Fairness-aware Adversarial Perturbation Towards Bias Mitigation for Deployed Deep Models

Zhibo Wang[†,‡], Xiaowei Dong[†], Henry Xue, Zhifei Zhang, Weifeng Chiu, Tao Wei, Kui Ren[‡]
[†]School of Cyber Science and Engineering, Wuhan University, P. R. China
[‡]School of Cyber Science and Technology, Zhejiang University, P. R. China
Ant Group, P. R. China,Adobe Research, USA
{zhibowang, xwdong}@whu.edu.cn, gkn1fexxx@gmail.com, zzhang@adobe.com
{weifeng.qwf, lenx.wei}@antgroup.com, kuiren@zju.edu.cn

Abstract

Prioritizing fairness is of central importance in artificial intelligence (AI) systems, especially for those societal applications, e.g., hiring systems should recommend applicants equally from different demographic groups, and risk assessment systems must eliminate racism in criminal justice. Existing efforts towards the ethical development of AI systems have leveraged data science to mitigate biases in the training set or introduced fairness principles into the training process. For a deployed AI system, however, it may not allow for retraining or tuning in practice. By contrast, we propose a more flexible approach, i.e., fairness-aware adversarial perturbation (FAAP), which learns to perturb input data to blind deployed models on fairness-related features, e.g., gender and ethnicity. The key advantage is that FAAP does not modify deployed models in terms of parameters and structures. To achieve this, we design a discriminator to distinguish fairness-related attributes based on latent representations from deployed models. Meanwhile, a perturbation generator is trained against the discriminator, such that no fairness-related features could be extracted from perturbed inputs. Exhaustive experimental evaluation demonstrates the effectiveness and superior performance of the proposed FAAP. In addition, FAAP is validated on real-world commercial deployments (inaccessible to model parameters), which shows the transferability of FAAP, foreseeing the potential of black-box adaptation.

1. Introduction

AI systems have been widely deployed in many high-stakes applications, *e.g.*, face recognition [3,21], hiring process [14, 15], health care [13], *etc*. However, some existing AI systems are found to treat individuals unequally based on protected attributes, *e.g.*, ethnicity, gender, and nationality. Such biases are referred to as unfairness. For instance, Amazon realized that their automatic recruitment system presents skewness between male and female candidates [12], *i.e.*, male candidates are with higher probability to be hired as compared to female candidates. The COMPAS, which is an assessment system of recidivating risk, is found to have racial prejudice [7]. Such unfairness has been a subtle and ubiquitous nature of AI systems, thus it is non-trivial to mitigate the unfairness, ideally without touching the deployed models.

Many works have been proposed to mitigate unfairness/biases, which can be divided into three categories according to the stage de-biasing is applied, *i.e.*, preprocessing, in-processing, and post-processing. From the perspective of pre-processing, [8, 16, 17, 27, 31] mitigated biases in the training dataset, thus mitigating the bias during training the model. For the in-processing methods, [1, 19, 30] introduced fairness-related penalties into the learning process to train a fairer model. These methods need to retrain or fine-tune the target models, while these are unsuitable if the models are deployed without access to their training set. [6] proposed a boosting method to postprocess a deployed deep learning model to produce a new classifier that has equal accuracy in different people groups. However, [6] needs to replace the final classifier and cannot ensure statistical and predictive parity, *e.g.*, individuals in different groups are equally treated in prediction.

To the best of our knowledge, existing works are not suitable to improve fairness at the inference phase without changing the deep model. Therefore, it is imperative to propose a practical approach to mitigate the unfairness of deployed models without changing their parameters and structures. Since deep models tend to learn spurious correlations between protected attributes and target labels from training data, *e.g.*, the race may correlate to criminal risk, the key to mitigating unfairness is to break such correla-

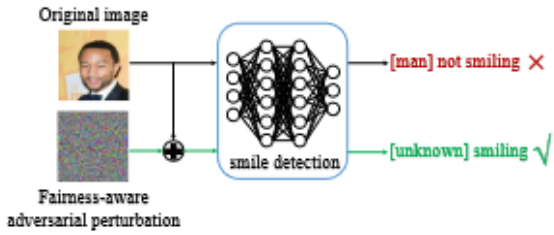

Figure 1. The illustration on a smile detection model. Original image is falsely recognized due to model unfairness, *i.e.*, tending to predict males as "not smiling". The fairness-aware adversarial perturbation generated by FAAP helps the input image to hide the protected attribute and get fair treatment.

tion. As we assume not modifying the model, the main challenge of achieving this goal is how to prevent the deployed model from extracting fairness-related information from inputs. Intuitively, the only thing we could modify is the input data during the inference stage of deployed models, *i.e.*, perturbing the inputs such that the model cannot recognize those protected attributes.

Based on the above idea, we propose the Fairness-Aware Adversarial Perturbation (FAAP), which learns to perturb input samples to blind deployed models on fairness-related features. As shown in Fig. 1, the deployed model can not distinguish the fairness-related feature (*e.g.*, gender) from the perturbed input image. Therefore, the predictions will not correlate to the protected attributes. The key idea is that perturbations can remap samples to tightly distribute along the decision hyperplane of protected attributes in the model latent space, making them difficult to be distinguished. To achieve this, we train a generator to produce adversarial perturbation. During the training process, a discriminator is trained to distinguish the protected attributes from the representations of the model, while the generator learns to deceive the discriminator, thus generating fairness-aware perturbation that can hide the information of protected attributes from the feature extraction process. Extensive experimental evaluation demonstrates the superior performance of the proposed FAAP and shows the potential in the black-box scenario, *i.e.*, mitigating unfairness of models without access to their parameters.

In summary, the main contributions of this paper are in three-folds:

- We give the first attempt to mitigate the unfairness from deployed deep models without changing their parameters and structures. This pushes the fairness research towards a more practical scenario.

- We propose the fairness-aware adversarial perturbation (FAAP), which designs a discriminator to distinguish fairness-related attributes based on latent representations from deployed models. Meanwhile, a generator is trained adversarially to perturb input data to prevent the deployed models from extracting fairness-related features. This design effectively decorrelates fairness-related/protected attributes from predictions.

- Extensive experiments demonstrate the superior performance of the proposed FAAP. In addition, evaluation on real-world commercial APIs shows the transferability of FAAP, which indicates the potential of further exploring our method in the black-box scenario.

2. Related work

This section overviews related works on unfairness mitigation that could be roughly divided according to targeting stages, *i.e.*, pre-processing (data pre-processing before training), in-processing (penalty design during training), and post-processing (prediction adjustment after training).

Pre-processing methods [16, 17, 31] aim to mitigate biases in the training dataset, *i.e.*, fairer training sets would train fairer models. Many methods have been proposed to de-bias training sets by fair data representation transformation or data distribution augmentation. Quadrianto et al. [16] used data-to-data translation to find middle-ground representation for different gender groups in training data, thus the model will not learn the tendency of gender. Ramaswamy et al. [17] generated paired training data to balance protected attributes, which would remove spurious correlation between target label and protected attributes. Zhang et al. [31] proposed to generate adversarial examples to supplement the training dataset, balancing the data distribution over different protected attributes.

In-processing approaches [1, 18, 19, 29, 30] introduce fairness principles into the training process, *i.e.*, training models by specifially designed fairness penalties/constraints or adverasial mechinsm. Zafar et al. [29] proposed to maximize accuracy under disparate impact constraints to improve fairness in machine learning. Brian et al. [1] and Zhang et al. [30] enforced the model to produce fair outputs with adversarial training techniques by maximizing accuracy while minimizing the ability of a discriminator to predict the protected attribute. Yuji Roh et al. [18] provided a mutual information-based interpretation of an existing adversarial training-based method for improving the disparate impact and equalized odds. Sarhan et al. [19] imposed orthogonality and disentanglement constraints on the representation and forced the representation to be agnostic to protected information by entropy maximization, then the following classifier can make fair predictions based on learned representation. This line of research aims at getting a fairer model by explicitly changing the training procedure. Different from this line of work, our method is applied af-

ter the training process and can improve fairness without changing the deployed model.

Post-processing works [6, 10] tend to adjust model predictions according to certain fairness criteria. Lohia *et al.* [10] proposed a post-processing algorithm that helps a model meet both individual and group fairness criteria on tabular data by detecting biases from model outputs and correspondingly editing protected attributes to adjust model predictions. However, this method needs to change protected attributes at the test time which is hard for computer vision applications. Michael *et al.* [6] proposed a method that can post-process a pre-trained deep learning model to create a new classifier, which has equal accuracy for people with different protected attributes. However, [6] needs to replace the final classifier, and equal sub-group accuracy can not ensure people in different groups have equal chance to get favorable predictions, *e.g.*, unequal false positive rate and false negative rate. We borrow ideas from this line of research, but we improve fairness from the data side, instead of manipulating the model or its prediction.

3. Preliminaries

3.1. Model fairness

In this paper, we focus on visual classification models because of exhaustive academic efforts on them, as well as their broad industrial applications. Moreover, it is important to achieve equal treatment for people with different protected attributes, *e.g.*, nationality, gender, and ethnicity. Therefore, demographic parity [28] and equalized odds [4] are adopted to measure model fairness.

In a binary classification task, *e.g.*, criminal prediction, suppose target label $y \in \mathcal{Y} = \{-1, 1\}$, protected attribute $z \in \mathcal{Z} = \{-1, 1\}$, where $y = 1$ is in favourable class (*e.g.*, lower criminal tendency) and $z = 1$ is in privileged group (*e.g.*, Caucasian).

Definition 1 (Demographic Parity). If the value of z does not influence assigning a sample to the positive class, *i.e.* model prediction $\hat{y} = 1z$, then the classifier satisfies demographic parity:

$$P(\hat{y} = 1|z = -1) = P(\hat{y} = 1|z = 1) \tag{1}$$

If a model satisfies demographic parity, samples in both the privileged and unprivileged groups have the same probability to be predicted as positive. **Definition 2** (Equalized Odds). If the value of z can not influence the positive outcome for samples given y, *i.e.* $\hat{y} = 1z \mid y$, then the classifier satisfies equalized odds:

$$P(\hat{y} = 1|y, z = -1) = P(\hat{y} = 1|y, z = 1), y = \{-1, 1\} \tag{2}$$

Equalized odds means that positive output is statistically independent to the protected attribute given the target label. Samples in both the privileged and unprivileged groups have the same false positive rate and false negative rate.

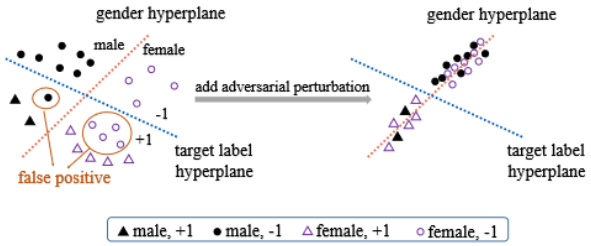

Figure 2. The basic idea of the proposed fairness-aware adversarial perturbation (FAAP). Gender bias exists in the left part, *i.e.*, the false positive rate of females is much higher than males. Without adjusting the decision hyperplanes of the deployed model, FAAP perturbs samples to decorrelate the target label and gender in latent space. In the right part, perturbed samples tightly distribute along the gender hyperplane, meanwhile, preserving the distinguishability along the target label hyperplane.

3.2. Adversarial examples

Recent studies show that deep learning models are vulnerable to adversarial examples [24]. Given a classification model $C(x)$, the goal of adversarial attacks is to find a small perturbation to generate an adversarial example x', to mislead classifier C. More specifically, there are two kinds of adversarial example attacks. For an input x with ground label y, targeted attack will let $C(x') = y'$ where $y' \neq y$ is a label specified by the attacker. On the contrary, in an untargeted attack, an attacker will mislead the classification model as $C(x') \neq y$. Typically, the l_p norm of the perturbation should be less than ϵ, *i.e.* $\|x - x'\|_p \leq \epsilon$. Many methods have been proposed to generate adversarial examples, such as PGD [11], CW [2] and GANs based method [26].

4. Fairness-aware adversarial perturbation

In this paper, we propose fairness-aware adversarial perturbation (FAAP) to mitigate unfairness born with deep models. This section will overview the proposed FAAP and detail the design of network and loss functions. Finally, we will further discuss the training strategy of FAAP.

4.1. Overview of FAAP

The unfairness could be caused by the bias in training sets (*e.g.*, skewed data distribution) and/or loose constraints in the training process. All of these lead to spurious correlations between target labels and protected attributes, *e.g.*, gender and ethnicity. In a dataset, females may have much more positive samples than males. As illustrated in Fig. 2 (left), the model learns such spurious gender correlation so that the false positive rate of the target label varies significantly for males and females. Therefore, the key of mitigating unfairness is to break spurious correlations between target labels and protected attributes.

In this paper, we propose the fairness-aware adversarial perturbation (FAAP) to mitigate model unfairness by hiding the information of protected attributes from the feature extraction process, so that the model would not correlate predictions with protected attributes. The basic idea is to leverage adversarial perturbation to remap the original samples to the position close to the decision hyperplane of the protected attribute in the latent space (*e.g.*, on the surface of gender hyperplane in the figure). Note that the distinguishability of these perturbed samples along the original target label decision hyperplanes should be preserved, as shown in Fig. 2 (right). In this way, the deployed model can not distinguish the protected attributes from the perturbed images during feature extraction. Therefore, the protected attribute would become uncorrelated to the target label. In other words, the model would fairly treat samples with different protected attributes.

The pipeline of FAAP is overviewed in Fig. 3, where there are two learnable components: 1) the generator that perturbs samples to regulate their distribution in the latent space, and 2) the discriminator that distinguishes the protected attribute. The deployed model is assumed to be a classification model that could be split into a feature extractor (*i.e.*, from image to latent space) and a label predictor (*i.e.*, from latent space to final label). Please note that we freeze the parameters of the deployed model. Sharing the spirit of general GANs during the training process, the discriminator is trained to distinguish the protected attribute from representations of the model, while the generator learns to fail the discriminator, thus synthesizing fairness-aware perturbation that reduces the information of the protected attribute in the latent representations.

4.2. Loss Functions

In this part, we detail the loss functions of the abovementioned FAAP. As illustrated in Fig. 3, we assume a classification model that is divided into a feature extractor g and a label predictor f . Given an input x, whose true label is y, the predicted label $\hat{y} = f(g(x))$. The generator G generates perturbation based on input x to obtain perturbed input $\hat{x} = x + G(x)$ subject to $\|\hat{x} - x\|_\infty \leq \epsilon$, and the discriminator D is applied on the latent representations $\hat{r} = g(\hat{x})$ to distinguish a certain protected attribute z.

Loss function of D: Intuitively, with a deployed model, the unfairness is mainly caused by the feature extraction process which tends to correlate the protected attribute to those predicted in the target label, *i.e.*, carrying distinguishable information from the protected attribute to the latent representations. Thus, the label predictor would utilize that distinguishable sensitive information to bias its final prediction. Based on the above hypothesis, we first need to let the discriminator D be aware of the protected attribute z in the latent representation, *i.e.*, perfectly predicting z. With such awareness, the generator$G$ is able to adversarially perturb inputs towards hiding the protected attribute in the latent representation. Therefore, the discriminator loss can be expressed as

$$\mathcal{L}_D = \mathcal{J}(D(g(\hat{x}), z)), \tag{3}$$

where $\mathcal{J}(,)$ denotes cross-entropy, $\hat{x}$ is the perturbed input, and z indicates the true label of the protected attribute.

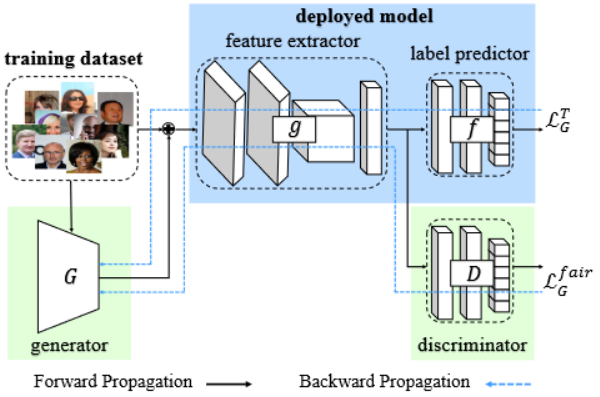

Figure 3. Overview of the proposed FAAP, which consists of two learnable components, *i.e.*, a generator for learning fairness-aware perturbation and a discriminator for distinguishing the protected attribute.

Loss functions of G: By contrast, the generator G aims to fail D, and an intuitive solution is to maximize $\mathcal{L}_D$ on perturbed samples $\hat{x}$. However, this will push the latent representations towards the opposite side of the protected attribute, *e.g.*, female flips to male. Therefore, we further let D make random guess on the representation of $\hat{x}$, increasing entropy of the protected attribute on perturbed samples. The fairness loss can be written as

$$\mathcal{L}_G^{fair} = -\mathcal{L}_D - \alpha\mathcal{H}(D(g(\hat{x}))), \tag{4}$$

where $\mathcal{H}()$ calculates the entropy, > 0 is a relatively small value controls the regularization of entropy loss. Besides $\mathcal{L}_G^{fair}$ that encourages fairness-aware perturbation, at the same time we need to preserve the model performance on the target label. The target label prediction loss is

$$\mathcal{L}_G^T = \mathcal{J}(f(g(\hat{x})), y). \tag{5}$$

Above all, the total loss for generator G in FAAP consists of $\mathcal{L}_G^{fair}$ and $\mathcal{L}_G^T$, which can be summarized as the following

$$\mathcal{L}_G = \mathcal{L}_G^{fair} + \beta\mathcal{L}_G^T, \tag{6}$$

where > 0 balances the performance of target label prediction and fairness.

4.3. Training of FAAP

Based on Eq. 3 and Eq. 6, in the training phase of FAAP, the generator and the discriminator are optimized alternatively. The generator G plays a min-max game with D where D maximizes the ability to predict protected attribute z while$G$ tries to minimize its ability. At the same time, G tries to let f still recognize the right target label for perturbed input data. Therefore, the objectives of FAAP can be formulated as follows:

$$\begin{aligned} \arg\max_{G}\min_{D} \mathcal{J}(D(\hat{r}), z) + \alpha\mathcal{H}(D(\hat{r})) - \beta\mathcal{L}_G^T, \\ s.t. \hat{r} = g(\hat{x}) = g(x + G(x)), \\ \|\hat{x} - x\|_\infty \leq \epsilon \end{aligned} \tag{7}$$

where D and G are updated alternatively during the optimization. Please note that is set to be 0 during updating D to allow D focus on distinguishing protected attributes. More detailed training algorithm of FAAP can be found in Algorithm 1.

Algorithm 1 Training of FAAP

Input: Feature extractor g and label prediction f of a deployed model, loss weights α and β, learning rates η_D and η_G, batch size n, maximum iteration N , and maximum perturbation magnitude ϵ. The training images x, true labels y, and protected attribute labels z.

Output: Generator G

Initialize the generator G and discriminator D.

for $i = 1, \cdots, N$ **do**

Get a batch of n inputs x_i and labels y_i and z_i

Get perturbed inputs $\hat{x}_i = x_i + G(x_i)$

Clip $\hat{x}_i$to meet $\|\hat{x}_i - x_i\|_\infty \leq \epsilon$

Get model feature $\hat{r}_i = g(\hat{x}_i)$

Calculate discriminator loss

$$\mathcal{L}_D = \frac{1}{n}\sum_{i=1}^{n} \mathcal{J}\left(D\left(\hat{r}_i\right), z_i\right)$$

Update D ¸$D_D\nabla_D\mathcal{L}_D$

Calculate fairness loss

$$\mathcal{L}_G^{fair} = -\frac{1}{n}\sum_{i=1}^{n}[\mathcal{J}\left(D\left(\hat{r}_i\right), z_i\right) + \alpha\mathcal{H}\left(D(\hat{r}_i)\right)]$$

Calculate target label prediction loss

$$\mathcal{L}_G^T = \frac{1}{n}\sum_{i=1}^{n} \mathcal{J}(f(\hat{r}_i), y_i)$$

Get total loss of G, $\mathcal{L}_G = \mathcal{L}_G^{fair} + \beta\mathcal{L}_G^T$

Update $G \leftarrow G_G\nabla_G\mathcal{L}_G$

end for

5. Experimental Evaluation

In this section, we first describe our experimental setup (Section 5.1) . Then, we quantitatively (Section 5.2) and qualitatively (Section 5.3) evaluate the proposed FAAP on different deployed models. Finally, we investigate the transferability of adversarial perturbation generated by FAAP on real-world commercial systems (Section 5.4).

5.1. Experimental Setup

Datasets. We adopt two face datasets in our evaluation, *i.e.*, Celeba[1] and LFW2[2] which carry those commonly protected attributes like gender. The CelebA dataset consists of 202,599 images along with 40 attributes per image, and LFW has 13,244 images along with 73 attributes per image. We take gender as the protected attribute to measure the fairness of model prediction for target labels. In CelebA, the *Smiling, Attractive,* and *Blond Hair* are chosen as target labels. Similarly, *Smiling, Wavy Hair,* and *Young* are selected as the target labels in LFW. We randomly divide the original training set of CelebA into two equal parts for training the deployed model and our FAAP, respectively. For LFW, it is randomly split to get a 6k training set, a 3.6k validation set, and the rest as the testing set. For convenience, all the images are resized to 224×224.

Training details. To investigate the effectiveness of FAAP in de-biasing models with different extent of unfairness, we train three kinds of models as the deployed models, *i.e.*, normal training model, fair training model, and unfair training model. The normal training model is trained normally by minimizing the loss on target label. This kind of model will learn the intrinsic bias in the training dataset, *e.g.*, the correlation between *Smiling* and *Male*. For the fair training model, we adopt adversarial training techniques [30] to train a fair model, which maximizes the classifier's ability to predict the target label, while minimizing the discriminator's ability to predict the protected attribute. This kind of model has better fairness than the normal training model. To valid our method against more unfair models, which could be from malicious manipulations, *e.g.*, data poison attack [23] and malicious training [22], we apply two methods to amplify unfairness in deployed models. One is to flip labels (denoted as **LF**), *e.g.*, randomly flipping the target labels. The other is to reverse the gradients of the discriminator in adversarial fair training (denoted as **RG**). These manipulations can strengthen the spurious correlation between target labels and gender.

For all deployed models, we use ResNet-18 [5] as the base architecture. We train all of these models for 30 epochs with a batch size of 64 using Adam optimizer with a learn-

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

[2] http://vis-www.cs.umass.edu/lfw/, attribute annotations are provided in [9]

ing rate of 5e-4. Once the training is finished, we fix the parameters of the deployed models. The generator G in FAAP has a similar architecture with [26]. Discriminator D is connected to the last convolution layer of the feature extractor. To mitigate unfairness without harming the visual quality of a specific image, we set the maximum perturbation magnitude ϵ to 0.05.

Evaluation metrics. For fairness evaluation, we use the difference in demographic parity (DP) and difference in equalized odds (DEO) to evaluate model fairness. Meanwhile, the accuracy (ACC) of predicting target labels will also be reported. The DP calculates the absolute difference between the acceptance rates for each gender. A larger DP indicates that samples in the privileged group have higher chances to be predicted as positive than those in the unprivileged group. Ideally, the DP is equal to zero. By contrast, the DEO computes the absolute difference between the false negative rates and the false positive rates for each gender. A larger DEO means that samples in the privileged group have higher false positive rates and/or lower false negative rates than those in the unprivileged group. Therefore, the lower DEO the better.

5.2. Quantitative Evaluation

Tables 1a to 1c show quantitative results of deployed models before and after embedded with the proposed FAAP on CelebA. We evaluate with three different target labels named *Smiling*, *Attractive* and *Blond Hair* respectively with the protected attribute *Male* ("+1" in *Male* means male and "-1" means female). Besides, we use three different kinds of models for each target label. As shown in Table 1, there exists gender bias in normal training models, *e.g.*, DP and DEO are larger than 0.5 when the target label is *Attractive*. Fair training can get a fairer model by incorporating adversarial fairness techniques into training procedures. For instance, we can see in Table 1c, fair training models have much lower DP (reduction from 0.5023 to 0.2745) and DEO (reduction from 0.5683 to 0.0724) than normal training models with a small drop in ACC (79.56% comparing to 82.43%). In contrast, unfair training amplifies gender bias and these models (LF and RG) show much more unfairness. For example, as shown in Table 1a, DP and DEO increase to about 0.25 with relatively high ACC (91.48%, 91.76% comparing to 92.61% of normal training model).

We evaluate our method FAAP on the above deployed models. Not surprisingly, FAAP can improve fairness and maintain target label prediction accuracy for a deployed model. From Tables 1a to 1c, we have the following observations. **(1) Normal training model.** For a normal training model, FAAP can improve its fairness and keep target label accuracy. We can see our method improves DP and DEO by 0.2319, 0.5062 respectively with accuracy loss less than 0.03 in Table 1b. **(2) Fair training model.** When adversarial fair training techniques are applied to the model training phase, our method can further improve the fairness of these models with slight accuracy drop, e.g., in Table 1c, FAAP still improve fairness (0.0083 and 0.0544 reduction in DP and DEO respectively) with slight accuracy degradation (from 94.41% to 94.05%). **(3) Unfair training model.** For an unfair training model, FAAP can significantly improve its fairness with slight accuracy degradation. For instance, in Table 1a, FAAP can decrease DEO to about 0.04, maintaining ACC above 91%. (4) **Comparison between Normal training+FAAP and Fair training.** It is better to take model fairness into consideration in the training phase. However, in Table 1 we can see that a deployed normal training model embedded with FAAP can get comparable fairness performance as a fair training model (*e.g.*, FAAP has even better DP and DEO in some cases) with almost the same accuracy (*i.e.*, the difference in ACC is less than 0.3% in most of the cases). For a deployed model, our method works after the training process without changing the model as compared to the fair training that needs to retrain or finetune the model. Similar observation can be observed in Tables 1d to 1f on LFW dataset as well.

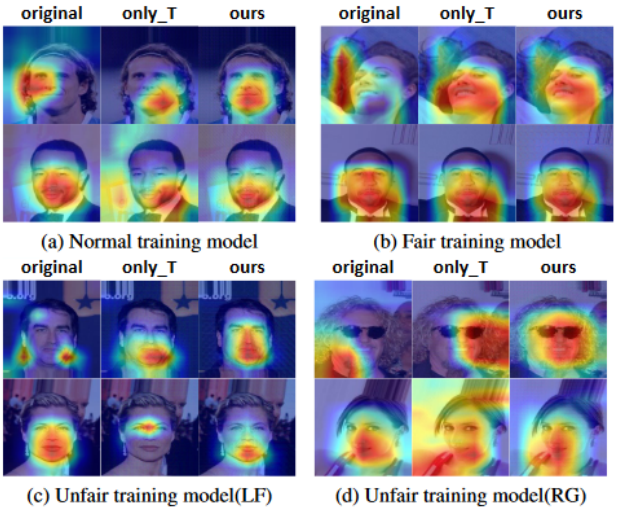

Figure 4. Grad-CAM results for three different models when the target label is *Smiling* in CelebA. "orginial" denotes raw data, "only$_T$" $denotesimagesperturbedbyG$ which is only optimized on $\mathcal{L}_G^T$ without $\mathcal{L}_G^{fair}$, "ours" denotes images perturbed with fairness-aware adversarial perturbation generated by G optimized on $\mathcal{L}_G$. (Better viewed in color)

5.3. Qualitative Evaluation

In this part, we further provide results of model explanation approaches Grad-CAM [20] and T-SNE [25] to better illustrate the effectiveness of our method.

Grad-CAM is a model explanation method by visualizing the regions of input data that are important for predic-

Smiling	ACC↑	DP↓	DEO↓
Normal training	92.61%	0.1748	0.0774
Normal training+FAAP	92.46%	0.1426	0.0327
Fair training	92.55%	0.1275	0.0308
Fair training+FAAP	92.49%	0.1326	0.0281
Unfair training (LF)	91.48%	0.2638	0.2737
Unfair training (LF)+FAAP	91.87%	0.1268	0.0381
Unfair training (RG)	91.76%	0.2439	0.2306
Unfair training (RG)+FAAP	91.78%	0.1321	0.0369

(a) Results on CelebA when the target label is *Smiling*

Attractive	ACC↑	DP↓	DEO↓
Normal training	82.43%	0.5023	0.5683
Normal training+FAAP	79.73%	0.2704	0.0621
Fair training	79.56%	0.2745	0.0724
Fair training+FAAP	79.31%	0.2244	0.0434
Unfair training (LF)	81.06 %	0.5566	0.7752
Unfair training (LF)+FAAP	79.08%	0.2890	0.1179
Unfair training (RG)	82.24%	0.5547	0.7217
Unfair training (RG)+FAAP	79.37%	0.2550	0.0539

(b) Results on CelebA when the target label is *Attractive*

Blond_Hair	ACC↑	DP↓	DEO↓
Normal training	95.63%	0.1787	0.5299
Normal training+FAAP	94.52%	0.1345	0.1013
Fair training	94.41%	0.1319	0.1587
Fair training+FAAP	94.05%	0.1236	0.1043
Unfair training (LF)	95.41%	0.1733	0.6728
Unfair training (LF)+FAAP	94.49%	0.1449	0.1321
Unfair training (RG)	95.66%	0.2041	0.6200
Unfair training (RG)+FAAP	94.26%	0.1305	0.1209

(c) Results on CelebA when the target label is *Blond_Hair*

Smiling	ACC↑	DP↓	DEO↓
Normal training	90.42%	0.3353	0.1472
Normal training+FAAP	89.80%	0.2910	0.0534
Fair training	90.08%	0.2704	0.0318
Fair training+FAAP	88.75%	0.2646	0.0136
Unfair training (LF)	89.23%	0.3678	0.2340
Unfair training (LF)+FAAP	88.10%	0.3026	0.1076
Unfair training (RG)	90.14%	0.3674	0.2257
Unfair training (RG)+FAAP	89.15%	0.2969	0.0782

(d) Results on LFW when the target label is *Smiling*

Wavy_Hair	ACC↑	DP↓	DEO↓
Normal training	78.69%	0.1707	0.1554
Normal training+FAAP	78.04%	0.1241	0.0651
Fair training	77.98%	0.1337	0.0800
Fair training+FAAP	77.67%	0.1094	0.0595
Unfair training (LF)	78.35%	0.2383	0.2919
Unfair training (LF)+FAAP	77.19%	77.19	0.1734
Unfair training (RG)	77.59%	0.2724	0.3692
Unfair training (RG)+FAAP	77.10%	0.2128	0.2508

(e) Results on LFW when the target label is *Wavy_Hair*

Young	ACC↑	DP↓	DEO↓
Normal training	83.81%	0.3511	0.5516
Normal training+FAAP	81.34%	0.2281	0.2914
Fair training	83.86%	0.2500	0.2870
Fair training+FAAP	80.71%	0.1515	0.1141
Unfair training (LF)	83.04%	0.4813	0.8196
Unfair training (LF)+FAAP	80.40%	0.2550	0.3786
Unfair training (RG)	83.72%	0.5002	0.8377
Unfair training (RG)+FAAP	82.30%	0.1970	0.3048

(f) Results on LFW when the target label is *Young*

Table 1. Results of deployed models before and after embedded with the proposed FAAP on CelebA (Tables 1a to 1c) and LFW (Tables 1d to 1f. For fairness criterion DP and DEO, the lower the fairer. For accuracy ACC, the higher the better.

tions [20]. We visualize a subset of test images that were originally false predicted by the deployed model but have been successfully recognized after perturbation in Fig. 4. For each deployed model, we provide explanations on raw data, images perturbed by G trained on $\mathcal{L}_G^T$ without $\mathcal{L}_G^{fair}$ and images perturbed by G optimized on $\mathcal{L}_G$. (1) **Normal training model.** As shown in Fig. 4(a), for a normal training model, our adversarial perturbation can help the model focus on the right area (mouth) and make correct predictions. The red area of images in "only T" deviates little from the mouth. (2) **Fair training model.** Since this kind of model has less gender bias than other models, as shown in Fig. 4(b), G optimized towards improving target label accuracy can get similar heat-maps as "ours". Both of them can help the deployed model focus on the right area. (3) **Unfair training model.** Unfair training models have larger gender tendency, thus we can see that perturbation in "only_T" will let the model make correct predictions but mislead the model to focus on the unrelated area (*e.g.*, eyes in Fig. 4(c), hair in Fig. 4(d)). In contrast, our method helps the model focus on the right area and make right predictions.

T-SNE is a method to visualize high-dimensional data from a low dimension view. To better demonstrate that our method can hide sensitive information for images by remapping them close to the protected attribute decision hyperplane while maintaining the distance to the target label decision hyperplane in latent space of the deployed model, we utilize T-SNE to get low-dimensional embedding of data feature representation. More specifically, we extract feature vectors of these images with/without adversarial perturbation and visualize them in a two-dimension diagram with T-SNE. **(1) Normal training model.** From Fig. 5(a) and Fig. 5(b) we can see that for a normal training model, samples with different target labels for smiling and attractive classification are linearly separable in latent space, meanwhile, samples with different gender before and after perturbation are mixed. In Fig. 5(c), even the feature representations of samples (yellow and purple points) in normal train-

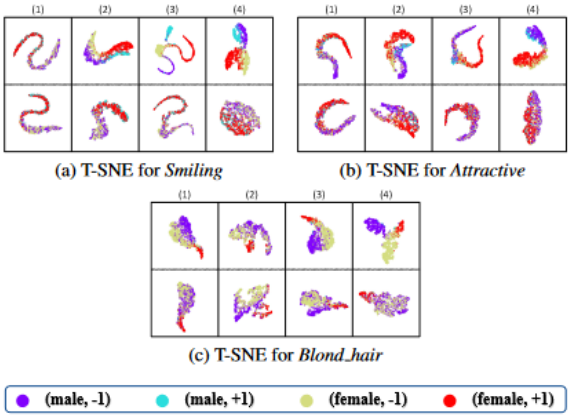

Figure 5. T-SNE results for three different models on *Smiling*, *Attractive* and *Blond_Hair* in CelebA. The upper row shows the results of the raw data and the bottom row shows the results of images perturbed with FAAP. In each sub-figure, the feature representation in column (1) is extracted from a normal training model, and column (2) from a fair training model, while column (3) from a LF model, column (4) from a RG model. (Better viewed in color)

ing model are linearly separated by the protected attribute hyperplane when the target label is Blond Hair, but FAAP can still effectively hide such sensitive information in latent feature space for samples. **(2) Fair training model.** Adversarial fair training can improve fairness, however, may slightly separate samples with different protected attributes (as shown in column (2) in Fig. 5(a) and Fig. 5(b)). In such a situation, our FAAP can make these samples become closer. **(3) Unfair training model.** In unfair training models, feature representations of original images with different protected attributes are almost linear separable on the protected attribute hyperplane ((male, -1) with (female, -1), (male, +1) with (female, +1)). Once perturbed with adversarial perturbation generated by FAAP, samples with different gender become almost indistinguishable and mixed but they are well separable on the target label hyperplane.

5.4. Transferability of FAAP

To demonstrate the transferability of adversarial perturbation generated by FAAP, we evaluate them on commercial face analyze APIs. At first, we investigate model fairness of these APIs in predicting "smiling". We upload testing dataset (about 20k images) from CelebA dataset to toady's commercial APIs, including Alibaba3 and Baidu4. For Alibaba's face analyze API, it returns binary results in which "0" means "not smiling" and "1" means "smiling". For Baidu's face analyze API, it returns three categories named "none", "smile" and "laugh". We assume "none" means not smiling and others mean smiling. We find these APIs have some extent of unfairness, i.e., DEO of them are about 0.1.

Alibaba's API	**ACC↑**	**DP↓**	**DEO↓**
original images	90.20%	0.1768	0.0952
after perturbation	89.94%	0.1475	0.0368

(a) Results on Alibaba's face analyze API

Baidu's API	**ACC↑**	**DP↓**	**DEO↓**
original images	90.47%	0.1817	0.1035
after perturbation	87.58%	0.1406	0.0387

(b) Results on Baidu's face analyze API

Table 2. Performance on commercial face analyze APIs.

Since this is a totally black-box scenario, we know nothing about the models behind these APIs. We try to train the generator with model ensemble techniques, taking the normal training model and the fair training model in Section 5.2 as surrogate models. Then we upload the perturbed images to these APIs and record results. Table 2 shows the results of these face analyze APIs on original and perturbed images. From Table 2a, we can see that FAAP improves DP by 0.0293 and decreases DEO to 0.0368 with only 0.0026 degradation in accuracy. Likewise, Table 2b shows 0.0411, 0.0648 improvement in DP and DEO while 0.0289 degradation in accuracy for Baidu. These results show the transferability of FAAP and the potential usage of FAAP in black-box scenarios.

6. Conclusion

This paper introduced the Fairness-Aware Adversarial Perturbation (FAAP) to mitigate unfairness in deployed models. More specifically, FAAP learns to perturb inputs, instead of changing the deployed models as the SOTA works, to disable deployed models from recognizing fairness-related features. To achieve this, we employed a discriminator to distinguish fairness-related attributes from latent representations of deployed models. Meanwhile, a generator was trained adversarially to deceive the discriminator, thus synthesizing fairness-aware perturbation that can hide the information of protected attributes. Extensive experiments demonstrated that FAAP can effectively mitigate unfairness, *e.g.*, improve DP and DEO by 27.5% and 66.1% respectively with only 1.5% accuracy degradation on average for normal training models.

In addition, evaluation on real-world commercial APIs showed significantly 19.5% and 61.9% improvement in DP and DEO with less than 1.7% degradation in accuracy, which indicates the potential usage of the proposed FAAP in the black-box scenario. However, the black-box exploration is a side product of our current design. Since we assume to access the deployed models although do not modify them, it is still impractical for certain real cases like those commercial APIs. Therefore, we are considering future works on the black-box setting with more specific designs.

Acknowledgments

This research was supported in part by National Key R&D Program of China (Grant No. 2020AAA0107705), National Natural Science Foundation of China (Grants No. 62122066, U20A20182, and 61872274,) Key R&D Program of Zhejiang (Grant No. 2022C01018), CCF-AFSG RF20200008, and the Fundamental Research Funds for the Central Universities (Grant No. 2021QNA5016).

References

[1] 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. 1, 2

[2] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 3

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[4] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*, pages 3323–3331, 2016. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5

[6] Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 247–254, 2019. 1, 3

[7] Lauren Kirchner, Jeff Larson, Surya Mattu, and Julia Angwin. How we analyzed the compas recidivism algorithm. https://www.propublica.org/article/how-we-analyzedthe-compas-recidivism-algorithm, 2016. Online; Accessed June 18, 2021. 1

[8] Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9572–9581, 2019. 1

[9] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. 5

[10] Pranay K Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R Varshney, and Ruchir Puri. Bias mitigation post-processing for individual and group fairness. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2847–2851, 2019. 3

[11] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 3

[12] David Meyer. Amazon reportedly killed an ai recruitment system because it couldn't stop the tool from discriminating against women. https://fortune.com/2018/10/10/amazon-ai-recruitment-bias-women-sexist/, 2018. Online, Accessed June 1, 2021. 1

[13] Beau Norgeot, Benjamin S Glicksberg, and Atul J Butte. A call for deep-learning healthcare. *Nature Medicine*, 25(1):14–15, January 2019. 1

[14] Alejandro Peña, Ignacio Serna, Aythami Morales, and Julian Fierrez. Bias in multimodal ai: Testbed for fair automatic recruitment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 129–137, 2020. 1

[15] Chuan Qin, Hengshu Zhu, Tong Xu, Chen Zhu, Liang Jiang, Enhong Chen, and Hui Xiong. Enhancing person-job fit for talent recruitment: An ability-aware neural network approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 25–34, 2018. 1

[16] Novi Quadrianto, Viktoriia Sharmanska, and Oliver Thomas. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2

[17] Vikram V. Ramaswamy, Sunnie S. Y. Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[18] Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning (ICML)*, pages 8147–8157, 2020. 2

[19] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision (ECCV)*, pages 746–761, 2020. 1, 2

[20] 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 *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. 6, 7

[21] Yichun Shi, Xiang Yu, Kihyuk Sohn, Manmohan Chandraker, and Anil K. Jain. Towards universal representation learning for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[22] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186. Association for Computing Machinery, 2020. 5

[23] David Solans, Battista Biggio, and Carlos Castillo. Poisoning attacks on algorithmic fairness. In Frank Hutter, Kristian Kersting, Jefrey Lijffijt, and Isabel Valera, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 162–177, 2021. 5

[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 3

[25] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11), 2008. 6

[26] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 3, 6

[27] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pages 570–575, 2018. 1

[28] 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 the 26th International Conference on World Wide Web (WWW)*, pages 1171–1180, 2017. 3

[29] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970, 2017. 2

[30] 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*, pages 335–340, 2018. 1, 2, 5

[31] Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*, pages 4346–4354, 2020. 1, 2