ADVERSARIAL FAIRNESS NETWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

Fairness is becoming a rising concern in machine learning. Recent research has discovered that state-of-the-art models are amplifying social bias by making biased prediction towards some population groups (characterized by sensitive features like race or gender). Such unfair prediction among groups renders trust issues and ethical concerns in machine learning, especially for sensitive fields such as employment, criminal justice, and trust score assessment. In this paper, we introduce a new framework to improve machine learning fairness. The goal of our model is to minimize the influence of sensitive feature from the perspectives of both data input and predictive model. To achieve this goal, we reformulate the data input by eliminating the sensitive information and strengthen model fairness by minimizing the marginal contribution of the sensitive feature. We propose to learn the sensitive-irrelevant input via sampling among features and design an adversarial network to minimize the dependence between the reformulated input and the sensitive information. Empirical results validate that our model achieves comparable or better results than related state-of-the-art methods w.r.t. both fairness metrics and prediction performance.

1 INTRODUCTION

In recent years, machine learning has achieved unparalleled success in various fields, from image classification, speech recognition, to autonomous driving. Despite the wide application and rapid development, the discrimination and bias that exist in machine learning models are attracting increasing attention in the research community. Recent models have been found to be biased towards some population groups when making the prediction. Hendricks et al. (2018) identified prediction bias towards gender in image captioning model, where the generation of caption is actually based on contextual information (*e.g.*, location and scenes) but not the visual evidence related with the person in the image. For example, the model is very likely to recognize the person in the image (without using the visual evidence of the person) as a woman if the location is kitchen, while recognize the person as a man if the scene shows snowboarding. In addition, model bias has also been discussed in recidivism prediction. ProPublica (J. Angwin & Kirchner, 2016) analyzed a widely used criminal risk assessment tool for future crime prediction and discovered discrimination among different races. For defendants that do not commit a future crime, the black people are more likely to be mistaken by the model as potential future criminals than the white people (*i.e.*, a higher false positive rate in the black people than the white people).

A model with merely good prediction performance (*e.g.*, high accuracy) is not convincing enough when we harness the power of machine learning. It is critical to guarantee that the prediction is based on appropriate information, and is not biased towards certain groups of population characterized by sensitive features like race and gender. To improve model fairness, recent works propose the strategies from different perspectives. For example in pre-processing, there are efforts on eliminating the bias in data with reweighing the samples (Kamiran & Calders, 2012; Nam et al., 2020), generating fair data (Jang et al., 2021; Sattigeri et al., 2019), or removing the disparity among groups (Feldman et al., 2015). Quadrianto et al. (2019a) improve fairness in image classification by minimizing the relevance between the reformulated input and the sensitive information. While in in-processing methods, there are works improving fairness by constraining the prediction not to be based on sensitive information (Zhang et al., 2018; Mary et al., 2019; Baharlouei et al., 2019; Cho et al., 2020). Adel et al. (2019) also propose an adversarial network that minimizes the influence of sensitive features to the prediction by characterizing the relevance between the latent data representation and

the sensitive feature. Oneto et al. (2019) learn a fair representation that is independent of the sensitive information in a multi-task learning setting. What's more, fairness in prediction can be achieved via post-processing methods (Pleiss et al., 2017) that modifies the model output for equalizing the probability of obtaining a favorable output, *e.g.*, getting approved for a loan.

Based on the targets of fairness, the motivation can be divided into group fairness and individual fairness. Group fairness (Li et al., 2021; Celis et al., 2021) guarantees that different groups of population have equalized opportunity of achieving a favorable prediction result. Whereas for individual fairness (Zemel et al., 2013), the goal is to guarantee that similar individuals get similar output. Based on the motivation of improving fairness, there are recent efforts for improving the long-term benefit of the protected groups (groups that are usually biased against by traditional models) (Liu et al., 2018; Mouzannar et al., 2019), which is different than the methods that focus more on the instant benefit of an equalized opportunity (Pleiss et al., 2017).

Previous models usually propose to improve the fairness w.r.t. either the data perspective or the model perspective, *i.e.*, modifying the input to reduce data bias or optimizing the model to reduce prediction bias. These strategies may not guarantee the learned input to be optimal for the model or the designed model to be optimal for the data, such that a fairness constraint in the model usually introduces deterioration in the prediction performance. In order to improve fairness while maintaining the predictive performance, we propose a new adversarial network to reduce the bias simultaneously from both the *data* perspective and the *model* perspective. By conducting sampling among features, we automatically reformulate the input with features that contain only sensitive-irrelevant information. By minimizing the marginal contribution of the sensitive feature, we strengthen model robustness towards the sensitive feature such that adding sensitive information will not affect the prediction results. The coupled optimization strategy from both the data and the model aspects improves fairness as well as prediction performance.

2 Problem Definition

For a given dataset $[\mathbf{x}^{(1)},\ \mathbf{x}^{(2)},\ \ldots,\ \mathbf{x}^{(n)}]$ consisting of n samples from the input space $\mathcal{X} \subset \mathbb{R}^d$, each sample $\mathbf{x}^{(i)} = [x_1^{(i)},\ x_2^{(i)},\ \ldots,\ x_d^{(i)}]^\top$ is characterized by d features. In a prediction problem, **prediction bias** exists when the model makes different prediction for different groups of samples (characterized by one feature) with all other features held constant. For example, the Home Mortgage Disclosure Act (HMDA) data shows the rate of loan rejection is twice as high for the black people as for the white people (Ladd, 1998).

The **sensitive feature** is the feature to characterize such **groups of population** of interest which we expect the prediction not to be biased towards. Examples of the sensitive feature include *race, gender, age.* The choice of sensitive features varies for different prediction problems. The **sensitive-relevant features** refers to the features that are not regarded as sensitive themselves, but indicate the information relevant to the sensitive feature. One straightforward idea to improve fairness is **fairness through blindness**, *i.e.*, simply exclude the sensitive feature from the input data. However, this cannot eliminate the prediction bias, as the sensitive-relevant features still provide sensitive information in the input data.

The goal of fairness varies in different applications, such as group/individual fairness, the long-term/instant benefit of fairness as introduced in the above section. Here in this work, we are interested in improving the fairness with instant benefit among different groups of population so that the model prediction is not based on the sensitive information, either from the sensitive or sensitive-relevant features.

In this paper, we propose to reduce such prediction bias from two aspects: reformulating the *input data* and strengthening the *model* fairness. We achieve the goal by simultaneously learning a new input $\tilde{\mathbf{x}}$ based on the original data $\mathbf{x}$ and building a prediction model $f^\phi : \mathcal{X} \to \mathcal{Y}$ with parameter ϕ, where $\mathcal{Y}$ is the output space, such that 1) the dependency between $\tilde{\mathbf{x}}$ and the sensitive information is minimized; 2) the influence of the sensitive information to the prediction of f^ϕ is minimized. By improving from both the input data and the model, the model prediction is based on the sensitive-irrelevant information and get enhanced robustness towards the sensitive feature.

To the best of our knowledge, our model is the first to use feature selection for learning a fair representation. Our model is the most related to Adel et al. (2019) and Quadrianto et al. (2019b),

where two related works focus on learning a fair representation in a latent data space. Our model is different in three major perspectives: 1) our model eliminates data bias in the original data space, which preserves the natural meaning of features. This makes the reformulated fair data easier understood and interpreted by end users - which is important in real-world applications; 2) our model optimizes w.r.t. the same predictor for both fairness and performance purposes - we don't need a separate sensitive attribute predictor as in previous work like Adel et al. (2019); 3) we focus on improving fairness in both data and model aspects (we reformulate the data with the features that get the most impacted by the addition of the sensitive feature, and building the predictive model to be the least affected by the sensitive feature).

3 ADVERSARIAL FAIRNESS NETWORK

As we discussed in the above section, the simple strategy of fairness through blindness cannot work with the existence of sensitive-relevant features. In order to reduce the prediction bias, we need to guarantee the prediction is not dependent on either the sensitive feature or the sensitive-relevant features. This is difficult to achieve since we usually do not have prior knowledge of what are the sensitive-relevant features. In this section, we propose a new FAIrness through AdverSarial network (FAIAS) model to improve the prediction fairness by improving both the data input and the model.

The goal of reducing the prediction bias from both the input and model aspects can be formulated as two folds: 1) from the perspective of input, we propose to learn the new input $\tilde{\mathbf{x}}$ based on the original data $\mathbf{x}$ such that $\tilde{\mathbf{x}}$ contains only sensitive-irrelevant information; 2) for the prediction model, we minimize the marginal contribution of the sensitive feature such that adding the sensitive feature does not change the model prediction too much.

We propose to learn the new input $\tilde{\mathbf{x}}$ by sampling the features in the original data $\mathbf{x}$, *i.e.*, selecting features with a selection function $S : \mathcal{X} \rightarrow \{0, 1\}^d$, such that the selected features contain only sensitive-irrelevant information.

Given a data sample $\mathbf{x} = [x_1, \ \ldots, \ x_d]^\top \in \mathcal{X}$, corresponding label $\mathbf{y} = [y_1, \ \ldots, \ y_c]^\top \in \mathcal{Y}$, and a selection set $\mathbf{s} = \{s_1, \ s_2, \ \ldots, \ s_m\} \subset \{1, \ 2, \ \ldots, \ d\}$, denote $f^\phi(\mathbf{x}, \mathbf{s}) = f^\phi([x_{s_1}, \ x_{s_2}, \ \ldots, \ x_{s_m}])$ as the output of function f^ϕ when the input contains only features selected by $\mathbf{s}$ (the value of not selected features is set to 0). For $t \notin \mathbf{s}$, the marginal contribution of the t-th feature to this input can be denoted as $\mathcal{L}\big(f^\phi(\mathbf{x}, S), f^\phi(\mathbf{x}, S \cup \{t\})\big)$, *i.e.*, the change in the output when adding the t-th feature. $\mathcal{L}$ is a loss function to describe the difference between $f^\phi(\mathbf{x}, S)$ and $f^\phi(\mathbf{x}, S \cup \{t\})$.

Denote the sensitive feature as x_k[1], the goal of FAIAS is to minimize the distance between the distribution $p(\hat{y}|x, S)$ and $p(\hat{y}|x, S \cup \{k\})$. In order to achieve this goal, we propose to minimize $\mathcal{L}\big(f^\phi(\mathbf{x}, S), f^\phi(\mathbf{x}, S \cup \{k\})\big)$, where S is the selection function that selects only features containing sensitive-irrelevant information. It is notable that reformulating the input with the selection function S has several advantages:

- Compared with learning a non-interpretable representation, the selection of features maintains *interpretation* of the input, since the natural meaning of features is kept;
- The selection function can be data-dependent, which maintains the *flexibility* such that we learn different sensitive-relevant features for different samples;
- Removing the sensitive-relevant features in the original data space is *theoretically supported* (Kusner et al., 2017), such that learning the observable non-descendants of sensitive feature (*i.e.*, sensitive-irrelevant features in our paper) only needs partial causal ordering without further causal assumptions.

We can approximate the selection function S using a continuous selector function $g^\theta : \mathcal{X} \rightarrow [0, 1]^d$ with parameter θ, that takes the feature vector as the input and output a probability vector $\mathbf{p} = [p_1, \ p_2, \ \ldots, \ p_d] \in \mathbb{R}^d$ showing the probability of sampling each feature to formulate the input. Then we conduct random sampling of the features based on the probability vector $\mathbf{p}$ and get the selection set $\mathbf{s}$. The probability of getting a joint selection vector $\mathbf{s} \in \{0, 1\}^d$ is

$$\pi_\theta(\mathbf{x}, \mathbf{s}) = \Pi_{j=1}^d \big(g_j^\theta(\mathbf{x})\big)^{s_j} \big(1 - g_j^\theta(\mathbf{x})\big)^{(1-s_j)}.$$

[1]For simplicity, here we only consider one sensitive feature. Our FAIAS model can easily apply to the case involving multiple sensitive features.

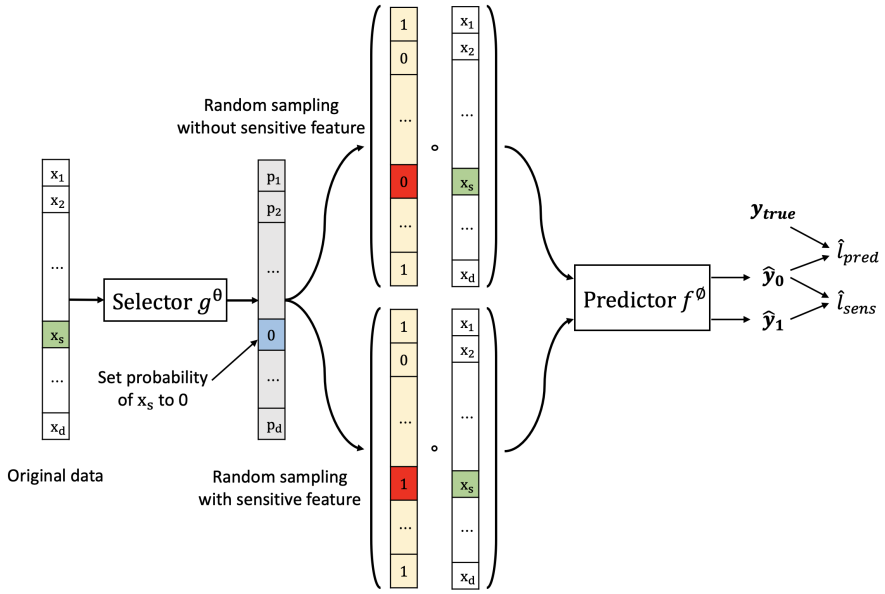


Figure 1: Illustration of the FAIAS model. FAIAS consists of a selector g^θ and a predictor f^ϕ. The selector g^θ takes the feature vector as an input and predict the probability for each feature to be selected, based on which we randomly sample the features. The predictor f^ϕ gets two inputs, one (shown in the upper dot product) is the reformulated input using the sampled features, the other (shown in the bottom dot product) is by adding the sensitive feature to the sampled features. The difference between the output of f^ϕ w.r.t. the two inputs is the sensitivity loss $\hat{l}_{sen}$, which shows the marginal contribution of the sensitive feature to the input. The prediction loss $\hat{l}_{pred}$ shows the prediction performance by using only sampled features.

To quantify the influence of sensitive feature, we consider the cross entropy loss for $\mathcal{L}$ and formulate the sensitivity loss as:

$$l_{sen}(\theta, \phi) \quad = \quad \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\Big[-\sum_{l=1}^{c} f_l^\phi(\mathbf{x},\mathbf{s})\log f_l^\phi(\mathbf{x},\mathbf{s}\cup\{k\})\Big],$$

which characterize the marginal contribution of sensitive feature x_k to model prediction given features selected by $\mathbf{s}$.

In order to optimize g^θ to approximate the selection function S and assign higher probability to only sensitive-irrelevant features, we propose an adversarial game between the selector function g^θ and the predictor function f^ϕ.

The goal of the prediction function f^ϕ is to minimize the sensitivity loss in equation 1 such that adding the sensitive feature does not influence the prediction too much. In contrast, we optimize the selector function g^θ to maximize the sensitivity loss in equation 1, so as to select the subset of features which can be influenced the most by adding the sensitive feature. In this way, the selector function g^θ can find the features that are not intrinsically relevant to the sensitive feature. If for example, the selected subset contains sensitive-relevant features, adding the sensitive feature will not bring too much change since the sensitive information is already indicated by the sensitive-relevant features. By updating the selector function g^θ to maximize the sensitivity loss, g^θ learns to exclude the sensitive information by assigning lower sampling probability to sensitive-relevant features and formulate the input on the basis of only sensitive-irrelevant information.

Moreover, we optimize the predictor f^ϕ and g^θ to minimize the following prediction loss:

$$l_{pred}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\Big[-\sum_{l=1}^{c} y_l \log f_l^\phi(\mathbf{x},\mathbf{s})\Big],$$

which measures the performance of the prediction model given the features selected by $\mathbf{s}$. Here we consider the cross entropy loss for l_{pred}. We illustrate the overview of FAIAS model in Figure 1.

Algorithm 1 Optimization Algorithm of FAIAS Model

Input dataset $\mathcal{Z} = (\mathcal{X} \times \mathcal{Y}) = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$, learning rate α_θ and α_ϕ.
Output selector g^θ and predictor f^ϕ.
Initialize parameter θ and ϕ randomly.
while not converge **do**
for t = $1,\ 2,\ \ldots,\ n_b$ **do**
for $(\mathbf{x}_{t_i}, \mathbf{y}_{t_i})$ in the t-th mini-batch $\mathcal{Z}_t$ **do**
1. Calculate the selection probability vector

$$g^\theta(\mathbf{x}_{t_i}) = [p_{t_i}^1,\ p_{t_i}^2,\ \ldots,\ p_{t_i}^d].$$

2. Sample the selection vector $\mathbf{s}_{t_i} \in \mathbb{R}^d$ with

$$\mathbf{s}_{t_i}^j \sim Bernoulli(p_{t_i}^j), \quad \text{for } j = 1,\ ,2,\ \ldots,\ d.$$

3. Calculate

$$\hat{l}_{pred}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}, \mathbf{y}_{t_i}) = -\sum_{l=1}^{c} (y_{t_i})_l \log f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}),$$

$$\hat{l}_{sen}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) = -\sum_{l=1}^{c} f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) \log f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\}).$$

end for
4. Update the parameter θ with gradient ascent

$$\theta \leftarrow \theta + \frac{\alpha_\theta}{n_b} \sum_i \big(\hat{l}_{sen}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) - \hat{l}_{pred}(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}, \mathbf{y}_{t_i})\big) \nabla_\theta \log \pi_\theta(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}).$$

5. Update the parameter ϕ with gradient descent

$$\phi \leftarrow \phi + \frac{\alpha_\phi}{n_b} \sum_i \sum_{l=1}^{c} y_l \frac{\nabla_\phi f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i})}{f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i})} + \frac{\alpha_\phi}{n_b} \sum_i \sum_{l=1}^{c} f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) \frac{\nabla_\phi f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\})}{f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\})}$$
$$+ \frac{\alpha_\phi}{n_b} \sum_i \sum_{l=1}^{c} \nabla_\phi f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i}) \log f_l^\phi(\mathbf{x}_{t_i}, \mathbf{s}_{t_i} \cup \{k\}).$$

end for
end while

The parameter θ and ϕ can be updated via gradient methods. We can easily derive the derivative of $l_{sen}(\theta, \phi)$ w.r.t. parameter θ and ϕ as

$$\nabla_\theta l_{sen}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)} \Big[\hat{l}_{sen}(\mathbf{x}, \mathbf{s}) \nabla_\theta \log \pi_\theta(\mathbf{x}, \mathbf{s})\Big],$$

and

$$\begin{aligned} &\nabla_\phi l_{sen}(\theta, \phi) \\ = \ &\mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)} \Big[-\sum_{l=1}^{c} \nabla_\phi f_l^\phi(\mathbf{x}, \mathbf{s}) \log f_l^\phi(\mathbf{x}, \mathbf{s} \cup \{k\}) - \sum_{l=1}^{c} f_l^\phi(\mathbf{x}, \mathbf{s}) \frac{\nabla_\phi f_l^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})}{f_l^\phi(\mathbf{x}, \mathbf{s} \cup \{k\})} \Big]. \end{aligned}$$

The derivative of $l_{pred}(\theta, \phi)$ w.r.t. θ is similar to $\nabla_\theta l_{sen}(\theta, \phi)$, that is,

$$\nabla_\theta l_{pred}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)} \Big[\hat{l}_{pred}(\mathbf{x}, \mathbf{s}, \mathbf{y}) \nabla_\theta \log \pi_\theta(\mathbf{x}, \mathbf{s})\Big],$$

and w.r.t. ϕ is

$$\nabla_\phi l_{pred}(\theta, \phi) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)} \big[-\sum_{l=1}^{c} y_l \frac{\nabla_\phi f_l^\phi(\mathbf{x}, \mathbf{s})}{f_l^\phi(\mathbf{x}, \mathbf{s})} \big].$$

In Algorithm 1, we summarize the optimization steps of FAIAS model. According to the update rules w.r.t. the gradients, the time complexity of our FAIAS model is linear w.r.t. the number of samples n, the number of parameters in θ and ϕ, as well as the number of iterations T.

4 EXPERIMENTS

In this section, we conduct experiments on three benchmark datasets to validate the performance of our FAIAS model. The experiments evaluate: 1) whether FAIAS improves the prediction fairness among different groups w.r.t. sensitive features; 2) how will the prediction performance get affected by including fairness constraints in the FAIAS model.

4.1 EXPERIMENTAL SETUP

Notably, our FAIAS model is proposed for group fairness, *i.e.*, minimizing the prediction bias w.r.t. a certain sensitive feature in both the pre-processing and in-processing steps. We compare our model with four recent methods for group fairness in pre-processing, in-processing, and post-processing steps, and one baseline method, which includes: **Adv_Deb** (Zhang et al., 2018), **CEOP** (Pleiss et al., 2017), **DIR** (Feldman et al., 2015), **Reweigh** (Kamiran & Calders, 2012), **LAFTR** (Madras et al., 2018), **Baseline**: a 5 layered neural network with 200 units for all hidden layer (same structure as the predictor f^{ϕ} in FAIAS) that adopts all features (including the sensitive feature) in training and prediction, *i.e.*, the difference between Baseline and FAIAS is that Baseline method use all features as the input, while FAIAS use only sensitive-irrelevant features.

We use three benchmark datasets to evaluate the model, which include: **Adult** (also know as Census Income) data from the UCI repository (Kohavi, 1996), **COMPAS**[2], **CelebA image dataset**[3] (Liu et al., 2015). We use **classification accuracy** and **true positive rate** to evaluate the model prediction performance in classification. Moreover, we adopt three different metrics to evaluate fairness among groups of population w.r.t. the sensitive feature in the data, which includes: **absolute equal opportunity difference**, **absolute average odds difference**, and **disparate impact**.

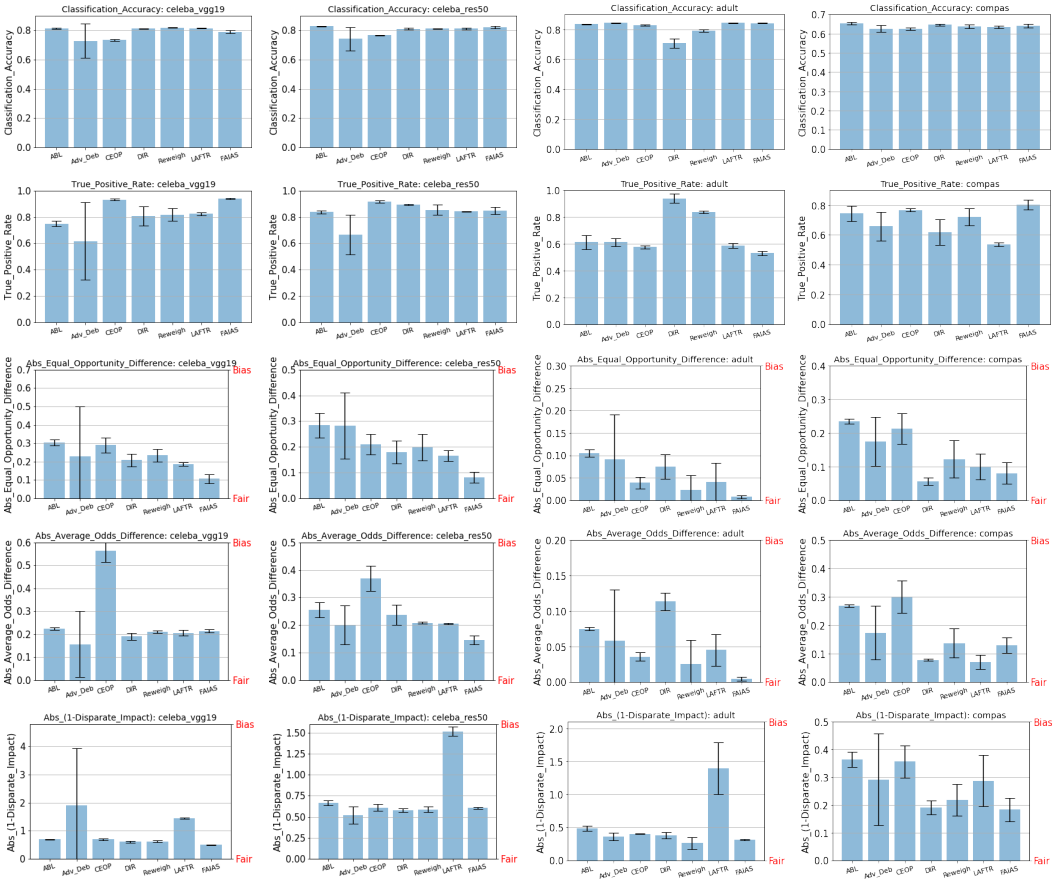
Detailed description of comparing methods, experimental setup, and evaluation metrics are in the Supplementary material. We use Tensorflow and Keras toolbox for implementing our code and run the algorithm on a machine with Quadro RTX 6000 GPU.

4.2 QUANTITATIVE COMPARISON ON BENCHMARK DATA

We compare the model performance and summarize the results in Figure 2. The results show that FAIAS achieves comparable or better classification results w.r.t. both the accuracy and the true positive rate, which indicate that the optimization on both the data and model perspective is successful in guaranteeing the prediction performance such that imposing the fairness constraints does not sacrifice the classification performance. We also use the three fairness metrics to evaluate if our model improves the prediction fairness by rendering equal prediction performance among different groups of population. We notice that FAIAS achieves equivalent or better results w.r.t. all three measurement metrics on the three benchmark datasets, such that the feature sampling via an adversarial network can eliminate the sensitive information and forces the prediction performance to be equalized among different groups of the population. Particularly, equal opportunity difference outperforms the other methods in all datasets. Baseline and FAIAS employ exactly the same classifier structure (f^{ϕ} in FAIAS) and the only difference lie in the input features (Baseline use all features as input). From the comparison, we can validate that the selector in FAIAS is properly filtering out the sensitive features and effectively increasing fairness without sacrificing classification performance. It is notable that though the removal of sensitive-relevant features sometimes harms the performance because sensitive-relevant features can be also target-relevant, it is beneficial for fairness (much improved fairness than ABL in Figure 2. Besides, based on the design of FAIAS, the model can select a set of features to improve fairness (by eliminating the sensitive-relevant features) while maintaining comparable discriminative power (by optimizing the predictor using the sensitive-irrelevant information). We show more results on VGG16, Adult and Compas datasets in the Supplementary material.

[2]https://github.com/propublica/compas-analysis

[3]http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html



(a) VGG19 (b) ResNet50 (c) Adult (*sex*) (d) Compas (*race*)

Figure 2: Comparison of classification performance (top two rows) and fairness (bottom three rows) on three benchmark datasets (with sensitive feature shown in the parenthesis). We use three pre-trained models (VGG19 and ResNet50) to extract 1,000 latent features for images in CelebA dataset (sensitive feature is *sex*). Higher accuracy and true positive rate indicates better classification performance. Lower values for all three fairness metrics shows better fairness.

Model	TPR Male	TPR Female	Abs Eq Opp Diff
VGG16 - Baseline	0.557 ± 0.009	0.908 ± 0.002	0.351 ± 0.008
VGG19 - Baseline	0.592 ± 0.009	0.882 ± 0.002	0.289 ± 0.010
ResNet50	0.644 ± 0.010	0.921 ± 0.003	0.289 ± 0.008
Fair ResDecomp (Quadrianto et al., 2019b)	0.614	0.852	0.238
VGG16 - FAIAS	0.823 ± 0.048	0.980 ± 0.007	0.156 ± 0.005
VGG19 - FAIAS	0.811 ± 0.051	0.979 ± 0.002	0.168 ± 0.051
ResNet50 - FAIAS	0.849 ± 0.004	0.940 ± 0.014	0.091 ± 0.054

Table 1: Comparison of true positive rate (TPR) and absolute equal opportunity difference (Abs Eq Opp Diff) on CelebA data. Higher TPR value indicates better performance, and lower Abs Eq Opp Diff means better fairness. Results of (Quadrianto et al., 2019b) are directly referred from its paper.

4.3 Image Classification with FAIAS

We further look into the image classification task on CelebA data and validate how FAIAS improve the prediction w.r.t. both classification performance and fairness. To evaluate the performance, we further compared our FAIAS model with related image classification models: pre-trained vanilla models (VGG16, VGG19 and ResNet50 as the baseline) and one state-of-the-art method that is

Figure 3: Visualization of reconstructed image based on different sets of feature. The first row shows the original image. The second row shows the reconstructed image from all 1,000 features. The third row shows the reconstructed image after zeroing out the 200 (female), 400 (male) sensitive-relevant features learned by FAIAS.

proposed to improve fairness in image classification with residual decomposition (abbreviated as Fair ResDecomp in the following) (Quadrianto et al., 2019b). The goal is to classify the attractiveness.

We summarize the evaluation results in Table 1. We can notice that baseline models introduce a large difference between the TPR in male and female, and are more likely to misclassify an attractive male as unattractive (much lower TPR for male than female). This indicates that the male is a unprivileged group for the attractiveness prediction. For the recent work (Quadrianto et al., 2019b), the model regularizes the image by projecting the residual to reduce the bias probability. We compare with this method as it use the same structure (VGG19) for their model. We can find that method in Quadrianto et al. (2019b) improves the absolute equal opportunity difference but sacrifices the performance of female TPR (the privileged group). In contrast, FAIAS model improves the prediction performance by increasing TPR of both male and female groups. Further, the results show that FAIAS narrows the gap between TPR of the two groups by introducing a large increase in the unprivileged group. This reveals that FAIAS model is able to enhance both fairness and prediction performance in the attractiveness classification.

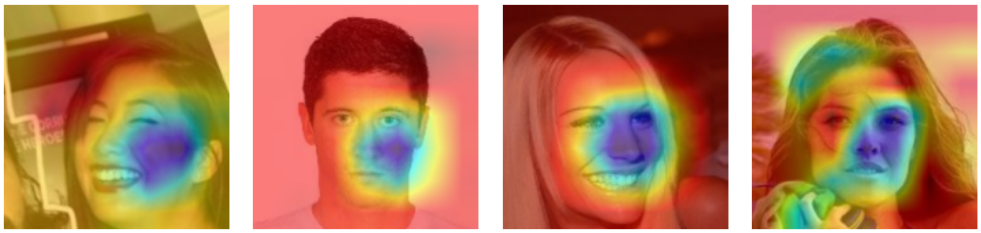
4.3.1 Qualitative Analysis of FAIAS

Image reconstruction To visualize the features that are selected by our FAIAS model, we built a deconvolution network based decoder to reconstruct an image from the 1,000 latent features and show the construction results in Figure 3. We notice that reconstructions without sensitive features makes gender-specific appearance to be blurry. Results show that one of the most sensitive-relevant feature for gender bias, *i.e.*, *length of hair* is modified. Short haired images get longer hair in the reconstruction with sensitive-irrelevant features. For example, the bold male in the first row and the female in the third row get longer hair in the sensitive-irrelevant reconstruction. In contrast, images with longer hair get hair length shortened or blurry around the hair region, *e.g.*, the female in the second row. This indicates that FAIAS successfully recognize *length of hair* as a sensitive-relevant feature that affects the group fairness. We empirically found that we need to remove more sensitive-relevant features from male data than female data. This is because the appearance related to male, *e.g.*, *beard* appears in a relatively subtle and local area than that of women's *e.g.*, *long hair*, we need

more features to filter it out. Removing the sensitive-relevant features makes more fair prediction in classification.

Visualization of Active Regions Adopting GradCAM (Selvaraju et al., 2017) visualization method, we visualize the heatmap in Figure 4. GradCAM visualizes the regions in the original input space that are important for predicting target feature of the CNN-based models. Feature with high probability output p from the selector g^θ implies that it is highly related to the performance and least related to sensitive information *i.e.*, sensitive-irrelevant. Therefore, pre-trained classifier (*e.g.*, ResNet50) should be also activated on 1) the target-related and sensitive-irrelevant region with features that have high p values; and 2) target-unrelated and sensitive-relevant region with features that have small p values in the original image space. In Figure 4a, we visualize GradCAM heatmap of features with high p values overlapped with original image. As expected, some specific facial areas are activated by the network to predict whether the person in the image is attractive or not. On the contrary, Figure 4b shows that the network is activated in not informative region or sensitive-relevant regions such as background and hair to predict features with small p values. This validates g^θ correctly masking out sensitive-relevant features.

(a) High p valued feature that expresses target label without sensitive relevant information.



(b) Low p valued feature that is sensitive relevant and could cause biased decision making.

Figure 4: Heatmap of GradCAM (Selvaraju et al., 2017) visualization of features with high and low p value. Face-related region is more activated with high p values and other region is focused with low p valued features.

5 Conclusion

In this paper, we propose a new adversarial network FAIAS for fairness. We formulate our model from both the data and model perspectives. Our FAIAS model consists of two components: a selector function and a prediction function, where the selector function is optimized on the data perspective to select features containing only sensitive-irrelevant information and the prediction function is optimized from the model perspective to minimize the marginal contribution of the sensitive feature and also improve the prediction performance. Extensive results validate that our FAIAS model achieves comparable or better results than all related methods w.r.t. both the prediction performance and fairness metrics.

Our FAIAS model is proposed for the supervised learning scenario, where we select the sensitive irrelevant features that maintain the discriminative power for classification tasks. In future work, we will explore the fair model in unsupervised learning and learn the set of meaningful and interpretable features that can preserve the data structure for unsupervised learning tasks like clustering, and eliminate the bias in the selected features.

REFERENCES

Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. One-network adversarial fairness. In *AAAI*, volume 33, pp. 2412–2420, 2019.

Sina Baharlouei, Maher Nouiehed, Ahmad Beirami, and Meisam Razaviyayn. Rényi fair inference. In *ICLR*, 2019.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Fair classification with noisy protected attributes: A framework with provable guarantees. In *ICML*, pp. 1349–1361. PMLR, 2021.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *NeurIPS*, 33:15088–15099, 2020.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *KDD*, pp. 259–268, 2015.

Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *ECCV*, pp. 793–811, 2018.

S. Mattu J. Angwin, J. Larson and L. Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Taeuk Jang, Feng Zheng, and Xiaoqian Wang. Constructing a fair classifier with generated fair data. In *AAAI*, volume 35, pp. 7908–7916, 2021.

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

Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pp. 202–207, 1996.

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, pp. 4066–4076, 2017.

Helen F Ladd. Evidence on discrimination in mortgage lending. *JEP*, 12(2):41–62, 1998.

Bo Li, Lijun Li, Ankang Sun, Chenhao Wang, and Yingfan Wang. Approximate group fairness for clustering. In *ICML*, pp. 6381–6391. PMLR, 2021.

Lydia T Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. *arXiv preprint arXiv:1803.04383*, 2018.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, pp. 3730–3738, 2015.

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

Jérémie Mary, Clément Calauzenes, and Noureddine El Karoui. Fairness-aware learning for continuous attributes and treatments. In *ICML*, pp. 4382–4391. PMLR, 2019.

Hussein Mouzannar, Mesrob I Ohannessian, and Nathan Srebro. From fair decision making to social equality. In *FAT*, pp. 359–368, 2019.

Jun Hyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: De-biasing classifier from biased classifier. In *NeurIPS*, 2020.

Luca Oneto, Michele Donini, Andreas Maurer, and Massimiliano Pontil. Learning fair and transferable representations. *arXiv preprint arXiv:1906.10673*, 2019.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NIPS*, pp. 5680–5689, 2017.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, June 2019a.

Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *CVPR*, pp. 8227–8236, 2019b.

Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM J Res Dev .*, 63(4/5):3–1, 2019.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *CVPR*, pp. 618–626, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, pp. 325–333, 2013.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *AIES*, pp. 335–340, 2018.