change if a characteristic of the face were different. For example, this approach allows us to ask "Would the smiling prediction change if this person's hair had been longer?"

CVPR 2019 Workshop on Fairness Accountability Transparency and Ethics in Computer Vision.

Our method leverages recent advances in generative models and post-hoc interepretability to facilitate an ethics-informed audit exploring normative constraints on what *should* be independent of the smiling attribute. First, we build a directed generative model of faces that maps latent codes, sampled from a fixed prior distribution, to images. Next, we infer direction in latent code space that correspond to manipulations of a particular image attribute in pixel space. We refer to these vectors as *attribute vectors*. Traversing these directions in latent space provides a straightforward method of generating images that smoothly vary one characteristic of a face, while preserving others. We propose several metrics that measure the sensitivity of a trained classifier to manipulations imposed by attribute vectors.

We test our method on a classifier trained to predict the smiling attribute from faces in the CelebA dataset [17] and find our approach reveals several biases in the smiling classifier.

2. Ethical Considerations

Intended use and limitations: The techniques proposed in this paper can be applied to detect and analyze unintended and undesirable bias in a wide variety of face-centric computer vision systems. While this type of analysis is an important part of designing fair and inclusive technology, it is not sufficient. Rather, it must be part of a larger, socially contextualized project to critically assess ethical concerns relating to facial analysis technology.

In this work, we focus our attention on smiling detection because we believe it has a range of beneficial applications and limited harmful applications, from aiding in the selection of images from a stream of images to augmented reality applications where a smile triggers added features. On the other hand, there are many facial analysis systems that do have harmful applications where foreseeable risk outweighs foreseeable benefit, e.g., where the facial analysis task itself has the potential to perpetuate, reinforce and amplify societal injustices, regardless of how balanced the classification performance is with respect to different demographics. For example, face recognition and gender classification technology represent two facial analysis domains with a great potential for abuse, with the highest risks often falling on already marginalized populations [12, 21, 25].

CelebA attributes: We utilize CelebA attribute annotations to infer directions in latent space that correspond to different factors of variation in the face images. All of the attributes within the dataset are operationalized as binary categories (i.e. the attribute is present or it is absent). We note that in many cases this binary scheme does not reflect the full diversity of real attributes and so the interpretation of category boundaries is contingent on the annotators and annotation scheme. The inherent subjectivity in the attribute annotations means that the factors of variation we manipulate in our experiments are tied to the ways the attributes have been operationalized and annotated within the CelebA dataset.

After careful consideration, we have chosen not to manipulate images based on the `Male` attribute. While this attribute may, on the surface, appear relevant to fairness evaluations, the risk and potential for negative impact outweighs the perceived benefit of examining this attribute in particular. First, the CelebA annotation scheme conceptualizes the `Male` attribute as both binary and perceptually obvious and discernible. Using a gender[1] binary in reporting results implicitly condones the classification of gender into two distinct and opposite categories, which perpetuates harm against individuals who exist outside the bounds of this categorization scheme and reinforces rigid social norms of gender expression. Second, due to the Hollywood skew of the dataset, a very narrow range of gender expression is reflected in the images. Consequently, applying our methodology to the `Male` attribute would result in manipulations that alter a specific set of correlated image features that, through a largely cisnormative and heteronormative cultural lens, are commonly interpreted as gender signifiers. Even with a nuanced discussion of what the `Male` attribute vector is (and is not) encoding, presenting these results risks reinforcing and implicitly condoning the views that individuals of different genders *do* and *should* present a certain way. Generally speaking, we recommend researchers avoid attempts to manipulate images based on unstable social constructs like gender. Instead, we suggest limiting manipulations to attributes defined by (mostly) objective facial characteristics.

[1]We note that it is unclear if the CelebA `Male` annotations are meant to reflect gender or sex (neither of which are binary). Since the annotations are based on face images, we assume the dataset curators were attempting to measure gender.

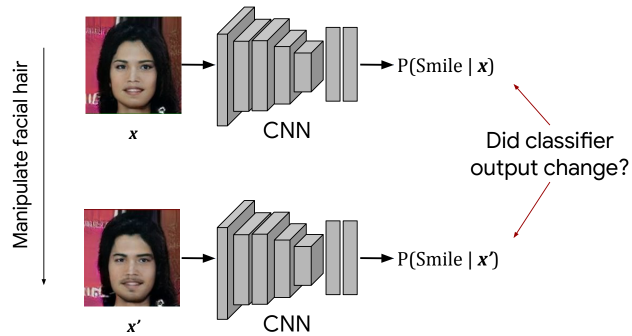


Figure 1: Counterfactual attribute sensitivity measures the effect of manipulating a specific property of an image on the output of a trained classifier. In this example, we consider the effect of adding facial hair on the output of a smiling classifier. If the classifier's output systematically changes as a result of adding or removing facial hair then a potentially undesirable bias has been detected since, all else being equal, facial hair should be irrelevant to the classification task.

3. Related work

Several works have focused on evaluating fairness of human-centric vision classifiers. [1, 20] compare error statistics of facial analysis systems across groups defined by Fitzpatrick skin type [4] and gender. [18] propose standardized model reporting guidelines that include evaluations disaggregated by groups defined along cultural, demographic, and phenotypic lines. Our work is complementary to disaggregated evaluations based on real images, while incorporating recent work in post-hoc interpretability for understanding biases. Specifically, the generative techniques we propose allow new types of evaluation not easily afforded by real image datasets: (i) We can assess how a prediction might change if characteristics of an image *that we believe should be irrelevant to classification* are altered in a specific targeted manner. This means we can identify a causal relationship between features in an image and the classifier's output; (ii) Our approach facilitates testing on faces with combinations of characteristics that may be underrepresented in the testing set.

There has been significant research effort devoted to developing post-hoc approaches for interpreting and explaining neural networks (e.g. see [7] for a review). The most common framework involves computing saliency maps to visualize the importance of individual pixels to a classifier's prediction. Gradient-based saliency approaches [24, 26, 23] attribute importance based on first-order derivatives of the input. Another set of approaches determine importance based on perturbing individual or groups of pixels and observing the change in classifier output [2, 5, 3].

The perturbation methods relate to our approach in that counterfactual-style inputs drive the search for regions that most affect the classifiers output. Our method is different in that, instead of searching over perturbations that maximally influence the classifiers output, we perturb the input along known and meaningful factors of variation to test the classifiers response.

Testing with Concept Activation Vectors (TCAV) [14] is an alternative interpretability framework that poses explanations at the level of high level concepts rather than pixels. TCAV defines a concept via a set of data instances (e.g., images) sharing a common characteristic. Similarly, our method infers attribute vectors based on two sets of images: (i) images with the attribute and (ii) images without the attribute. TCAV uses directional derivatives to reveal the 'conceptual sensitivity' of a model's prediction of a class (e.g `Smiling`) to a concept by the two sets of images). Our approach extends this technique to measure the effect of an attribute vector *directly on the model predictions*. TCAV provides a general framework for model interpretability, and our approach narrows in on cases where generative techniques produce high quality and perceptually realistic images. In the settings where our method is applicable, it provides direct and actionable information about a model's decisions. We believe the two approaches are complementary and can work in combination: TCAV can point to a potential bias, and our method can provide clear interpretable evidence of bias in classifier predictions.

Our method is related to [13, 16], who propose the notion of counterfactual fairness in the context of causal modeling. Here, the fairness of a model is determined by assessing the effect of counterfactual interventions on a sensitive attribute (e.g., race or gender) over a causal graph. Recently, [15, 19] have argued this framework is based on a misunderstanding of how race operates in our society and the flawed assumption that social groups like race and gender are entities that can have counterfactual causality. Crucially, our approach manipulates observable characteristics of a face, rather than conceptualizing the manipulation at the level of social groups. Of course, some of these attributes (e.g., presence of a beard) may be strongly correlated with a particular social group (e.g. gender identities), but they neither constitute nor are determined by the group.

[6] consider counterfactual fairness in text classification. They propose both evaluation and training techniques based on counterfactual manipulation of terms referencing different identify groups in sentences. Our work can be understood as the image analog, with one key difference: rather than manually perturbing the input data, we utilize a generative model to manipulate face characteristics in the images.

Our approach is most similar to [22], which explores how the output of a gender classifier changes in response to the alteration of various characteristics of the same person (e.g., hair length, facial expression, etc.). Our approach differs in that we use generative techniques to systematically vary characteristics of a face image, rather than relying on a manually curated set of real images.

4. Methods

We now outline our *counterfactual face attribute manipulation* framework, which is motivated by questions of the form: how would the prediction change if a single factor of variation in the image were altered?

4.1. Generative model

Generative adversarial networks (GANs) [8] are a framework for learning generative models whereby two networks are trained simultaneously in an adversarial manner. The generator G takes as input a latent code vector $z \in \mathcal{Z}$ sampled from a fixed prior $p(z)$ and outputs an image $x \in \mathcal{X}$. The discriminator D takes as input an image $x \in \mathcal{X}$ that is sampled from either the training set or the generative distribution defined by G. D is trained to discriminate between images from these two distributions while G is trained simultaneously to generate samples that fool D into thinking they come from the data distribution.

In this work, we utilize the architecture and progressive training procedure introduced in [11] and train with the Wasserstein GAN loss [**?**] with a gradient penalty [9]:

$$\begin{aligned}\min_G \max_D & \mathbb{E}_{x\sim p(x)}[D(x)] - \mathbb{E}_{z\sim p(z)}[D(G(z))] \\ & + \lambda \mathbb{E}_{x\sim p_I(x)}[(||\nabla_x D(x)||_2 - 1)^2]\end{aligned} \tag{1}$$

Here, $p(x)$ is the data distribution, $p(z)$ is the prior noise distribution and $p_I(x)$ is given by sampling uniformly along lines between points in the true and generative distributions.

The basic GAN framework lacks an inference mechanism, i.e., a method of identifying the latent code z that generated a particular image x. Inference is central to our method since we use annotated data to estimate directions in latent space that correspond to different factors of variation in the dataset. In order to obtain an approximate inference mechanism we train an encoder E to map from images to latent codes. Specifically, for a fixed generative model, we train the encoder with an ℓ_2 loss to predict the latent code z that generated $x = G(z)$:

$$\min_E \mathbb{E}_{z\sim p(z)} ||z - E(G(z))||_2 \tag{2}$$

4.2. Face attribute vectors

Inspired by [14], we compute attribute vectors by training a linear classifier to separate the latent codes inferred from images annotated with the attribute from latent codes inferred from images annotated without the attribute. We

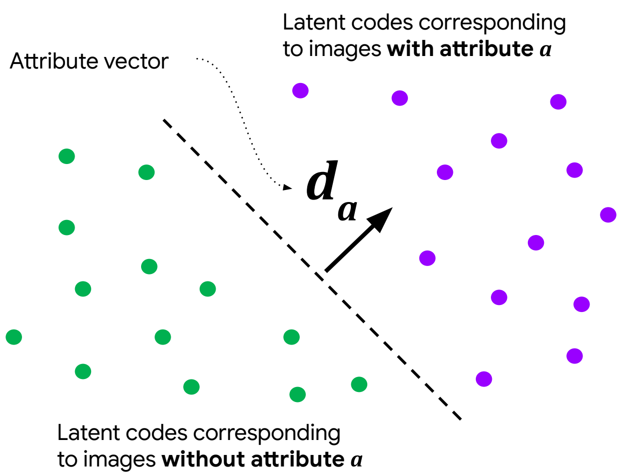


Figure 2: Attribute vectors are computed by training a linear classifier to discriminate between latent codes corresponding to images with and without a given attribute a. The attribute vector d_a is taken to be the vector orthogonal to the decision boundary, normalized to have unit ℓ_2 norm.

take the attribute vector d_a to be the vector orthogonal to the linear classifier's decision boundary. We normalize all attribute vectors to have unit ℓ_2 norm.

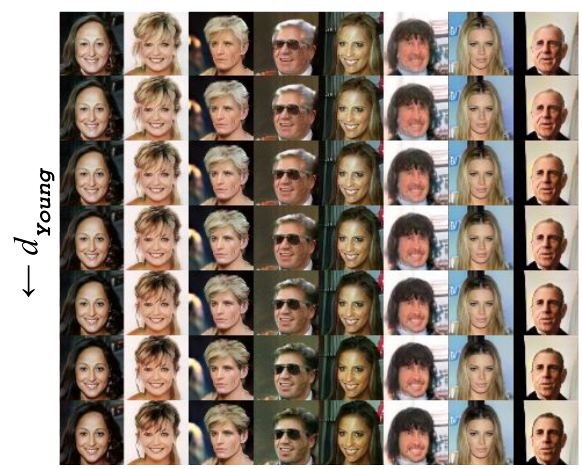
In our formulation, we examine the effect of a range of CelebA attributes (e.g. `Young`, `Heavy_Makeup`, `No_Beard`, etc.) on `Smiling` predictions. Since the `Smiling` attribute is highly correlated with several other attributes in CelebA (see Figure 6 in the Appendix for the full attribute correlation matrix), we ensure the `Smiling` attribute is equally represented in both the positive and negative classes of the other attributes we model to avoid inadvertently encoding it in the attribute vector. We also balance the `Male` attribute in an attempt to disentangle the set of attributes highly correlated with it.

A face attribute vector d_a can be traversed, positively or negatively, to generate images that smoothly vary along the factor of variation defined by the attribute. For example, given an image $x = G(z)$, we can synthesize new images that differ along the factor of variation defined by d_a via $G(z + i * d_a)$. We consider $i \in [-1, 1]$ and note that as i increases above 0, the attribute will be present in increasing amounts. Similarly, as i decreases below 0, the attribute will be removed. Despite being operationalized as binary within the CelebA dataset, the vast majority of characteristics captured by the annotations are continuous. Thus, this continuous traversal of attribute vectors gives rise to perceptually meaningful alterations in the images. For example, Figure 3(a) demonstrates images synthesized by perturbing $z's$ along the $d_{\texttt{Young}}$ direction. We emphasize again that many of the annotations are subjective, and thus, these manipulations reflect how the particular attributes were operationalized and annotated within the CelebA dataset.

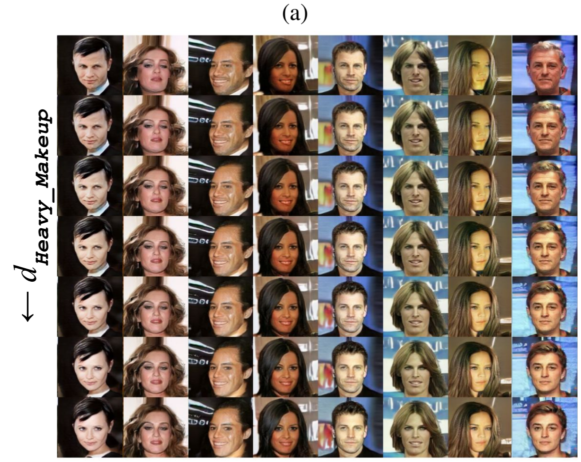
4.3. Measuring sensitivity to counterfactual face attribute manipulation

We now describe our method of testing classifier sensitivity to counterfactual face attribute manipulation. Let f denote a trained classifier that takes an image $x \in \mathcal{X}$ as input and outputs a continuous value representing the probability of a smile being present in the image:

$$f(x) = P(Smile = 1|x) \in [0, 1] \tag{3}$$



(a)



(b)

Figure 3: Traversing in latent code space along the (a) `Young` and (b) `Heavy_Makeup` attribute direction. The center row shows images generated from random $z \sim p(z)$. Rows moving down (up) show images generated by traversing the `Young`/`Heavy_Makeup` attribute vector in a positive (negative) direction. We see age/makeup related facial characteristics being smoothly altered while the overall facial expression remains constant.

Given a threshold, $0 \leq c \leq 1$, binary classifications are obtained:

$$y(x) = \mathbb{I}[P(Smile = 1|x) \geq c] \in \{0, 1\} \tag{4}$$

We are interested in how both $f(x)$ and $y(x)$ are affected as the input x changes in a controlled manner. We measure the sensitivity of the *continuous valued output* of f to image manipulations by attribute vector d as:

$$S_f(d) = \mathbb{E}_{z\sim p(z)}[f(G(z + d)) - f(G(z))] \tag{5}$$

In other words, $S_f(d)$ compares the classifiers output given an image $G(z)$ with the classifiers output given the image manipulated by an attribute vector $G(z + d)$. By comparing images generated from the same z, this metric isolates the effect of altering a specific set of characteristics (defined by attribute vector d) on the classifier's output. $S_f(d)$ will be close to 0 if the classifier is blind to the change resulting from attribute vector d. In contrast, a $S_f(d)$ of large magnitude indicates f is sensitive to attribute d. If d influences something relevant to the classification task (e.g., mouth shape in the context of smiling classification) a non-zero $S_f(d)$ is expected. However, if d encodes characteristics that we believe should be irrelevant to classification (e.g., hair color in the context of smiling classification) then a large $|S_f(d)|$ indicates a potentially undesirable bias in the classifier.

We define a function to indicate if the classifier's binary prediction changes as a result of manipulating an image with attribute vector d:

$$I(z, d) = \mathbb{I}[y(G(z + d))! = y(G(z))] \tag{6}$$

Whereas $S_f(d)$ captures the change in probabilities predicted by the classifier, we now define two metrics that capture the frequency with which manipulating images with attribute vector d changes the binary classification. We break this down into classifications that flip from 'not smiling' to 'smiling':

$$S_y^{0\to 1}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))==0} I(z, d) \tag{7}$$

and classifications that flip from 'smiling' to 'not smiling':

$$S_y^{1\to 0}(d) = \mathbb{E}_{z\sim p(z)|y(G(z))==1} I(z, d) \tag{8}$$

5. Results

CelebA [17] is a dataset of celebrity faces annotated with the presence/absence of 40 face attributes. We train a convolutional neural network to predict the `Smiling` attribute on 128x128 images. We train with a cross-entropy objective and use a threshold of $c = 0.5$ for final predictions. Error statistics, disaggregated by a select set of attributes, are given in Table 2 in the Appendix.

We train a progressive GAN on 128x128 CelebA images with a 128-dimensional zero-mean Gaussian prior. For each attribute in the CelebA dataset, we estimate the directions in latent code space corresponding to these factors of variation via the procedure described in Section 4.2. We infer 800 latent codes for each of the positive (attribute is present) and negative (attribute is not present) classes. We select a random 80% of the vectors for training and evaluate classification accuracy on the remaining 20%. The classification accuracy of each linear classifier is shown in Figure 7 in the Appendix.

To illustrate the visual effect of the generative face attribute manipulation, Figure 3 shows samples generated from the model by traversing the latent code space along the $d_{\texttt{Heavy\_Makeup}}$ and $d_{\texttt{Young}}$ directions respectively. We see age/makeup related facial characteristics being smoothly altered while the overall facial expression remains constant.

Many of the CelebA attributes are highly correlated in the dataset (see Figure 6 in the Appendix for the full correlation matrix). The GAN learns to separate some factors (as evidenced, for example, by the consistent facial expressions in Figure 3), but does not cleanly disentangle all factors of variation in human faces. For example, there are many facial attributes that, through the societal production of the gender binary, have come to signify traditionally masculine or feminine characteristics. Social norms around gender expression lead these features to be highly correlated in society and the CelebA dataset reflects, if not amplifies, these correlations. Empirically, we observe these attributes remain entangled to differing degrees in the generative model. Consequently, traversing one attribute vector (e.g. `Heavy_Makeup`) sometimes changes a larger set of facial characteristics (e.g. removing/adding a beard, altering hair length, etc.) that tend to be correlated with the attribute in question. Future work is aimed at further disentangling factors of variation that are highly correlated in the dataset.

Our approach involves traversing directions in latent code space to generate images that vary smoothly along some factor of variation. We performed a simple sanity check to ensure that images generated from nearby latent codes are consistently classified by the smiling classifier. First, we construct a large set of generated images that are classified as smiling. Then we generate new images by sampling latent codes uniformly from lines connecting two different codes of the generated images. We find that 98% of the new images are also classified as smiling. We similarly test this for images classified as not smiling and find the same results.

Next, we assess the sensitivity of the smiling classifier to the counterfactual face attribute manipulation, using the

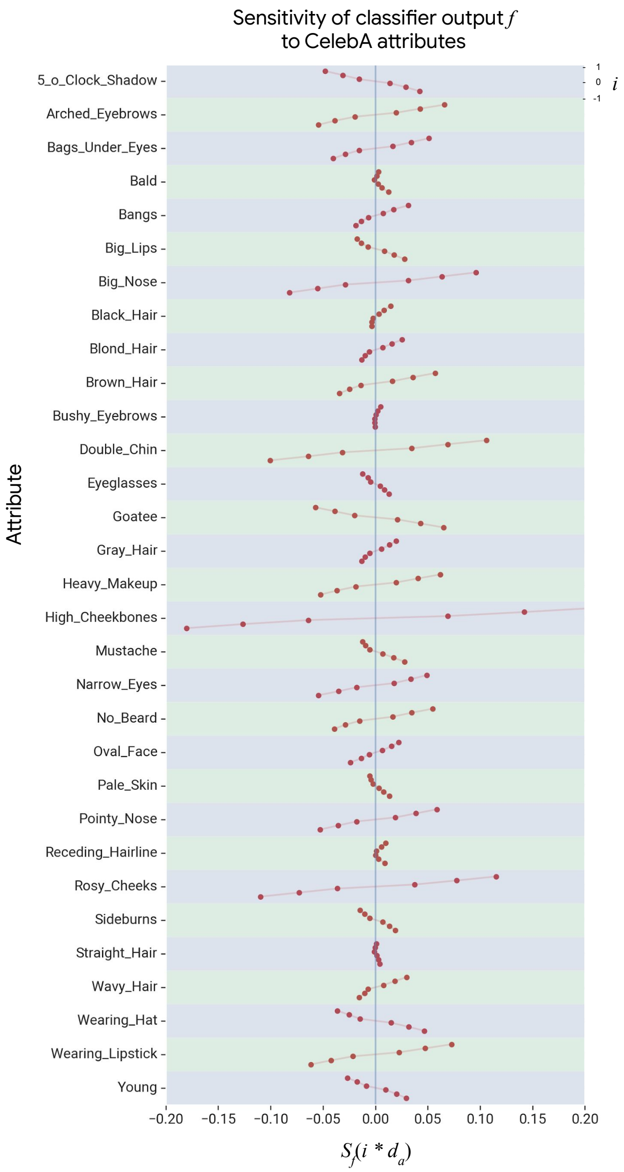


Figure 4: Effect of traversing different attribute vectors on the continuous output of the smiling classifier. For each attribute, the y-axis denotes i equally space within $[-1, 1]$, i.e. the degree to which the attribute has been added ($i > 0$) or removed ($i < 0$) in the image.

measures proposed in Section 4.3. Recall, we can use attribute vector d_a to manipulate attribute a in an image. For example, given image $x = G(z)$, we can lighten hair color by generating

$$G(z + i * d_{\texttt{Blond\_Hair}}), i \in (0, 1].$$

Conversely, generating

$$G(z + i * d_{\texttt{Blond\_Hair}}), i \in [-1, 0)$$

CelebA attribute defining d_a	$S_y^{1\to 0}(d_a)$	$S_y^{0\to 1}(d_a)$
`Young`	**7.0%**	2.6%
`5_o_Clock_Shadow`	**11.8%**	2.2%
`Goatee`	**12.4%**	0.9%
`No_Beard`	0.8%	**11.8%**
`Heavy_Makeup`	1.6%	**12.4%**
`Wearing_Lipstick`	1.7%	**16.3%**

Table 1: Effect of manipulating images with CelebA attribute vectors on smiling classifier's predictions. Vectors for `Young`, `5_o_Clock_Shadow`, and `Goatee` flip the classifier from 'smiling' to 'not smiling', while vectors for `No_Beard`, `Heavy_Makeup`, and `Wearing_Lipstick` flip from 'not smiling' to 'smiling'

will darken hair color in the image. Figure 4 plots the sensitivity $S_f(i * d_a)$ for all attributes a in the CelebA dataset and for $i \in [-1, 1]$. Interestingly, we observe that linear changes to z result in linear changes in the output of the classifier.

Next, we consider the actual change in classification that results from perturbing images in direction d_a. Figure 8 in the Appendix plots these results for a large set of CelebA attributes. Blue bars plot $S_y^{1\to 0}$, i.e., the frequency with which images originally classified as smiling are subsequently classified as not smiling after manipulating images with the attribute vector. Red bars indicate $S_y^{0\to 1}$, i.e., the frequency with which images originally classified as not smiling are subsequently classified as smiling after manipulating images with the attribute vector. The left hand side of the plot indicates the effect of negatively shifting along d_a, i.e. $S_y(-d_a)$ and the right hand side indicates the effect of positively shifting along d_a, i.e. $S_y(d_a)$. Note that whereas Figure 4 plots classifier sensitivity for incremental shifts along d_a, here we plot the effects of a unit step in either the positive or negative direction.

There are several attributes in the CelebA dataset that one might consider from a fairness perspective. We summarize these results in Table 1. For example, the `Young` attribute captures something related to age. Again, we emphasize that the attribute vector $d_{\texttt{Young}}$ reflects how the young attribute was operationalized and measured within the CelebA dataset. Category boundaries of `Young` $= 1$ and `Young` $= 0$ are contingent on the annotators and the distribution of faces that fall into each category in dataset might not accurately represent faces of different ages in the real world. That said, our results indicate the smiling classifier is sensitive to the `Young` direction, with 7% of images changing from a smiling to not smiling classification as a result of manipulating the images in this direction. In contrast, only 2.6% of images flip from a not smiling to smiling classification.

A fair smiling classifier should also perform consistently

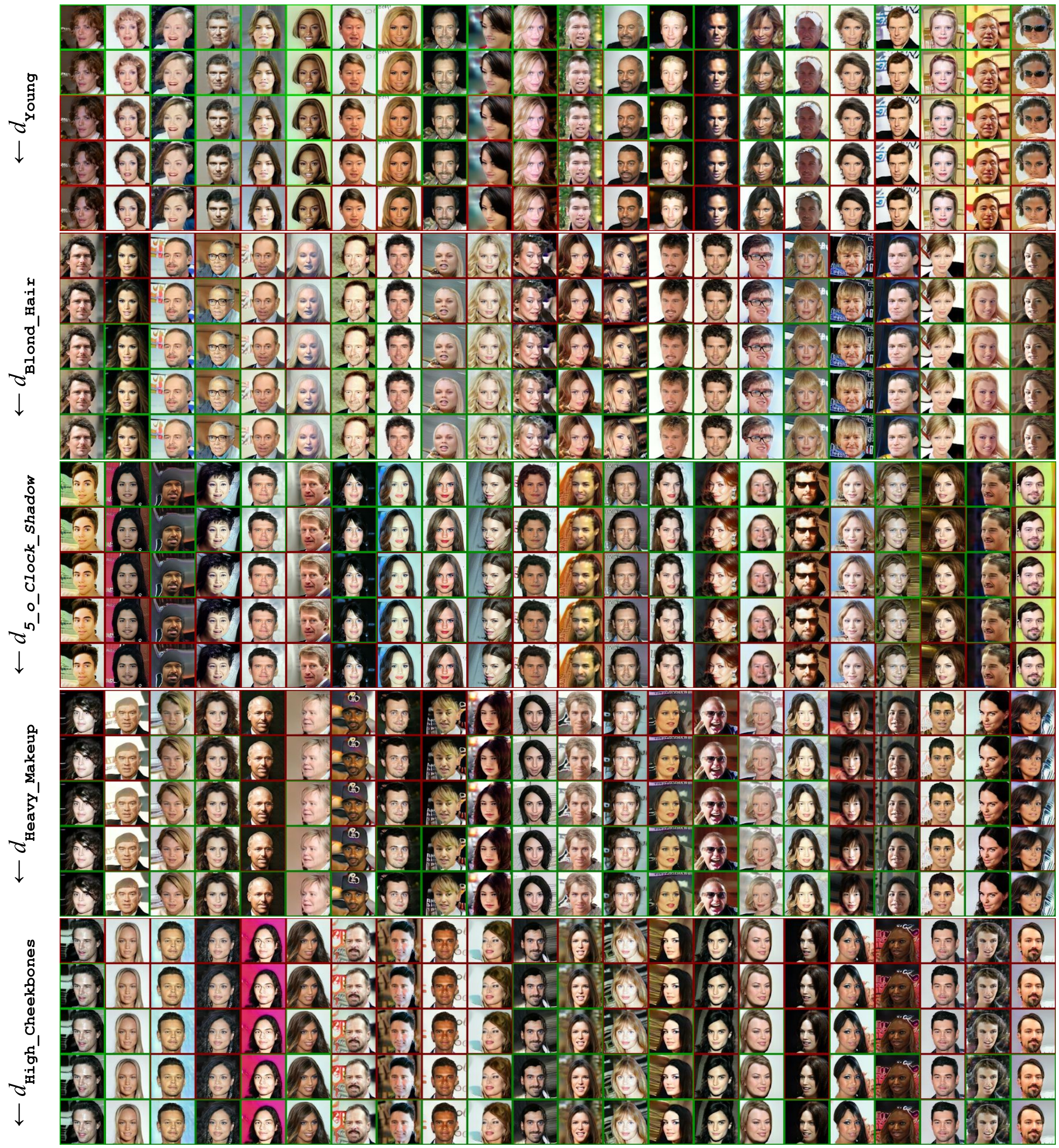


Figure 5: Examples where smiling classifier's prediction flips when traversing different attribute vectors. Green (red) boxes indicate a smiling (not smiling) prediction.

regardless of the gender or gender expression of the individual in the photo. The CelebA `Male` attribute captures something related to these concepts. However, as discussed in Section 2, we chose not to use the `Male` annotations. We emphasize that image manipulations based on the `Male` attribute would not alter anything inherent to gender identity or gender expression. Rather, these manipulations would simply reflect a specific set of correlated features that occur in different proportions in the images annotated `Male` $= 0$ and `Male` $= 1$. Table 1 highlights the effect of more objective attributes, such as facial hair and makeup on smiling classification. We observe that adding facial hair flips classifications from smiling to not smiling. In contrast, adding makeup flips classifications from to smiling to smiling.

Both Figure 4 and Figure 8 suggest the smiling classifier is sensitive to different attributes. We now turn to visual inspection to (i) assess the perceptual effect of traversing the attribute vectors and (ii) ensure the basic mouth/facial expression that signifies the presence/absence of a smile does not change when traversing the attribute vectors. Figure 5 shows several examples of images generated by traversing different attribute vectors. Green or red boxes outline each image to indicate the smiling classifier's prediction for the image. Red boxes indicate a not smiling prediction and green boxes indicate a smiling prediction. We observe the basic facial expression remains constant when traversing many of the attribute vectors (e.g., `Young`, `Blond_Hair`, `5_o_Clock_Shadow`, `Heavy_Makeup`), despite other characteristics of the face changing. This indicates the generative model has sufficiently disentangled the factor of variation defined by the CelebA attribute from the features of a smile. However, despite the faces (perceptually) having nearly identical mouth expressions, the classifier is nonetheless sensitive to the changes.

Visual inspection also identifies some attribute vectors that have encoded something about the mouth shape. For example, we observe that the mouth/facial expression changes significantly when traversing the `High_Cheekbones` attribute vector. Subsequently, manipulating faces according to the `High_Cheekbones` attribute vector does not convey anything meaningful about the classifier's bias.

6. Conclusion

We have proposed a technique that leverages the power of generative models to evaluate and reveal biases in a face attribute classifier. This work represents the first step of a larger research agenda focused on auditing algorithms and post-hoc interpretability and is introduced here as a proof-of-concept.

We now highlight several future research directions:

- Leverage human raters to further understand potential bias embedded in the attribute vectors. This is important to help distinguish the classifier bias from potential bias in the generator / attribute annotations.
- Leveraging the same generative techniques to visualize dataset bias.
- Training the generative model on a larger and more diverse set of faces than is available for the classifier.

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. 1, 2

[2] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *ICLR*, 2019. 2

[3] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NIPS*, 2017. 2

[4] Thomas Bernard Fitzpatrick. The validity and practicality of sun-reactive skin types i through vi. *Archives of dermatology*, 124 6:869–71, 1988. 2

[5] Ruth Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, 2017. 2

[6] Sahaj Garg, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed Chi, and Alex Beutel. Counterfactual fairness in text classification through robustness. In *AIES*, 2019. 3

[7] Leilani Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. *arXiv:1806.00069*, 2018. 2

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 3

[9] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 5769–5779, USA, 2017. Curran Associates Inc. 3

[10] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 793–811, Cham, 2018. Springer International Publishing. 1

[11] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv: 1710.10196*, 10 2017. 3

[12] Os Keyes. The misgendering machines: Trans/hci implications of automatic gender recognition. *PACMHCI*, 2:88:1–88:22, 2018. 2

[13] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Proceedings of the 31st International Conference*

on Neural Information Processing Systems, NIPS'17, pages 656–666, USA, 2017. Curran Associates Inc. 3

[14] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *ICML*, 2018. 3

[15] Issa Kohler-Hausmann. Eddie murphy and the dangers of counterfactual causal thinking about detecting racial discrimination. *Northwestern University Law Review*, 113(5), 2019. 3

[16] Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, 2017. 3

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 1, 5

[18] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *FAT*, 2019. 2

[19] Emanual Moss. Toward a theory of race for fairness in machine learning. *ACM FAT* Tutorial*, 2019. 3

[20] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *AIES*, 2019. 1, 2

[21] Concerned Researchers. On recent research auditing commercial facial analysis technology. *https://medium.com/@bu64dcjrytwitb8/on-recent-research-auditing-commercial-facial-analysis-technology-19148bda1832*, 2019. 2

[22] Kerry Rodden. Is that a boy or a girl? exploring a neural networks construction of gender., 2017. 3

[23] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, Oct 2017. 2

[24] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv:1312.6034*, 2013. 2

[25] Brad Smith. Facial recognition technology: The need for public regulation and corporate responsibility. *https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/*, 2018. 2

[26] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017. 2

[27] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. Predictive inequity in object detection. *arXiv:1902.11097*, 2019. 1

[28] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 01 2017. 1

A. Appendix

Data split	**Accuracy**	**FPR**	**FNR**
Total	89.330%	14.416%	6.929%
`Young=0`	87.170%	17.727%	8.193%
`Young=1`	90.022%	13.391%	6.509%
`Male=0`	89.962%	16.548%	4.984%
`Male=1`	88.356%	11.948%	11.190%
`Heavy_Makeup=0`	88.735%	12.346%	9.769%
`Heavy_Makeup=1`	90.203%	19.046%	4.098%
`Wearing_Lipstick=0`	88.254%	12.071%	11.267%
`Wearing_Lipstick=1`	90.315%	17.521%	4.195%
`No_Beard=0`	88.942%	10.683%	11.699%
`No_Beard=1`	89.396%	15.263%	6.352%
`5_o_Clock_Shadow=0`	89.303%	14.739%	6.823%
`5_o_Clock_Shadow=1`	89.569%	12.014%	8.128%
`Goatee=0`	89.363%	14.667%	6.731%
`Goatee=1`	88.634%	10.500%	13.016%

Table 2: CelebA smiling attribute classification results.

Figure 6 shows the correlation matrix of CelebA attributes, computed from the training images.

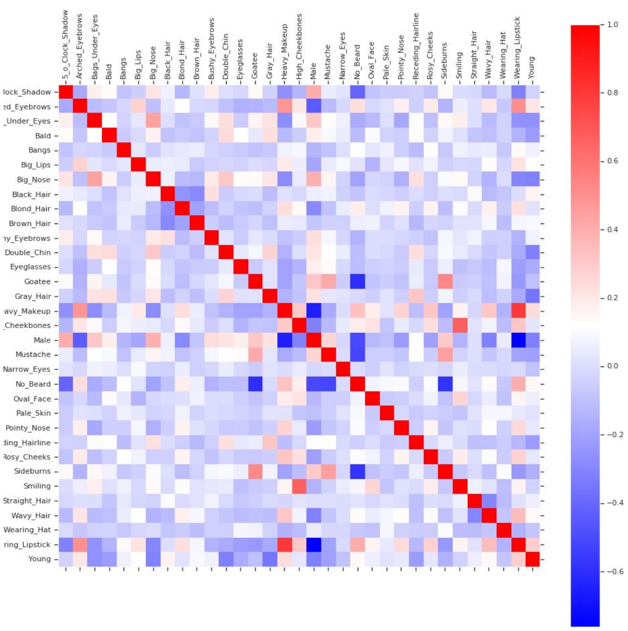


Figure 6: Empirical correlation matrix of CelebA attributes.

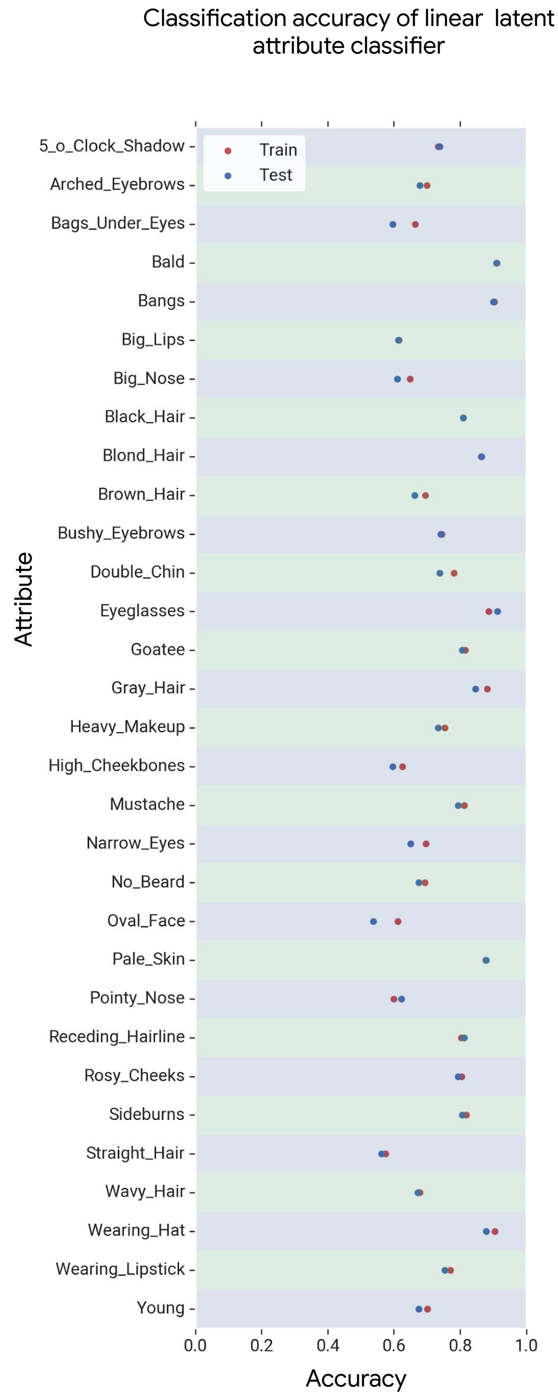


Figure 7: Classification accuracy of each linear classifiers trained to separate latent codes computed from images annotated with and without the attribute. Recall, the vector orthogonal to the decision boundary gives the attribute vector.

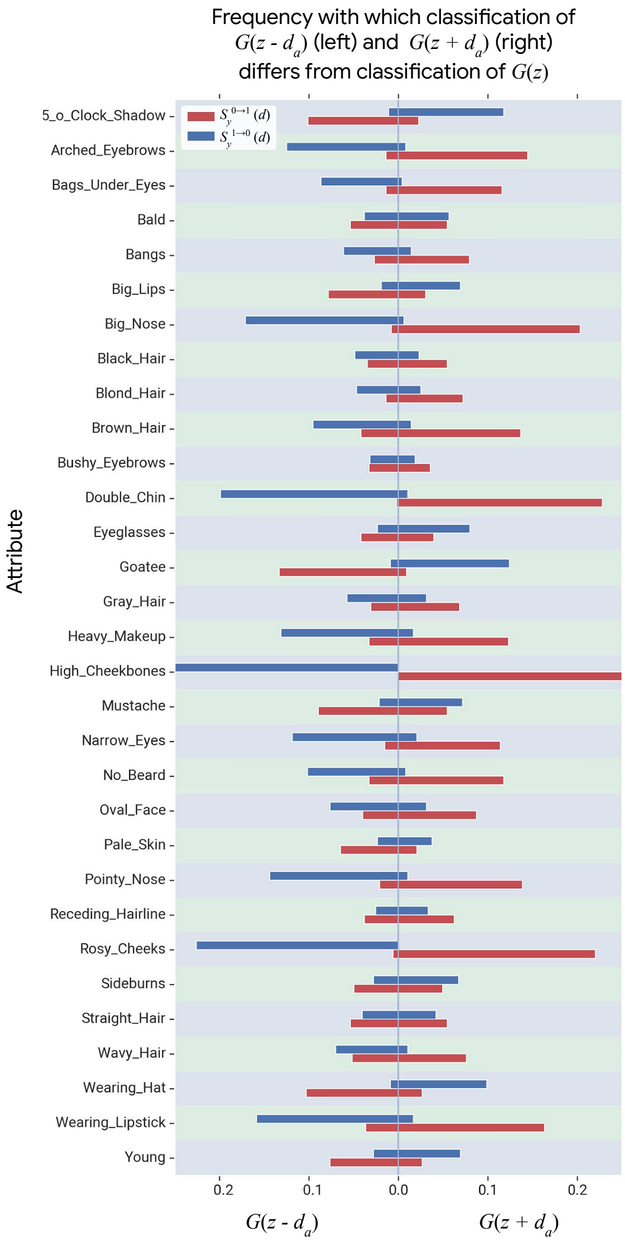


Figure 8: Effect of traversing different attribute vectors on the smiling classifier's binary prediction.