FAIRNESS GAN: GENERATING DATASETS WITH FAIRNESS PROPERTIES USING A GENERATIVE ADVERSARIAL NETWORK

Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney
IBM Research
Yorktown Heights, NY 10598
{psattig@us.,shoffman@,ecvijil@us.,krvarshn@us.}ibm.com

ABSTRACT

We introduce the Fairness GAN, an approach for generating a dataset that is plausibly similar to a given multimedia dataset, but is more fair with respect to protected attributes in decision making. We propose a novel auxiliary classifier GAN that strives for demographic parity or equality of opportunity and show empirical results on several datasets, including the CelebFaces Attributes (CelebA) dataset, the Quick, Draw! dataset, and a dataset of soccer player images and the offenses they were called for. The proposed formulation is well-suited to absorbing unlabeled data; we leverage this to augment the soccer dataset with the much larger CelebA dataset. The methodology tends to improve demographic parity and equality of opportunity while generating plausible images.

1 INTRODUCTION

Automated essay scoring in high-stakes educational assessment (Shermis, 2014; Perelman, 2014) and automated employment screening based on voice and video (Shahani, 2015; Chandler, 2017) are examples of decision making with multimedia inputs supported by machine learning algorithms that raise concern about perpetuating and scaling unwanted bias present in historical data and running afoul of, e.g., Title VI (education) and Title VII (employment) of the Civil Rights Act of 1964 in the United States.

Systematic discrimination against groups defined by protected attributes such as race, gender, caste and religion that directly prevents favorable outcomes such as being hired, paroled, or given a loan has always been a problem in human decision making, but has come to the fore as such decision making is being shifted from people to machines (Williams et al., 2018). This realization has spurred a recent flurry of research in the data mining and machine learning literature. Designing the right goals and arriving at the right problem formulation is the first step towards learning a automatic decision making system (Passi and Barocas, 2019). It is now well-known that there are three main classes of intervention to introduce fairness into supervised machine learning pipelines: pre-processing, in-processing, and post-processing (d'Alessandro et al., 2017).

Much of the fairness literature focuses on structured data. Specifically, existing pre-processing approaches have not been applied to multimedia data and are usually not scalable to the high dimensionality that comes with such data. Moreover, the subset of existing pre-processing methods that generate a new dataset by editing feature values are unlikely to produce realistic pre-processed datasets when applied to multimedia data even if they could scale. (The algorithms of Kamiran and Calders (2012), which do not work with the features, are the only ones of which we are aware that can be tractably applied to high-dimensional data.) To address these limitations, we propose a new debiasing approach based on generative adversarial networks (GANs) (Goodfellow, 2017). We name this methodology the *Fairness GAN*.

Recent contributions examining fairness from the perspective of adversarial learning include Edwards and Storkey (2016), Beutel et al. (2017), Zhang et al. (2018), Madras et al. (2018), Wadsworth et al. (2018), Celis and Keswani (2019), and Adel et al. (2019). These methods are similar to our proposed approach by including a classifier trained to perform as poorly as possible on predicting the outcome from the protected attribute, but are different from our proposed method in two key ways. First, they are not intended to create a releasable new dataset that plausibly approximates a given original biased dataset but with more favorable fairness metrics; they are intended for learning fair latent representations that are not in the same space as the given original, which restricts the transparency of the transformation and the flexibility in the use of the dataset. As such, they are not GANs with a real/fake discriminator component that aims to generate realistic samples. Second, they are applied to low-dimensional structured data or word embeddings.

The method of Edwards and Storkey (2016) provides fairness in the sense of demographic parity; Beutel et al. (2017) allows for fairness in the sense of equality of opportunity; and Zhang et al. (2018) (and its simplification by Wadsworth et al. (2018)) further allows for fairness in the sense of equality of odds. Herein, we propose versions of the Fairness GAN for the first two of these definitions and leave equality of odds for future work.

The existing methods (Edwards and Storkey, 2016; Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018) can handle the case when the protected attribute is known only for a subset of samples and Madras et al. (2018) additionally considers transfer learning. In our work, we consider a similar but distinct case that is relevant because we are attempting to generate plausible signals: the dataset labeled with outcomes and protected attributes is small, but there exists a much larger multimedia dataset with similar features but no outcomes. Here our solution is to augment the dataset for training the generator with samples from the larger dataset.

The GAN literature includes approaches for conditional manipulation of multimedia data (Perarnau et al., 2016), unpaired translation (Zhu et al., 2017), and their combination (Lu et al., 2017), which have some relationship to the method proposed herein. One differentiating aspect of our work is that we do not have a collection of samples from a target distribution that we are trying match. Also, we are never in position to manually manipulate or change the value of some descriptive attribute of the multimedia data, and in fact our work right now is limited to the setting in which all predictive features are multimedia ones: there are no other metadata-like features. Additionally, in the fairness setting, there are objective measures of quality that are defined over a collection of samples and cannot be judged on a single sample alone.

One very recent piece of work by Xu et al. (2018) is also a GAN for fairness developed concurrently with our work. Their architecture is developed for low-dimensional structured data and only applies to demographic parity. While our work is geared towards high dimensional image data and apart from demographic parity our architecture also supports equality of opportunity. Another major difference is the way the demographic parity is imposed. We are striving for independence between the outcome Y and the protected variable C therefore the fairness discriminator is only fed the outcome Y. This helps in retaining information about the protected variable C such as gender in the generated image X and preserving the natural image structure. This crucially enables the use of the generated fair image datasets for downstream tasks.

We experiment with several datasets, including the CelebFaces Attributes (CelebA) dataset (Liu et al., 2015), a dataset of images of soccer players (Silberzahn et al., 2017), and the Quick, Draw! dataset of hand-drawn sketches.[1] Further discussion about datasets is provided in the data description section.

2 Fairness Definitions

Consider the classification task of predicting a true outcome $Y \in \{0, 1\}$ from features X. The prediction is denoted $\hat{Y}$. Consider also a protected attribute $C \in \{0, 1\}$. The fairness notion *demographic parity* is defined as:

$$\Pr[\hat{Y} \mid C = 0] = \Pr[\hat{Y} \mid C = 1], \tag{1}$$

i.e. equality of selection rates across the two groups delineated by the protected attribute. The fairness notion *equality of opportunity* is defined as:

$$\Pr[\hat{Y} \mid C = 0, Y = 1] = \Pr[\hat{Y} \mid C = 1, Y = 1], \tag{2}$$

[1] https://github.com/googlecreativelab/quickdraw-dataset

i.e. equality of false negative rates across the two groups. In order to quantify how close we are to achieving these, we will take the absolute difference of the left-hand side and right-hand side of (1) or (2), respectively.

3 AC-GAN Background

The AC-GAN was recently developed to improve the training of GANs for image synthesis when the images come with a class label C such as 'monarch butterfly,' 'daisy,' and 'grey whale' (Odena et al., 2017). As in all GANs, the AC-GAN has a discriminator $D(\cdot)$ and a generator $G(\cdot)$ that work against each other, and are trained using so-called 'real' samples distributed according to X_{real}. Every generated sample is a function of a noise realization z and additionally a class label realization c, i.e. $X_{\text{fake}} = G(c, z)$.[2] The objective functions for training the generator and discriminator are composed of the typical GAN objective, the log-likelihood of the correct source $S \in \{\text{real}, \text{fake}\}$:

$$L_S = E[\log P(S = \text{real} \mid X_{\text{real}})] + E[\log P(S = \text{fake} \mid X_{\text{fake}})] \tag{3}$$

as well as the log-likelihood of the correct class:

$$L_C = E[\log P(C = c \mid X_{\text{real}})] + E[\log P(C = c \mid X_{\text{fake}})]. \tag{4}$$

The discriminator maximizes $L_S + L_C$ and the generator minimizes $L_S - L_C$.

4 Fairness GAN Formulation

The proposed Fairness GAN builds upon the AC-GAN. Notational differences are as follows. First, instead of using C to indicate a class label, we use C to denote the protected attribute label such as gender or caste. Second, we have an additional outcome variable Y which is the decision such as hiring or loan approval.

The objective of the Fairness GAN is to take a given real dataset $(C_{\text{real}}, X_{\text{real}}, Y_{\text{real}})$ and learn to generate debiased data $(X_{\text{fake}}, Y_{\text{fake}})$ such that the joint distribution of the features and outcome of the generated data (conditioned on the protected attribute) is close to that of the real data while yielding decisions that have either demographic parity or equality of opportunity. Ideally, the outcome produced by the data generator would be independent of the conditioning protected attribute; we pursue this ideal by reversing the motive of the AC-GAN and introducing an auxiliary classifier trained to predict outcome from protected attribute as poorly as possible.

The generator of the Fairness GAN produces both the features and outcome variables: $(X_{\text{fake}}, Y_{\text{fake}}) = G(c, z)$. It contains two variations of the log-likelihoods of the correct source, one pair for joint (X, Y) samples with source variable S_J:

$$L_{S_J}^R = E[\log P(S_J = \text{real} \mid X_{\text{real}}, Y_{\text{real}})] \tag{5}$$

$$L_{S_J}^F = E[\log P(S_J = \text{fake} \mid X_{\text{fake}}, Y_{\text{fake}})], \tag{6}$$

and one pair for multimedia X features alone with source variable S_X:

$$L_{S_X}^R = E[\log P(S_X = \text{real} \mid X_{\text{real}})] \tag{7}$$

$$L_{S_X}^F = E[\log P(S_X = \text{fake} \mid X_{\text{fake}})]. \tag{8}$$

We include a pair of class-conditioned losses to add structure to the GAN and help with training and generating plausible images. These objective functions are the same as in the AC-GAN:

$$L_C^R = E[\log P(C = c \mid X_{\text{real}})] \tag{9}$$

$$L_C^F = E[\log P(C = c \mid X_{\text{fake}})]. \tag{10}$$

Finally and most importantly, for fairness, we include a pair of losses to encourage demographic parity:

$$L_{DP}^R = E[\log P(C = c \mid Y_{\text{real}})] \tag{11}$$

$$L_{DP}^F = E[\log P(C = c \mid Y_{\text{fake}})]. \tag{12}$$

[2] We have overloaded the variable C because it will take the role of the protected attribute in the proposed Fairness GAN.

The argument supporting these terms is the same as given by Zhang et al. (2018).

The discriminator maximizes:

$$L^D = L^R_{S_J} + L^F_{S_J} + L^R_{S_X} + L^F_{S_X} + L^R_C + L^R_{DP},$$

and the generator minimizes:

$$L^G_{DP} = L^F_{S_J} + L^F_{S_X} - L^F_C + L^F_{DP}.$$

To summarize, the discriminator cost term (L^D) that is being maximized contains the standard AC-GAN terms for the discriminator plus the fairness term. The first two terms ($L^R_{S_J} + L^F_{S_J}$) are the expected probability of the discriminator assigning the correct source label (real or fake) to the joint samples (X,Y). The next two terms ($L^R_{S_X} + L^F_{S_X}$) are the expected probability of the discriminator assigning the correct source label using only X. We found that having these terms separately on X helped image quality in general. The fifth term (L^R_C) is typical in AC-GAN; it ensures that the discriminator can determine the conditioning variable C from X. The last term (L^R_{DP}) is to train the discriminator to predict the conditioning variable from the outcome Y alone for real data.

The generator cost term (L^G_{DP}), that is being minimized, consists of the standard AC-GAN terms for the generator plus the fairness term. The first term ($L^F_{S_J}$) and second term ($L^F_{S_X}$) leads to the misclassification of the fake joint (X,Y) and X samples by the discriminator, respectively. The third term (L^F_C) is typical in AC-GAN; it ensures that the discriminator can determine the conditioning variable C from fake X (generated images). The last term (L^F_{DP}) is aimed at reducing the ability of the discriminator to correctly predict the conditioning variable from the outcome Y alone for generated dataset. At convergence, the generator produces (X, Y) pairs where the Y alone cannot lead to good prediction of the protected variable C.

For equality of opportunity, we take a cue from Beutel et al. (2017) and Zhang et al. (2018), and activate the fairness loss on the generator only for samples where $Y_{fake} = 1$. In the binary case, where $Y_{\text{fake}} \in [0, 1]$, the generator minimizes:

$$L^G_{EOPP} = L^F_{S_J} + L^F_{S_X} - L^F_C + Y_{\text{fake}} L^F_{DP}.$$

5 Data Description

CelebA CelebA consists of 202,599 color images of the faces of celebrities downloaded from the internet, cropped and resized to 64 pixels by 64 pixels (Liu et al., 2015). The images come with 40 binary attributes annotated by a “professional labeling company,” further described by Böhlen et al. (2017) as “a group of 50 paid male and female participants, aged 20 to 30, and recruited from mainland China during a 3 month development phase.”

One of the 40 attributes is *male* and another is *attractive*. In one set of experiments, we use *male* as the protected attribute $C \in \{0, 1\}$ without further delving into the social construction of gender or commenting on why the attribute is named as it is. In another set of experiments, we use skin tone as the protected attribute. CelebA contains multiple images of the same celebrity; we manually annotated one image for each of the 10,177 unique celebrities and propagated the annotation to the rest of the images. We used the Fitzpatrick skin type scale to do the annotation, with types I (ivory), II (beige), and III (light brown) categorized as $C = 0$ and types IV (medium brown), V (dark brown), and VI (very dark brown) categorized as $C = 1$.

We use *attractive* as the outcome decision $Y \in \{0, 1\}$ and assume it represents the labeler’s judgment on the celebrity’s attractiveness. Several concerns with this attribute are presented by Böhlen et al. (2017), but we treat it as a decision just like a hiring decision or a decision to accept an individual into a program.

Soccer (Many Analysts, One Dataset) We also consider a second dataset of images of people: soccer players from European professional leagues along with a record of their cautionable and sending-off offenses. This dataset was originally assembled for a unique crowdsourcing experiment testing whether different statisticians will find evidence of racial discrimination in the calling of offenses by referees (Silberzahn et al., 2017).

The players have been labeled according to skin tone by two annotators, Lisa and Shareef, with five possible values {0 (very light skin), 0.25 (light skin), 0.5 (neither light nor dark skin), 0.75 (dark skin), 1 (very dark skin)}. We count players with average annotation value less than 0.5 as light and players with average greater than or equal to 0.5 as dark, and use this as the protected attribute C.

The dataset contains counts of yellow cards, second yellow cards, and straight red cards given to the player. We aggregate the count of all of these offenses over all matches and set $Y = 0$ for players with more than 0.12 offenses per match and set $Y = 1$ for players with less than or equal to 0.12 offenses per match. The value 0.12 is fairly arbitrary, but a choice that does not lead to severe class imbalance.

The images are a mix of action shots and posed profile pictures from which we extract faces using a pre-trained Viola-Jones face detector and scale them to 64 pixels by 64 pixels. The face detector does not find any faces in 5.3% of the images, without any discernible bias related to C or Y. We drop these samples, yielding 1501 total samples in the dataset. In images that the detector finds more than one face, we manually select the main player. We note that the profile image of the player is not the direct basis of individual yellow cards and red cards given by the referee.

Quick, Draw! The final dataset we consider is not images of people, but sketches drawn by people. The Google Quick, Draw! dataset contains 50 million quickly drawn sketches of objects from 345 categories.

The sketches are captured as vectors of pen movements and also released as 28 pixel by 28 pixel bitmap images. Unfortunately, it was not clear whether the correspondence between the image and the metadata was maintained in the released bitmap images. So, we created our own 64 pixel by 64 pixel single-channel bitmap images from the vector representation. Each sample is labeled by the country of the user that submitted the sketch. We use this variable as the protected attribute C. We use the binary assessment of the quality of the sketch as the decision Y (essentially a performance evaluation that could potentially be used for evaluating candidates for employment or admission to educational institutions). This quality assessment, the variable named ‘recognized,’ seems to be the output of an automated decision making model, but that is irrelevant to our work and could just as easily have been a human judgment.

We focus on the category ‘power outlet’ as it is known to have differential recognition performance on submitters from Great Britain (poor), and Canada and the United States (good) (TailSpectrum, 2016). We use a binary C with Great Britain and Canada since the two countries have a similar number of samples for power outlet sketches.

6 Technical Description

The architecture of the generator and discriminator are based on the network structures in Miyato et al. (2018), Miyato and Koyama (2018), and Gulrajani et al. (2017). The image generation path of the generator transforms the noise vector into image X using a linear layer followed by 4 up-sampling ResNet blocks (He et al., 2016). To generate class-conditional images, conditional batch normalization is employed (Dumoulin et al., 2017). The same noise vector is passed through 2 dense layers to generate the outcome variable Y. ReLU functions are used as the activation functions in the intermediate layers for both the image and outcome generation paths. Meanwhile, $\tanh$ activation is used as the last layer for both paths.

The discriminator provides four outputs: probability distribution over sources of the joint samples, probability distribution over sources of the image samples, probability distribution over classes conditioned on image samples, and probability distribution over classes conditioned on outcome samples: $P(S_J \mid X, Y)$, $P(S_X \mid X)$, $P(C \mid X)$, $P(C \mid Y) = D(X, Y)$. The first three outputs share a common network $\phi(.)$ that transforms X. $P(S_X \mid X)$ and $P(C \mid X)$ are obtained using independent linear layers over $\phi(X)$. $P(C \mid Y)$ is obtained by passing Y through 2 dense layers with ReLU activation function.

We use an architecture similar to projector discriminator (Miyato and Koyama, 2018) to obtain $P(S_J \mid X, Y)$. First, the outcome variable Y is embedded in a space with same dimensionality as $\phi(X)$. This is followed by combining the embeddings of X and Y by inner product based interaction $(f(X, Y) = Y\mathbf{V}_Y\phi(X) + \mathbf{V}_X\phi(X))$ instead of the common approach of concatenation.

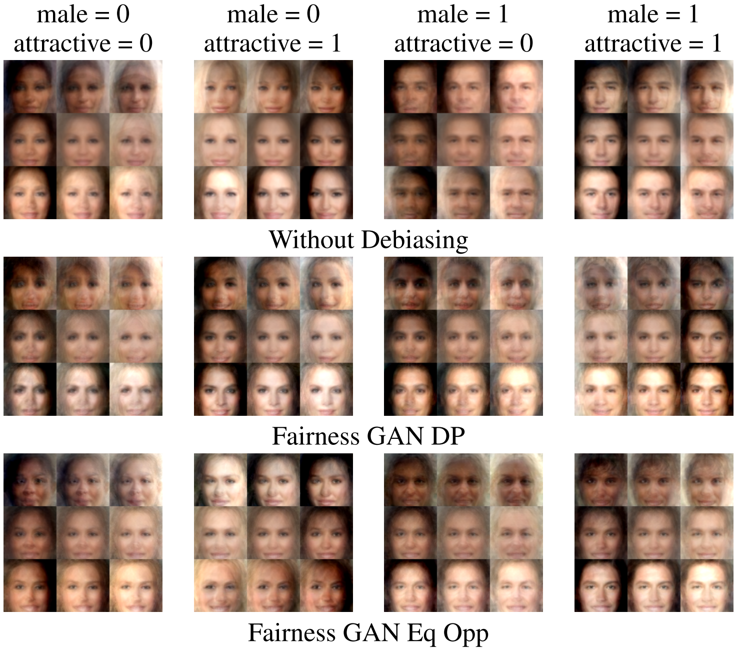


Figure 1: Eigenfaces from the CelebA dataset (male, attractive).

In the case of the small cardinality soccer dataset, we leverage the much larger CelebA dataset to aid the learning of the generator. Skin tone is the protected attribute C for both datasets, but CelebA is unlabeled with respect to the yellow card and red card outcome Y. To overcome the lack of outcome labels in CelebA, we tweak the discriminator loss to:

$$L^D = L_{S_J}^{R^{lab}} + L_{S_J}^{F} + L_{S_X}^{R^{lab+unl}} + L_{S_X}^{F} + L_{C}^{R^{lab+unl}} + L_{DP}^{R^{lab}},$$

where lab denotes the *labeled* soccer dataset and unl denotes the *unlabeled* CelebA dataset.

We treat Y as a continuous variable to get around the issues of back-propagating through discrete variables. We also soften these values to range between $(-0.8, 0.8)$ and add stochasticity in the form of Gaussian noise with standard deviation 0.01 for training stability. An alternative is to keep Y discrete via the Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017), but given that it is just a single scalar variable alongside a high-dimensional X, it is simpler to treat Y as continuous.

The classifiers for evaluation have the same architecture as the discriminator output $P(C \mid X)$. For stable GAN training, the discriminator weights are regularized using spectral normalization (Miyato et al., 2018). Cross entropy loss is used for classification losses while hinge version of the adversarial loss is used for source (real or fake) losses (Lim and Ye, 2017). We use Adam optimizer with hyperparameters set to $\beta_1 = 0.0$ and $\beta_2 = 0.9$. We also decay the learning linearly with the initial value set to 2×10^{-4}.

Because of the small number of real samples in the soccer dataset, it is impractical to train a classifier with a similar architecture to the GAN discriminator like we do with the other datasets. Therefore, we chose to only train the weights for the final fully-connected layer and use the weights from a classifier trained on CelebA faces for the previous layers.

We apply the Fairness GAN to the four datasets described above. We perform a random 90/10 training/testing partition of the data (70/30 partition for soccer) and use the training set for independently learning two GANs: one with demographic parity loss and one with equality of opportunity loss. Debiased datasets are generated from the learned generators. Then three separate classifiers are trained on the training partition of the original data without debiasing and the two generated datasets with debiasing. Finally, these three separate classifiers are evaluated on the same testing partition of the original data. The reported performance is based on an average across iterations of the classifier.

7 Empirical Results

Table 1 presents the different error rates for the three classifiers conditioned on the protected attribute, the overall unconditional error rate, and the values of demographic parity loss and equality of opportunity loss for all four datasets.

Table 1: Error rates, demographic parity, and equality of opportunity.

CelebA (male, attractive)	Without Debiasing		Fairness GAN DP		Fairness GAN Eq Opp		Reweighing	
	male = 0	male = 1	male = 0	male = 1	male = 0	male = 1	male = 0	male = 1
False Positive Rate	0.4043	0.0927	0.5185	0.2356	0.4232	0.1672	0.1445	0.1152
False Negative Rate	0.1213	0.4222	0.1821	0.4074	0.2119	0.4373	0.3659	0.3535
Error Rate	0.2196	0.1749	0.2989	0.2785	0.2853	0.2346	0.2890	0.1746
Unconditional Error Rate	0.2023		0.2910		0.2657		0.2447	
Demographic Parity	0.0447		**0.0204**		0.0507		0.1144	
Equality of Opportunity	0.3009		0.2253		0.2253		**0.0123**	

CelebA (skin tone, attractive)	Without Debiasing		Fairness GAN DP		Fairness GAN Eq Opp		Reweighing	
	dark	light	dark	light	dark	light	dark	light
False Positive Rate	0.1186	0.2035	0.3296	0.4761	0.2896	0.3755	0.1126	0.1507
False Negative Rate	0.3099	0.1917	0.3279	0.2194	0.3652	0.2799	0.3495	0.2764
Error Rate	0.1846	0.1973	0.3290	0.3413	0.3157	0.3253	0.1943	0.2167
Unconditional Error Rate	0.1953		0.3394		0.3238		0.2132	
Demographic Parity	0.0127		0.0123		**0.0096**		0.0224	
Equality of Opportunity	0.1182		0.1085		0.0853		**0.0730**	

Soccer	Without Debiasing		Fairness GAN DP		Fairness GAN Eq Opp		Reweighing	
	dark	light	dark	light	dark	light	dark	light
False Positive Rate	0.2029	0.1428	0.1591	0.3466	0.3445	0.4814	0.2035	0.1412
False Negative Rate	0.8202	0.8089	0.8850	0.6899	0.5492	0.5651	0.7700	0.8215
Error Rate	0.5459	0.4387	0.5624	0.4991	0.4582	0.5186	0.5182	0.4434
Unconditional Error Rate	0.4602		0.5118		0.5064		0.4584	
Demographic Parity	0.1072		0.0633		**0.0604**		0.0749	
Equality of Opportunity	**0.0113**		0.1950		0.0160		0.0515	

Quick, Draw!	Without Debiasing		Fairness GAN DP		Fairness GAN Eq Opp		Reweighing	
	GB	CA	GB	CA	GB	CA	GB	CA
False Positive Rate	0.2638	0.3697	0.2951	0.3482	0.4213	0.4921	0.2844	0.3109
False Negative Rate	0.0716	0.0189	0.1957	0.1343	0.0348	0.0111	0.0645	0.0548
Error Rate	0.1096	0.0509	0.2203	0.1565	0.1113	0.0549	0.1132	0.0850
Unconditional Error Rate	0.0864		0.1938		0.0890		0.1008	
Demographic Parity	0.0587		0.0638		0.0563		**0.0281**	
Equality of Opportunity	0.0527		0.0614		0.0237		**0.0096**	

We visualize the image samples in Figure 1 as follows. We compute the eigenfaces or eigensketches and show the mean image in the center of a 3 by 3 grid (Turk and Pentland, 1991). The images to the left and right show variation along the first principal component. The images to the top and bottom show variation along the second principal component. The diagonal images show variation along both the first and second principal components together. In Table 1, we additionally provide a comparison to the Reweighing method of Kamiran and Calders (2012), the only existing preprocessing method we are aware of that could tractably be applied to multimedia data (since it does not work with the features).

On CelebA (male, attractive), we see that debiasing takes the demographic parity value from 0.0447 to 0.0204 and the equality of opportunity from 0.3009 to 0.2253. Reweighing actually makes demographic parity worse, but does an excellent job at equality of opportunity. The overall accuracy does suffer in achieving these goals. With GAN debiasing, the false positive rate for non-males is quite high, which actually results in more favorable outcomes for this group. (Different definitions of fairness act in different, perhaps unexpected, ways (Friedler et al., 2018)). Looking at the eigenfaces, we see that the demographic parity GAN makes both unattractive and attractive males presumably less attractive by feminizing their features: fuller lips, less defined jawline, and bigger eyes. The change caused by the equality of opportunity GAN is much less pronounced on the unattractive males, which makes sense because equality of opportunity is only concerned with $Y = 1$. One of the desired properties of the Fairness GAN approach is to produce a realistic dataset that can be examined in a transparent way; our examination of lips and jawlines is exactly such a transparent examination not possible with latent representation based approaches.

CelebA (skin tone, attractive) already has excellent demographic parity, and the GAN improves it a bit. The GAN improves the equality of opportunity from 0.1182 to 0.0853. Here too, Reweighing makes demographic parity worse and has the best equality of opportunity performance. The eigenfaces (see Figure 2 in Appendix) show that the GANs equalize the skin tones across the groups. In fact, after debiasing for equality of opportunity, it appears that the dark attractive mean face has slightly lighter tone than the light attractive mean face.

The soccer dataset presents an inherently difficult task because a face image is not a particularly useful feature to predict cautionable and sending-off offenses. Nonetheless, even for such a challenging dataset, the Fairness GAN does improve demographic parity quite a lot. The equality of opportunity is already extremely small, and the equality of opportunity GAN maintains a small value; the demographic parity GAN, which does not have equality of opportunity in its objective causes it to degrade significantly—once again illustrating that the different fairness definitions behave in weird ways sometimes. Reweighing is poorest on this dataset. It is observed that the GANs serve to equalize skin tone in these eigenfaces (see Figure 3 in Appendix) as well.

Finally, with the Quick, Draw! dataset, we notice that the GAN is not able to improve the demographic parity, but is able to improve the equality of opportunity from 0.0527 to 0.0237. We suspect that too much sample distortion would be required to improve the demographic parity. The eigensketches (see Figure 4 in Appendix) are very interesting to examine. The unrecognized sketches from both Great Britain and Canada have no structure to them but the recognized sketches do: more single socket square three-pronged patterns from Great Britain and double socket flat blade patterns from Canada (corresponding to those countries' respective power outlets). After debiasing, the Great Britain eigensketches have more of the double socket character and the Canada ones have more of the single socket square character. It is likely because of this mixture characteristic that Reweighing is superior on this dataset.

8 Conclusion

In this paper, we have examined fairness in the scenario of binary classification with multimedia features and have developed a GAN-based pre-processing approach to improve demographic parity or equality of opportunity by learning to generate a fairer dataset in the original input feature space. We use the proposed algorithm, the first application of GANs to algorithmic fairness, to process several attributed image datasets with varied properties, outcome variables, and protected attributes by adapting a unique combination of several recent techniques from the GAN literature. The empirical results are generally positive (superior to Reweighing on face datasets for demographic parity), but do leave room for improvement.

The work so far illuminates several directions for future research, e.g., considering other modalities of multimedia data in addition to images, adding the equality of odds fairness definition (Zhang et al., 2018), pursuing the Gumbel-Softmax trick for discrete outcome variables (Jang et al., 2017; Maddison et al., 2017), and formulating a Fairness GAN for continuous protected attributes (Louppe et al., 2017).

References

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *Thirty-Third AAAI Conference on Artificial Intelligence*, 2019.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Proc. Workshop Fairness, Accountability, Transparency Mach. Learn.*, Halifax, Canada, August 2017.

Marc Böhlen, Varun Chandola, and Amol Salunkhe. Server, server in the cloud. who is the fairest in the crowd? arXiv:1711.08801, November 2017.

L Elisa Celis and Vijay Keswani. Improved adversarial learning for fair classification. *arXiv preprint arXiv:1901.10443*, 2019.

Simon Chandler. The AI chatbot will hire you now. https://www.wired.com/story/the-ai-chatbot-will-hire-you-now/, September 2017.

Brian d'Alessandro, Cath O'Neil, and Tom LaGatta. Conscientious classification: A data scientist's guide to discrimination-aware classification. *Big Data*, 5(2):120–134, June 2017.

Vincent Dumoulin, Jonathon Shlens, and Manjunath Kudlur. A learned representation for artistic style. In *Proc. Int. Conf. Learn. Repr.*, Toulon, France, April 2017.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *Proc. Int. Conf. Learn. Repr.*, San Juan, Puerto Rico, May 2016.

Sorelle A. Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P. Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. arXiv:1802.04422, February 2018.

Ian Goodfellow. NIPS 2016 tutorial: Generative adversarial networks. arXiv:1701.00160, April 2017.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Adv. Neur. Inf. Process. Syst.*, pages 5769–5779, Long Beach, USA, December 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 770–778, Las Vegas, USA, June 2016.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *Proc. Int. Conf. Learn. Repr.*, Toulon, France, April 2017.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, October 2012.

Jae Hyun Lim and Jong Chul Ye. Geometric GAN. arXiv:1705.02894, May 2017.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3730–3738, Santiago, Chile, December 2015.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. In *Adv. Neur. Inf. Process. Syst.*, pages 981–990, Long Beach, USA, December 2017.

Yongyi Lu, Yu-Wing Tai, and Chi-Keung Tang. Conditional cycleGAN for attribute guided face image generation. arXiv:1705.09966, May 2017.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *Pro. Int. Conf. Learn. Repr.*, Toulon, France, April 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. arXiv:1802.06309, February 2018.

Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *Proc. Int. Conf. Learn. Repr.*, Vancouver, Canada, April–May 2018.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *Proc. Int. Conf. Learn. Repr.*, Vancouver, Canada, April–May 2018.

Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proc. Int. Conf. Mach. Learn.*, pages 2642–2651, Sydney, Australia, August 2017.

Samir Passi and Solon Barocas. Problem formulation and fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48. ACM, 2019.

Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M. Álvarez. Invertible conditional GANs for image editing. In *Proc. NIPS Workshop Adversarial Training*, Barcelona, Spain, December 2016.

Les Perelman. When 'the state of the art' is counting words. *Assessing Writing*, 21:104–111, July 2014.

Aarti Shahani. Now algorithms are deciding whom to hire, based on voice. https://www.npr.org/sections/alltechconsidered/2015/03/ 23/394827451/now-algorithms-are-deciding-whom-to-hire-based-on-voice, March 2015.

Mark D. Shermis. State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, 20:53–76, April 2014.

Raphael Silberzahn et al. Many analysts, one dataset: Making transparent how variations in analytical choices affect results. PsyArXiv:qkwst, September 2017.

TailSpectrum. Quickdraw with Google didn't recognise my drawing of a power outlet because I drew a 3-pinned rectangular plug. https://www.reddit.com/r/britishproblems/comments/5deygp/quickdraw_with_google_didnt_recognise_my_drawing, November 2016.

Matthew Turk and Alex Pentland. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3(1):71–86, Winter 1991.

Christina Wadsworth, Francesca Vera, and Chris Piech. Achieving fairness through adversarial learning: an application to recidivism prediction. In *Proc. Workshop Fairness, Accountability, Transparency Mach. Learn.*, Stockholm, Sweden, July 2018.

Betsy Anne Williams, Catherine F. Brooks, and Yotam Shmargad. How algorithms discriminate based on data they lack: Challenges, solutions, and policy implications. *J. Inf. Policy*, 8:78–115, 2018.

D. Xu, S. Yuan, L. Zhang, and X. Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, Dec 2018. doi: 10.1109/BigData.2018.8622525.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proc. AAAI/ACM Conf. Artif. Intell., Ethics, Society*, New Orleans, USA, February 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 2242–2251, Venice, Italy, October 2017.

A Appendix

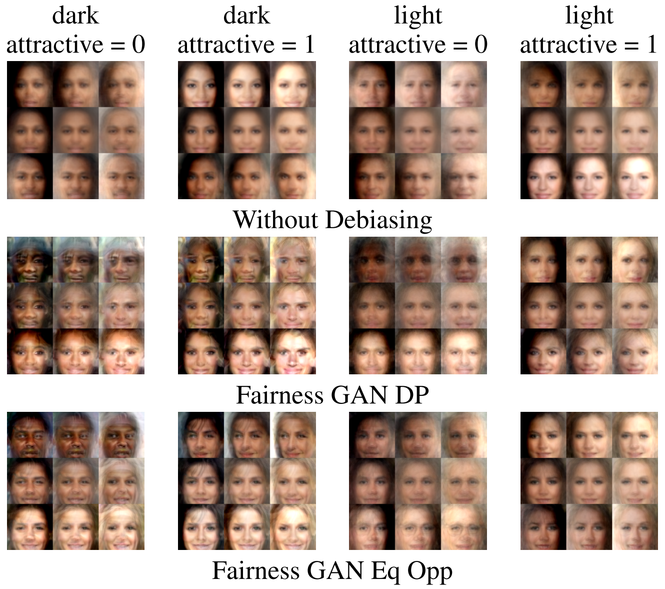


Figure 2: Eigenfaces from the CelebA dataset (skin tone, attractive).

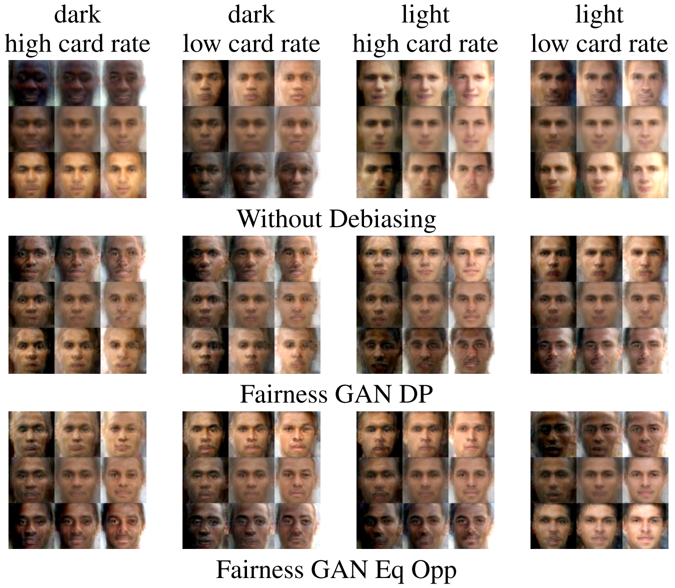
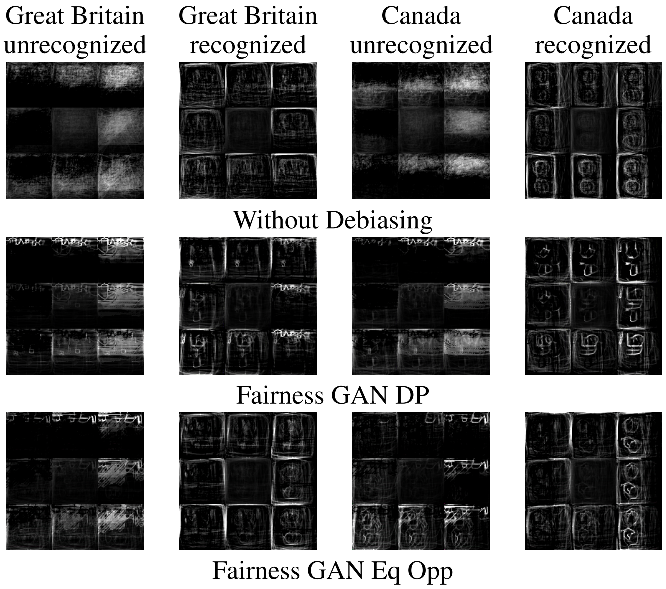


Figure 3: Eigenfaces from the soccer dataset.



Without Debiasing

Fairness GAN DP

Fairness GAN Eq Opp

Figure 4: Eigensketches from the Quick, Draw! dataset's power outlet category.