Benchmarking Bias Mitigation Algorithms in Representation Learning through Fairness Metrics

Anonymous authors
Paper under double-blind review

Abstract

With the recent expanding attention of machine learning researchers and practitioners to fairness, there is a void of a common framework to analyze and compare the capabilities of proposed models in deep representation learning. In this paper, we evaluate different fairness methods trained with deep neural networks on a common synthetic dataset to obtain a better insight into the working of these methods. In particular, we train about 2000 different models in various setups, including unbalanced and correlated data configurations, to verify the limits of the current models and better understand in which setups they are subject to failure. In doing so we present a dataset, a large subset of proposed fairness metrics in the literature, and rigorously evaluate recent promising debiasing algorithms in a common framework hoping the research community would take this benchmark as a common entry point for fair deep learning.

1 Introduction

The emergence of deep learning models has brought up questions on the fairness of these models with respect to sensitive attributes. There have been studies on the reliability and bias of deep learning approaches, such as the bias of the learning algorithms due to matching a data distribution in adversarial models (Cohen et al., 2018), visual question answering (Agrawal et al., 2018), image search setups (Kay et al., 2015), and gender classification (Buolamwini & Gebru, 2018). Recent approaches (Madras et al., 2018; Zhao et al., 2020; Creager et al., 2019; Zhang et al., 2018) address the bias in deep learning models and propose fair models that better guarantee fairness criterion while maintaining accuracy. The general idea in these models is applying a fairness learning technique either on the learned latent representation (Madras et al., 2018; Zhao et al., 2020; Creager et al., 2019) or the output eligibility criterion (Zhang et al., 2018), given an input.

Adversarial learning techniques (Goodfellow et al., 2014; Ganin & Lempitsky, 2015) have been extensively used to learn the data distribution of interest, such as learning disentangled latent space (Kim & Mnih, 2018; Chen et al., 2018; 2016), or subtraction of a feature from the latent space (Lample et al., 2017; Denton et al., 2017). Adversarial learning has also been used in fairness models to either remove sensitive information from the latent space (Madras et al., 2018; Zhao et al., 2020; Zhang et al., 2018) or disentangle latent features into sensitive and non-sensitive attributes (Creager et al., 2019). The goal in these approaches is to remove the sensitive information, such as gender, from the latent space that is later used for other tasks, such as eligibility classification. In this paper, we aim to verify whether well-performing and ubiquitous models can effectively reduce bias when evaluated using well-established fairness metrics. Here we focus on deep learning models for classification and their adversarial bias mitigation counterparts.

The difficulty of a task is determined by the distribution of the dataset, which in turn, is a function of the label imbalance, features imbalance, dependency of sensitive attributes to the rest of features, and distribution shift from training to development phase to name a few. In addition, in the current state of fairness community problems with reproducibility and inconsistency in the experimentation and dataset setups makes comparison difficult. Given the importance and possibly tangible impact of proposed algorithms when in production, we advocate for a rigorous and unified evaluation of the capabilities of debiasing models. We argue the need for a systematic analysis that would assess different bias mitigation approaches under the perspective of different fairness metrics. This would help elucidate the most promising research contributions.

To analyse debiasing models we propose a dataset that facilitates creation of different biased setups, such as data imbalance or correlation among eligibility and sensitive or non-sensitive attributes. In particular, contrary to real datasets where the intensity of each feature cannot be controlled, this dataset allows changing only one component while keeping all other components unchanged, something that allows study of different variations of biased setups.

To evaluate models under different biased setups, we provide an in-depth analysis of baselines and recently proposed bias mitigation models in the literature. In particular, we evaluate three promising debiasing models (six variants) together with a baseline model using a unified set of fairness metrics and report results by carrying extensive hyper-parameter search in all cases, ensuring that the drawn conclusions can be attributed to modeling or loss choices. In doing so, we evaluate these models under different setups by training about 2000 models, in which we transition from balanced setups towards challenging unbalanced and correlated setups, where eligibility criterion is correlated with sensitive or non-sensitive attributes. Given the importance of fairness and the serious implications of their misuse, we intentionally try to push these models to their breaking point using extreme settings from our dataset. We consider different ratios for each sensitive group to create “unfair datasets”. This is noteworthy that our analysis is not to undermine the effectiveness of any of these methods, rather setting the expectations and boundaries for different use cases and encouraging the community to follow. To summarize, our contributions are:

- We show when there is a correlation between eligibility and the sensitive attribute, models exploit it for prediction of accuracy and use it even in test cases when such correlation does not exist, which leads to unfair predictions.
- We show that even when there are non-sensitive attributes in the dataset or small visual features in the input image that correlate with eligibility, models can still exploit it and hence be biased.
- We demonstrate that the choice of seed can affect the results significantly and different models show a considerable variation of performance given the seed.
- We provide a deep learning codebase composed of six debiasing models and a baseline to the fairness community for research and evaluation of fairness models.
- We provide a dataset with different controllable sets of features and correlation among them, to facilitate research on fairness models in different unfair setups.

2 Related Works

There is an increasing literature studying how biased datasets can bias learning algorithms to discriminate (Bolukbasi et al., 2016; Cohen et al., 2018; Agrawal et al., 2018). These studies have led researchers to propose new evaluation tools and datasets (Buolamwini & Gebru, 2018), with the goal of identifying potential machine learning error rate gaps among different groups. To remedy this issue, numerous contributions have emerged in order to probe machine learning systems at different levels and reduce discrimination from the perspective of modeling. Friedler et al. (2019) evaluate the fairness of pre-processing, in-processing, and post-processing machine learning approaches; however, no deep learning model is evaluated in their study. Moreover, the models are not evaluated while changing the level of dataset bias or correlation among features. Kamishima et al. (2012) propose a regularization approach for models with probabilistic discriminative nature and evaluate it on logistic regression models; (Zafar et al., 2017a) proposes a flexible mechanism to optimize accuracy in the presence of non-convex fairness constraints (or vice-versa, optimize fairness with accuracy constraints) and evaluate it on logistic regressions and support vector machines.

In addition, several contributions have attempted to leverage adversarial learning to mitigate biases in learned representations or in classifiers. Adversarial learning was originally introduced within the framework of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014)and it has been leveraged in the bias mitigation frameworks to make different groups indistinguishable from one another with respect to a sensitive attribute. In particular, (Edwards & Storkey, 2016) learns adversarially to debias encoded features to be used in classification tasks, while being fair with respect to demographic parity. Madras et al. (2018) adopt a group normalized ℓ_1 loss and adapt it for each fairness metric of interest. Beutel et al. (2017) debias the encoded features using an adversarial training procedure for classification tasks to achieve equality of opportunity. Zhang et al. (2018)

directly pass the classifier's output, rather than the encoded features, to a discriminator to debias the classifier with respect to different fairness metrics.

Despite the recent progress in fair deep learning models, the comparison of bias mitigation algorithms in a common setup is not straightforward, especially in biased and correlated dataset setups. Verma & Rubin (2018) evaluate how fair an off-the-shelf logistic regression model is, given a set of fairness metric definitions. In this work, we evaluate some of the most recent and promising deep learning models in a common deep representation setup.

Adversarial techniques can be categorized into three main approaches: (i) applying debiasing through adversarial training directly on the class label, where the class label is an indicator of eligibility (Zhang et al., 2018), (ii) mitigating bias through enforcing group-fairness on the learned latent space where it is directly used for classification of eligibility (Beutel et al., 2017; Madras et al., 2018; Zhao et al., 2020), and (iii) discarding sensitive representation features for downstream tasks after disentangling learned latent space into sensitive and non-sensitive ones (Creager et al., 2019; Träuble et al., 2020). This line of research is motivated by recent development in disentangled representation learning (Kim & Mnih, 2018; Chen et al., 2018; Locatello et al., 2019). In this paper, we take some of the promising models from the above-mentioned categories and assess their merits in challenging scenarios, in particular, in under-represented and correlated dataset setups. This exercise is not to undermine the applicability or achievements of these models but rather to understand the robustness and boundaries of these models under a variety of settings.

3 Models

In this study, we empirically analyze the performance of a baseline MLP (multi-layer perceptron) model without any bias mitigation learning criteria and compare it with LAFTR (Madras et al., 2018), which applies adversarial learning to achieve group fairness, CFAIR (Zhao et al., 2020) which performs conditional alignment of representations to achieve a better accuracy-fairness trade-off, and finally, FFVAE (Creager et al., 2019), which disentangles latent representation into sensitive and non-sensitive attributes.

For simplicity, we introduce the following shorthand notation. We denote by $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$ the set of inputs and outputs, respectively. Two sets of random variables X and Y take associated values $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Variable y determines eligibility which in our MNIST-based dataset refers to a digit is even or odd. We denote the sensitive binary variable by S taking values $s \in \{0, 1\}$. Moreover, p indicates the output probability of the classifier as a real number and $\hat{Y}$ is the class predicted random variable based on p, which takes values $\hat{y} \in \{0, 1\}$ using a threshold. For each s and y, the term $\mathcal{D}_s^y$ is the conditional distribution of the joint distribution $\mathcal{D}$ over $X \times Y \times S$, given $Y = y, S = s$.

Baseline Model. We use an MLP as our baseline model, which given an input image x, predicts the probability p of the input being odd or even. This probability is then transformed into a classification prediction $\hat{y}$. This model does not leverage any bias mitigation algorithm and is meant to show how fair a base deep learning model would perform under different fairness criteria.

Learning Adversarially Fair and Transferable Representations (LAFTR). LAFTR (Madras et al., 2018) is an adversarial based bias mitigation algorithm within the scope of representation learning. In the supervised version of LAFTR, given an input image x, it first learns a latent encoded representation z that is passed to the discriminator to be debiased. The learned representation is then passed to a classifier to predict the task of interest y. The discriminator is trained by minimizing

$$\mathbb{E}_{x,y,s \in \mathcal{D}} \mathcal{L}_{\mathcal{S}}(D(z, y), s) \tag{1}$$

where y is only passed in debiasing models aimed for *equality of odds* and *equality of opportunity*. The encoder and classifier are trained jointly by minimizing

$$\mathbb{E}_{x,y,s \in \mathcal{D}} \mathcal{L}_{\mathcal{Y}}(C(z), y) - \gamma \mathcal{L}_{\mathcal{S}}(D(z, y), s) \tag{2}$$

where $z = E(x)$ is the encoded feature, which is passed to both the classifier C and the discriminator D. The first term measures the classification loss and the second term gets the adversarial gradients from the discriminator regarding the sensitive attribute s.

Following the original paper, we considered four variants of this model. LAFTR-DP objective is defined as

$$\mathbb{E}_{x,s\in\mathcal{D}}\mathcal{L}_S^{\text{DP}}(D(z),s) = 1 - \sum_{s\in\{0,1\}} \mathbb{E}_{x,s\in\mathcal{D}_s}|D(z)-s| \tag{3}$$

and LAFTR-EqOpp0 objective is considered as

$$\mathbb{E}_{x,y,s\in\mathcal{D}}\mathcal{L}_S^{\text{EqOpp0}}(D(z,y),s) = 1 - \sum_{s\in\{0,1\},y=0} \mathbb{E}_{x,s\in\mathcal{D}_s^y}|D(z)-s| \tag{4}$$

$\mathcal{L}_S^{\text{EqOpp1}}$ is obtained by replacing $y = 1$ in Eq. (4) and $\mathcal{L}_S^{\text{EqOdd}}$ is the sum of $\mathcal{L}_S^{\text{EqOpp0}}$ and $\mathcal{L}_S^{\text{EqOpp1}}$.

Conditional Learning of Fair Representations (CFAIR). Proposed by (Zhao et al., 2020) this model leverages two adversarial networks h_0 and h_1, predicting sensitive attribute s respectively for class labels $Y = 0$ and $Y = 1$. CFAIR depends on an objective function called the balanced error rate (BER) (Feldman et al., 2015; Menon & Williamson, 2018), which guarantees small joint error across demographic groups. The BER represents the sum of false positive rate and false negative rate. Therefore, it is equal to minimizing the below two conditional errors. $\text{BER}_{\mathcal{D}}(\hat{Y}\|Y)$ is defined as

$$\text{BER}_{\mathcal{D}}(\hat{Y}\|Y) \propto p(\hat{Y}=1|Y=0) + p(\hat{Y}=0|Y=1). \tag{5}$$

and $\text{BER}_{\mathcal{D}}(\hat{S}\|S)$ is defined similarly, where $\hat{S}$ is the predicted sensitive random variable. CFAIR is optimized based on the following min-max formulation.

$$\min_{C,E}\max_{h_0,h_1} \quad \text{BER}_{\mathcal{D}}(C(E(X))\|Y) - \gamma\left(\text{BER}_{\mathcal{D}^{y=0}}\left(h_0(E(X))\|S\right) + \text{BER}_{\mathcal{D}^{y=1}}\left(h_1(E(X))\|S\right)\right) \tag{6}$$

This approach proposes that using the balanced error rate along with the conditional alignment helps in achieving equalized odds across the groups without impacting demographic parity.

Flexibly Fair Representation Learning by Disentanglement (FFVAE). Inspired by FactorVAE (Kim & Mnih, 2018), FFVAE (Creager et al., 2019) performs disentanglement by factorizing latent space. It learns a disentangled representation of the inputs, which is flexibly fair in the sense that it can be easily modified at test time to achieve demographic parity across various groups.

In our setup, x is the input image, which is considered to implicitly contain both sensitive and non-sensitive attributes. Moreover, $a = (a_1, \ldots, a_N)$, $b = (b_1, \ldots, b_N)$, and z are the sensitive attribute, the sensitive latent, and non-sensitive latent of x, respectively, with N indicating the number of sensitive or non-sensitive features, depending on the dataset.

FFVAE trains an encoder $q(z, b|x)$, a decoder $p(x|z, b)$, as well as an adversarial network. The latent representation is disentangled into sensitive and non sensitive latent attributes by encouraging both $\text{MI}(b, z)$ and $\text{MI}\left(b_i, a_j\right), \forall i \neq j$ to be low, where MI represents mutual information. FFVAE objective is defined as

$$\mathcal{L}_{\text{FFVAE}}(p,q) = \mathbb{E}_{q(z,b|x)}[\log p(x \mid z,b) + \alpha \log p(a \mid b)]] - D_{KL-\text{FFVAE}} \tag{7}$$

where

$$D_{KL-\text{FFVAE}} = \gamma D_{KL}(q(z,b)\|q(z)\prod_j q\left(b_j\right)) + D_{KL}(q(z,b \mid x)\|p(z,b)) \tag{8}$$

The first term of Eq. (7) consists of a reconstruction term (on left) and a *predictiveness* term $p(a \mid b)$, which aligns sensitive attributes to its respective sensitive latents. Eq.(8) has two terms; the *disentanglement* term on the left decorrelates the sensitive latent representation b from z using an adversarial network and the second term is the Kullback–Leibler divergence between the prior and the latent distribution.

In addition to the above-mentioned models, we also did experiments with (Zhang et al., 2018) (based on the code released by authors), however, the model was very unstable on our dataset configurations even after extensive hyper-parameter search. We hypothesize that this is due to the application of adversarial training directly to the class labels, which makes the model unstable, as indicated by the authors. Due to unstable results, we dropped this model from our evaluation.

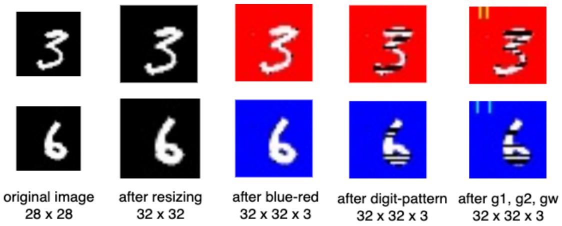

Figure 1: We show the conversion process used to generate our dataset and add our dataset attributes in the last 3 columns to the resized image. *e-o-ratio* affects the number of lines observed on the image. g_1/g_2 and g_w respectively affect the location and size of the green pixels (small vertical lines on top-left of the image).

4 Setup

Machine learning algorithms have attracted much attention and found application in a wide range of highly risky domains such as medical tests and criminal justice. It is important to make sure that the unfairness present in datasets (Cohen et al., 2018; Agrawal et al., 2018; Kay et al., 2015; Buolamwini & Gebru, 2018), is not learned and later inferred by powerful deep learning models used for decision making (Barocas & Selbst, 2016; Berk et al., 2018; of the President, 2016). One way to achieve this goal is to measure the ability of the current prominent debiasing models in ignoring the correlation existing in the datasets and evaluate these models in setups where we have a mechanism to control the correlations among dataset features.

Colored-MNIST. In order to evaluate the debiasing models in challenging setups and be capable of controlling different dataset configurations, we design a variant of the MNIST dataset (LeCun & Cortes, 2010), called Colored-MNIST, where we introduce different types of correlations between sensitive attribute, dataset features, and eligibility criterion. For an input image x, the label $y \in \{0, 1\}$ indicates eligibility or ineligibility, given x is even or odd. We use the background colors as the protected or sensitive attribute $s \in \{0, 1\}$, where blue indicates the unprivileged group and red indicates the privileged group. The task of interest is to classify whether, given an input image x, the class label y is odd or even. While a model improves its classification accuracy, we aim at observing its performance given different fairness metrics. We would like to also observe how the model's classification accuracy and fairness measures change in different dataset scenarios. A common practice is to show reasonable performance on a few standard datasets. These, however, do not cover the majority of possible distributions of a dataset that a model can encounter in real-world development and production. Here we are using this dataset as a benchmarking tool that will allow control over these factors to give a better sense of what a model is capable of in different challenging setups.

4.1 Dataset features

Label (y): The label indicates whether the digit is even or odd. In our dataset, even is eligible ($y = 1$) and odd is ineligible ($y = 0$).

Background color (bck): It refers to the background color (sensitive attribute) which is various shades of blue and red colors.

clr-ratio: shown in pair (b_e, b_o), $b_e(b_o)$ refers to the ratio of the even (odd) digits with blue backgrounds and the rest of the digits will have red backgrounds. Using this feature we control the correlation between the sensitive attribute and the eligibility.

e-o-ratio: This feature indicates the number of drawn lines on the digit. *e-ratio* (*o-ratio*) refers to the ratio of even (odd) digits that contain 6 to 20 lines. The rest of the digits have between 2 to 6 lines. Using *e-o-ratio*, we can completely correlate $(1, 0)$ or completely decorrelate $(0.5, 0.5)$ the number of lines on the digit (as a non-sensitive attribute) with the eligibility.

Green width (g_w): This value indicates the number of green pixels (either 1 or 5) in the image and is visualized as a small vertical line (see Figures 1 and 5). The location of green pixels is indicated by g_1 and g_2.

$\boldsymbol{g_1}$**:** This variable indicates the location of the first green column in the image and is set to one of $\{2, 22\}$. For each given even (odd) digit, the location of the first green column is randomly drawn in the range of $[0, g_1/2)$ ($[g_1/2, g_1)$). Hence, if $g_1 = 2$, the location of the green column is an easy-to-spot indicator of eligibility since the same pixel location (0 or 1) indicates eligibility/ineligibility, while if

$g_1 = 22$, the model needs to observe over many cases, hence it requires a stronger analysis across different cases.

$\boldsymbol{g_2}$**:** This variable indicates the location of the second green column in the image and is set to $g_1 + digit$ where $digit$ varies from 0 to 9. We used g_2 in all our experiments along with g_1.

By using g_1, g_2, and g_w, we evaluate the following: when we have features that are visually small (1 or 5 pixels big), can models exploit the correlation between g_1 and eligibility, in an easy-to-spot case ($g_1 = 2$) and a more challenging case ($g_1 = 22$)?

4.2 Fairness Metrics

We use the same set of fairness metrics, as reported in LAFTR, CFAIR, and FFVAE, to evaluate the debiasing models. Table 1 presents the fairness metrics as well as their mathematical notations and abbreviations. In Table 2 we provide a comprehensive list of fairness metrics in the literature, which we have implemented in our framework and will release to the community. Due to the limit of space and more widespread usage of the metrics presented in Table 1, we only report on these metrics. The provided tool by (Friedler et al., 2019) is used to compute all of the metrics.

Fairness Criteria	**Formulation**	**Abbreviation**
Demographic Parity	$1 - \vert p(\hat{Y} = 1 \vert S = \text{Protected}) - p(\hat{Y} = 1 \vert S = \text{Unprotected}) \vert$	DP
Equality of Opportunity w.r.t $y = 1$	$1 - \vert p(\hat{Y} = 1 \vert Y = 1, S = \text{Unprotected}) - p(\hat{Y} = 1 \vert Y = 1, S = \text{Protected}) \vert$	EqOpp1
Equality of Opportunity w.r.t $y = 0$	$1 - \vert p(\hat{Y} = 1 \vert Y = 0, S = \text{Unprotected}) - p(\hat{Y} = 1 \vert Y = 0, S = \text{Protected}) \vert$	EqOpp0
Equality of Odds	$0.5 \times [\text{EqOpp0} + \text{EqOpp1}]$	EqOdd
unprotected-accuracy	$p(\hat{Y} = y \vert Y = y, S = \text{Unprotected})$	up-acc
protected-accuracy	$p(\hat{Y} = y \vert Y = y, S = \text{Protected})$	p-acc
accuracy	$0.5 \times [\text{up-acc} + \text{p-acc}]$	acc

Table 1: X, Y, S denote the input sample, ground truth label, and the sensitive attribute. $\hat{Y}$ and p are the model's prediction and the output probability of the model. For all metrics, 1 indicates the perfect and 0 the worst value.

5 Experimental Results

In this section, we evaluate the baseline MLP model as well as the debiasing models (LAFTR-DP, LAFTR-EqOpp0, LAFTR-EqOpp1, LAFTR-EqOdd, CFAIR, and FFVAE) on the Colored-MNIST dataset and evaluate their performance on different fairness metrics. The goal is to provide an evaluation baseline for different debiasing models and meanwhile measure how fair they remain in unbalanced and biased situations. In particular, we change the dataset from a balanced setup to different unbalanced setups. In the balanced setup, the ratio of unprivileged (blue) and privileged (red) background is at 50% for both eligible and ineligible groups, giving *clr-ratio* of $(0.5, 0.5)$, and the data features and eligibility are uncorrelated with *e-o-ratio* being $(0.5, 0.5)$. In the unbalanced cases, while we deviate from the balanced setup during training, at test time, we evaluate the models in a balanced setup, meaning the dataset is at 50% for eligible and ineligible groups as well as for privileged and unprivileged groups. This way, we report the model's performance on a balanced setup of different sub-groups. Experimental and model details are provided in Section B.

Impact of reducing representation of unprivileged group. In this experiment, we change *clr-ratio* from the balanced setup of $(0.5, 0.5)$ to unbalanced cases of $(0.1, 0.1)$, $(0.01, 0.01)$, and $(0.001, 0.001)$. In these setups, the unprivileged group with the blue background is less and less represented in both eligible and ineligible groups, hence evaluating the models where the unprivileged groups are under-represented, but represented equally in both eligible and ineligible groups. Figure 2 compares two models side-by-side, Figure 6 and Tables 4 to 10 in the Supplementary in Section C.1 show all models and give detailed results. We initially averaged metrics over three seeds, then for each metric, the best value is reported, meaning the performance of the best-performing model on the fairness metric (chosen across hyper-parameters on the validation set) is reported on the test set. This would allow us to report the best performance on each metric, without choosing how to compromise between accuracy and a fairness metric (which usually have a trade-off and improving one deteriorates the other). We want to show that there exists a bias even when the best variation of the model is considered. Note that this is even in favor of the models as the best possible result is reported. Hence in each column we report the best obtained result, which can correspond to a different hyper-parameter or early-stopping point. As observed in the results, through the transition from balanced to unbalanced setups, both accuracy and fairness metrics drop. While LAFTR models maintain better fairness metrics compared to other models, they perform poorly on accuracy metrics. After LAFTR, FFVAE and then CFAIR, can less and less preserve fairness metrics while in the same order they improve on accuracy. The baseline model (MLP), without any debiasing, as expected performs the best on the accuracy, while fails more than other models in maintaining fairness.

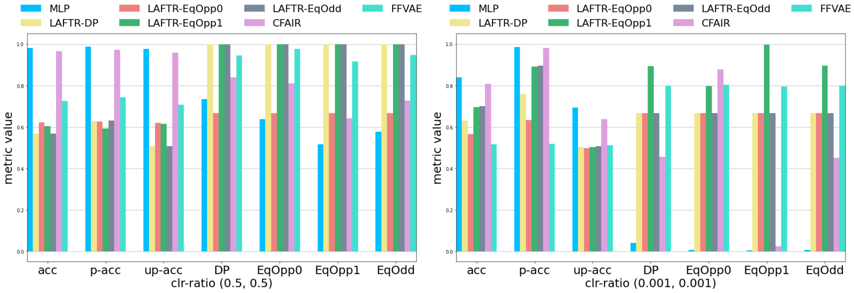

Figure 2: Comparing different models while decreasing minority representation.

Impact of correlation of sensitive attribute with eligibility. In this setup, we change *clr-ratio* from the balanced setup of $(0.5, 0.5)$ to unbalanced cases of $(0.1, 0.9)$ and $(0.01, 0.99)$, where the unprivileged group is under-represented in the eligible and over-represented in the ineligible groups. In unbalanced settings, the resistance of models to the correlation between sensitive attribute (background color) and the eligibility metric is evaluated. The goal is to evaluate the following: when such cases arise, do models perform fairly in test setups, where everything is completely balanced? and hence can they avoid carrying the correlation of training to test setups? Figure 3 compares two models side-by-side, Figure 7 and Tables 11 to 17 in Section C.2 of Supplementary show all models and detailed results. In this setup, FFVAE preserves better the fairness metrics compared to other models. LAFTR and CFAIR models perform similarly or sometimes even worse than the baseline model on fairness metrics, while the baseline model preserves better the accuracy. Missing bars in the plots imply zero metric values and biased predictions on the balanced test set. The results show that even though debiasing models are provided with the sensitive attribute during training, they still fail to decorrelate the sensitive attribute from eligibility in highly correlated setups.

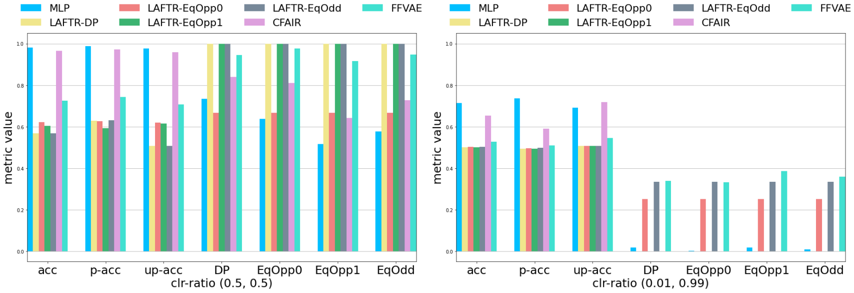

Figure 3: Comparing different models while shifting correlation of sensitive attribute (bck) and the eligibility.

Impact of correlation of non-sensitive attribute with eligibility. In this setup, we keep *clr-ratio* at $(0.5, 0.5)$, however, change *e-o-ratio* from balanced setting of $(0.5, 0.5)$ to unbalanced settings of $(0.9, 0.1)$. This would make the number of lines on the digits correlated with eligibility. The goal is to evaluate the following: to what extent a model can exploit the correlation between a non-sensitive attribute and the eligibility criterion and hence be biased at test time? Figure 8 and Tables 18 to 24 in Section C.3 show results for all models. While MLP, CFAIR, and FFVAE do not show much change in this setup, LAFTR models show more deviation, indicating they are more sensitive to such correlations. For example, LAFTR-DP dropped in fairness measures, while LAFTR-EqOpp0's performance improved. As we show later, this is mainly due to the high variance of this model under different seeds.

Impact of small features in the input images. In this setup, we change g_1 from 22, where the location of the green pixel has a more complex relation with eligibility, to 2, where the location of the green pixel (0 for eligibility and 1 for ineligibility) is an easy-to-spot indicator of eligibility that can be more easily learned by models and hence induce bias. The question is whether models find such small but non-trivial correlated features with eligibility? If the answer is yes, then information even in the small portion of the input can cause biased performance of the model. In this setup we run an experiment where for each one of $g_1 = 2$ and $g_1 = 22$ we consider two cases with $g_w = 1$ and $g_w = 5$, respectively showing only 1 and 5 green pixels in the image. This setup would verify whether the model observes the green part of the image and if yes how much impact the size of that region will have. Figure 4 shows two models, Figure 9 and Tables 25 to 31 show two extra models and give detailed results for all models. While models like MLP, CFAIR, and FFVAE show a small difference in two setups, LAFTR models become less fair in the biased setup of $g_1 = 2$, indicating this model is more prone to pick up on such small details, when it can spot it. Results with $g_w = 5$ show very close performance to $g_w = 1$, indicating even one pixel can be enough for models to extract small details from the images. The results show that when there are small and sometimes not-noticeable features in the dataset, they can cause bias in the learning algorithm.

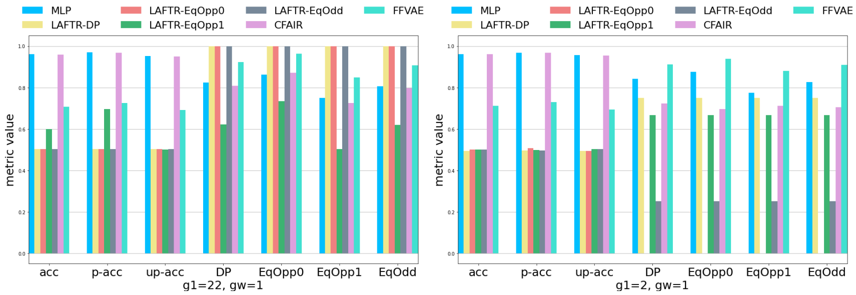

Figure 4: Impact of small visual components on different models' performance.

6 Discussion

6.1 Sources of Bias

In our experiments, we have considered four different sources of bias.

1. **Small percentage of the unprivileged group.** This case causes bias due to the small ratio of the unprivileged group compared to the privileged group. In this case, the bias can be two-fold, one is due to the imbalance between the two groups, another can be due to rareness of the data from the unprivileged group. While the clr-ratios (0.1, 0.1), (0.01, 0.01), and (0.001, 0.001), all are subject to group imbalance, the latter two cases have 250 and 25 samples from the unprivileged group, compared to 24,750, and 24,975 samples respectively from the privileged group. So, the model has to also handle the secondary aspect of the bias, which is different.
2. **Correlation of a feature with eligibility.** In some data the unprivileged group might be less eligible, e.g. due to historic reasons of lack of access to facilities or limited resources compared to privileged group. Such correlations in the data can be picked up by the model and cause bias at inference.
3. **Impact of size of the correlated visual features.** While large features that correlated with eligibility and cause bias can be easily observed by a model, small visual features can also have correlations with eligibility. Given the size of the correlated features, some models can act more fair while others might remain biased.
4. **Impact of unnoticed and small features.** The debiasing algorithms work based on the assumption that the sensitive or biased attributes are known a-priori and hence can be addressed by knowing these features and removing them from the learned representation. However, some features, especially small ones, might not be noticed by the model or the annotator. For example, wearing earrings, can correlate with gender, or glasses with age. Such cases might not be observed and addressed properly by the debiasing model, if they are not know a-priori.

We have studied the impact of the four above-mentioned sources of bias, and have observed a stronger level of bias in the first two cases. Specially, the correlation of big visual sensitive attributes with eligibility, and imbalance among the privileged and unprivileged groups lead to a higher bias. This trend was observed in all models. We observed smaller visual components that are correlated with eligibility cause less bias, as the model can more easily get rid of them. Some models such as LAFTR suffered still in these cases. We believe study of such sources of bias can help better evaluate the debiasing models and advocate their study in future works.

6.2 Model Stability

Variation due to seed. In deep learning models, seed is a great source of variation due to initialization of the model parameters and the sampling done during training, which can converge the model to different solutions in the course of training. To observe its impact, we evaluated how much standard deviation changes when the seed used for a model differs. The results for three seeds considered are shown in Figure 10 for all of the setups described in our experiments. We observed the choice of seed can impact some models more than others. In particular, LAFTR was the least stable model in terms of variation of results given different seeds. We advocate verification of any debiasing method under different seeds, as this indicates how much the model is susceptible to small training variations and hence not suitable for bias mitigation. This is complementary to other studies of stability, such as cross-validation, studied in Friedler et al. (2019).

Correlation between dataset features and model's prediction. If a model's prediction is correlated with dataset features, it indicates it can rely on biases in data in its predictions and hence be unfair when applied. To verify this, for each model, we measured the correlation between dataset features and fairness metrics using Spearman correlation matrices, measured over all of the experiments presented in this section. The results are presented in Figure 11 for the 64-dim architecture. On average the models that capture less correlation between dataset features and fairness metrics act fairer in the experiments, as we also observe from results presented above.

Instability due to the debiasing method. Debiasing algorithm can cause instability due to the way the model mitigates bias. In particular, adversarial models can be unstable in practice and caution needs to be made in properly using them. We found Mubal (Zhang et al., 2018) to be the least stable model, as it applies the adversarial loss directly onto the class label. Applying the adversarial loss to the learned latent space, as done in LAFTR, made this process more stable. However, even this approach performed poorly in many setups and was subject to high variations with a small change in the dataset. Even among different LAFTR variants, the LAFTR model with a given fairness criteria introduced into its objective function, was not always the LAFTR variant performing the best on that particular metric. We believe this is due to application of both accuracy and adversarial loss to the same latent space, which causes a competition between the two losses. CFAIR has a more stable learning due to using two adversarial networks for two eligibility classes and using a balanced error rate as its loss function. Finally, FFVAE, obtained the most stable performance due to disentangling of attributes into sensitive and non-sensitive and also decoding features, which helps the model capture all of the information of the input image.

7 Conclusion

With representation learning algorithms getting more accepted as automatic decision-making tools, we propose an evaluation scenario to better understand the limits of these methods. We systematically compare and analyze a few of the better-known models using our proposed dataset, where we control the correlation in-between different dataset components. Although these models proved their effectiveness to the community and can solve and improve upon some of our dataset variants, we show that we can purposefully push them to their breaking point. Our results indicate that in different setups the models can learn and hence induce a bias when correlations in the dataset exist. This suggests that not all models are suitable for all scenarios and using unsuited models for certain tasks could cause unwanted consequences. We also note that with the abundance of fairness metrics, there exists room for improvement when it comes to model selection as we cannot pinpoint a single metric that works best across all the metrics. This becomes imperative as specifically, practitioners need to be fair across more than one metric and find a trade-off between fairness and accuracy. Finally, we will release our dataset and codebase to the community in order to provide a framework for the evaluation of deep fairness models.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2018.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.. California Law Review*, 104(IR):671, 2016. URL `http://lawcat.berkeley.edu/record/1127463`.

Shai Ben-David, Nadav Eiron, and Philip M. Long. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66:496–514, 2003.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, pp. 004912411878253, 2018.

Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 4356–4364, 2016.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91, 2018.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Advances in Neural Information Processing Systems*, pp. 2610–2620, 2018.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pp. 2172–2180, 2016.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Joseph Paul Cohen, Margaux Luck, and Sina Honari. Distribution matching losses can hallucinate features in medical image translation. In *International conference on medical image computing and computer-assisted intervention*, pp. 529–536. Springer, 2018.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. volume 97 of *Proceedings of Machine Learning Research*, pp. 1436–1445. PMLR, 2019.

Emily L Denton et al. Unsupervised learning of disentangled representations from video. In *Advances in neural information processing systems*, pp. 4414–4423, 2017.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference on Learning Representations*, 2016.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

V. Feldman, V. Guruswami, P. Raghavendra, and Yi Wu. Agnostic learning of monomials by halfspaces is hard. *2009 50th Annual IEEE Symposium on Foundations of Computer Science*, pp. 385–394, 2009.

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. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 329–338. ACM, 2019.

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pp. 1180–1189. PMLR, 2015.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *AISTATS*, 2010.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Matthew Kay, Cynthia Matuszek, and Sean A Munson. Unequal representation and gender stereotypes in image search results for occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 3819–3828, 2015.

Hyunjik Kim and A. Mnih. Disentangling by factorising. In *ICML*, 2018.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic DENOYER, et al. Fader networks: Manipulating images by sliding attributes. In *Advances in Neural Information Processing Systems*, pp. 5963–5972, 2017.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. volume 97, pp. 4114–4124. PMLR, 2019.

David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pp. 3381–3390, 2018.

A. Menon and R. Williamson. The cost of fairness in binary classification. In *FAT*, 2018.

Executive Office of the President. Big data: A report on algorithmic systems, opportunity, and civil rights. 2016.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations, 2020.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Anirudh Goyal, Francesco Locatello, Bernhard Schölkopf, and Stefan Bauer. Is independence all you need? on the generalization of representations learned from correlated data, 2020.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, FairWare '18, pp. 1–7, 2018. ISBN 978-1-4503-5746-3.

M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of International Conference on World Wide Web*, pp. 1171–1180, 2017b.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340. ACM, 2018.

Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. Conditional learning of fair representations. In *International Conference on Learning Representations*, 2020.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. In *The 2nd workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML) at ICML'15*, 2015. doi: arXiv:1505.05723.