SUPPLEMENTARY TO “ADVERSARIAL FAIRNESS NETWORK”

Anonymous authors
Paper under double-blind review

1 DERIVATIVE OF LOSS TERMS

First we look into the derivative of $l_{sen}(\theta, \phi)$ w.r.t. θ,

$$\begin{aligned}\nabla_\theta l_{sen}(\theta,\phi) &= \nabla_\theta \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\Big[-\sum_{l=1}^{c} f_l^{\phi}(\mathbf{x},\mathbf{s})\log f_l^{\phi}(\mathbf{x},\mathbf{s}\cup\{k\})\Big] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},\mathbf{y})\Big(\sum_{\mathbf{s}\in\{0,1\}^d}\pi_\theta(\mathbf{x},\mathbf{s})\frac{\nabla_\theta\pi_\theta(\mathbf{x},\mathbf{s})}{\pi_\theta(\mathbf{x},\mathbf{s})}\hat{l}_{sen}(\mathbf{x},\mathbf{s})\Big)dxdy \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\Big[\hat{l}_{sen}(\mathbf{x},\mathbf{s})\nabla_\theta\log\pi_\theta(\mathbf{x},\mathbf{s})\Big].\end{aligned}$$

Similarly, the derivative of $l_{pred}(\theta, \phi)$ w.r.t. θ can be written as

$$\begin{aligned}\nabla_\theta l_{pred}(\theta,\phi) &= \nabla_\theta \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\big[-\sum_{l=1}^{c} y_l\frac{\nabla_\phi f_l^{\phi}(\mathbf{x},\mathbf{s})}{f_l^{\phi}(\mathbf{x},\mathbf{s})}\big] \\ &= \int_{\mathcal{X}\times\mathcal{Y}} p(\mathbf{x},\mathbf{y})\Big(\sum_{\mathbf{s}\in\{0,1\}^d}\pi_\theta(\mathbf{x},\mathbf{s})\frac{\nabla_\theta\pi_\theta(\mathbf{x},\mathbf{s})}{\pi_\theta(\mathbf{x},\mathbf{s})}\hat{l}_{pred}(\mathbf{x},\mathbf{s})\Big)dxdy \\ &= \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim(\mathcal{X}\times\mathcal{Y})}\mathbb{E}_{\mathbf{s}\sim\pi_\theta(\mathbf{x},\cdot)}\Big[\hat{l}_{pred}(\mathbf{x},\mathbf{s})\nabla_\theta\log\pi_\theta(\mathbf{x},\mathbf{s})\Big].\end{aligned}$$

Thus, to learn θ that maximizes l_{sen} and minimize l_{pred}, we update θ as

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

2 EXPERIMENTAL DETAILS

2.1 COMPARING METHODS

- **Adversarial de-biasing model** (abbreviated as Adv_Deb in the comparison)Zhang et al. (2018): an in-processing model that proposes to maximize the predictive performance while minimizing the adversary’s ability to predict the sensitive features;
- **Calibrated equal odds post-processing** (abbreviated as CEOP in the comparison)Pleiss et al. (2017): a post-processing model that proposes to minimize the error disparity among different groups indicated by the sensitive feature;
- **Disparate impact remover** (abbreviated as DIR in the comparison) Feldman et al. (2015): a model that proposes to minimize the disparity in the outcome from different groups via pre-processing;
- **Reweighing method** Kamiran & Calders (2012): a pre-processing method that eliminates the discrimination bias among different groups by reweighing and re-sampling the data;
- **Learning Adversarially Fair and Transferable Representations** (abbreviated as LAFTR in the cmoparison) Madras et al. (2018): fair representation learning by adversarial network that adopts fairness metric as adversarial objective;

- **Baseline method without fairness constraint**: 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.

2.2 DATASET

- **Adult** (also know as Census Income) data from the UCI repository (Kohavi, 1996): The data contains 48,842 instances described by 14 features (workclass, age, education, sex, race, etc.) and the goal is to predict whether income exceeds 50K USD per year. The feature *sex, race* is used as the sensitive feature;
- **Compas**[1]: The data includes 6,167 samples described by 401 features with the outcome showing if each person was accused of a crime within two years. The feature *sex, race* is used as the sensitive feature in this data;
- **CelebA image dataset**[2] (Liu et al., 2015): The data consists of 202,599 face images of the celebrities. The images are annotated with 40 attributes (face shape, eyeglasses, smiling, etc.). Similar to (Quadrianto et al., 2019), our goal in this data is to predict whether the person in the image is attractive or not. We use three pre-trained models (VGG16 (Simonyan & Zisserman, 2014), VGG19 (Simonyan & Zisserman, 2014), and ResNet50 (He et al., 2016)) to extract latent features for images in CelebA dataset. The feature *sex* is used as the sensitive feature.

2.3 FAIRNESS METRICS

We use three fairness metrics in evaluation, which include:

- **Absolute equal opportunity difference**: the absolute difference in true positive rate among different population groups;

$$\Big|P(\hat{Y}=1|A=1,Y=1)-P(\hat{Y}=1|A=0,Y=1)\Big|$$

- **absolute average odds difference**: the absolute difference in balanced classification accuracy among different population groups;

$$\frac{1}{2}\sum_{y\in\{0,1\}}\Big|P(\hat{Y}=1|A=1,Y=y)-P(\hat{Y}=1|A=0,Y=y)\Big|$$

- **disparate impact**: proportion of individuals that receive a positive output for two groups: an unprivileged group and a privileged group. The calculation is the proportion of the unprivileged group that received positive outcome divided by the proportion of the privileged group that received positive outcome.

$$\frac{P(Y=1|A=0)}{P(Y=1|A=1)}$$

2.4 EXPERIMENTAL SETUP

Features in the data are normalized to the range of $[0,\ 1]$. For image data, we process the data with backbone architectures (VGG16, VGG19, and ResNet50) to extract 1,000 features of the image as an input to FAIAS model. We run all comparing methods 5 times with 5 different random splits of the data and report the average performance and the standard deviation on the test set. We implement the comparing methods via the AI Fairness 360 toolbox (Bellamy et al., 2018). For other methods involving a hyper-parameter, i.e., the threshold value in CEOP, DIR, and Reweighing method, we tune the hyper-parameter in the range of $\{0,\ 0.1,\ 0.2,\ \ldots,\ 1\}$ and use the best hyper-parameter achieving the best balanced classification accuracy on the validation set. For our FAIAS model, we

---

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

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

	Adult				COMPAS			
Models	Acc	Eq.Opp	Avg.Odds Diff	1-DisImp	Acc	Eq.Opp	Avg.Odds Diff	1-DisImp
ABL	0.843	0.279	0.200	0.850	0.671	0.398	0.325	0.497
Adv_Deb	-	-	-	-	-	-	-	-
CEOP	0.251	0	0	-	0.549	0	0	-
DIR	0.785	0.300	0.311	0.825	0.657	0.252	0.238	0.665
Reweigh	0.795	0.250	0.259	0.773	0.638	0.262	0.245	0.499
LAFTR	-	-	-	-	-	-	-	-
FAIAS	0.842	0.159	0.125	0.808	0.646	0.164	0.135	0.272

Table 1: Comparison of performance and fairness of the methods with multiple sensitive attributes on Adult and Compas dataset. Higher accuracy (Acc.) indicates better classification performance. Lower values for all three fairness metrics (Eq. Opp., Avg. Odds Diff., 1 - DisImp) shows better fairness.

construct the predictor as a 5-layer neural network with 200 nodes in each layer. We adopt scaled exponential linear units (SELU) (Klambauer et al., 2017) as the activation function of the first 4 layers and the softmax function for the last layer. We use Adam optimizer (Kingma & Ba, 2014) and set the learning rate as 0.01 for FAIAS model and 0.004 for the selector. In validation, we include all feature with a sampling probability higher than 0.5 in FAIAS.

3 DISCUSSION ON DETECTED FEATURES

Here we discuss the detected sensitive relevant feature in Adult dataset related with the results shown in Figure 2 in the main paper. The classification goal in Adult data is to predict whether the income of an individual exceeds 50K USD per year. The sensitive feature is *sex*.

The bottom-5 scored features in FAIAS are: “education=Doctorate”, “native-country=Dominican-Republic”, “native-country=Honduras”, “occupation=Tech-support”, “marital-status=Separated”. While the top-5 scored features: “marital-status=Widowed”, “native-country=Portugal”, “capital-gain”, “workclass=Self-emp-not-inc”, “education=11th”. The bottom-5 scored features are the most sensitive-relevant features selected by the selector. It is notable that features like ’education=Doctorate’ and ’occupation=Tech-support’ are detected, and these features are highly indicative of the sensitive feature sex. This confirms the validity of the selected sensitive-relevant features. As for the top-5 features (the most sensitive-irrelevant features), we observe that features like ’capital-gain’ are detected, which is important for accurate classification of individual income per year. The interpretation results support the validity of selected sensitive-relevant and irrelevant features.

4 RESULTS WITH MULTIPLE SENSITIVE FEATURES

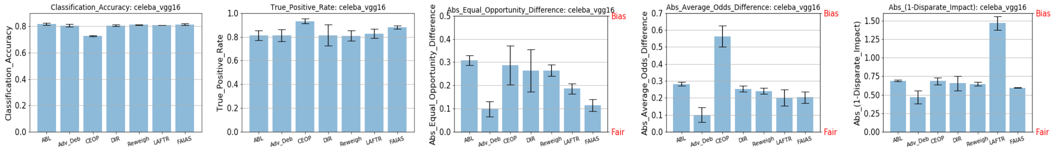
We can easily extend FAIAS to multiple sensitive attribute scenarios. The only change from the single sensitive feature case is the number of sensitive attributes (red entries in Figure 1 of the main paper). Below we show results on Adult, and Compas datasets with gender and sex as sensitive attributes (which results in 4 demographic combination groups). To measure fairness in multiple sensitive attributes, we use the disparity between the maximum and minimum value among demographics. For example, equal opportunity is formulated as

$$\max_a P(\hat{Y} = 1 | A = a, Y = 1) - \min_a P(\hat{Y} = 1 | A = a, Y = 1). \tag{1}$$

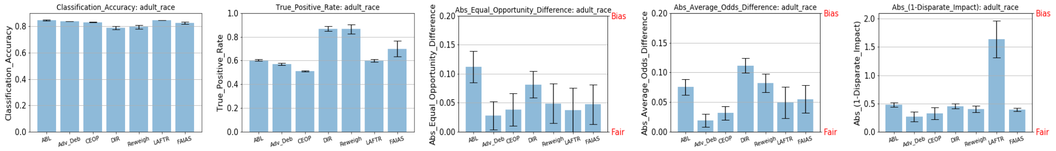
Note that results in Table. 1 shows that the fairness measures of ABL increased significantly compared to a single sensitive attribute case in the main paper. We can observe consistent improvement in fairness from FAIAS while keeping comparable accuracy. Note that for some methods, multiple sensitive attributes scenario is not applicable.

5 MORE EXPERIMENTAL RESULTS

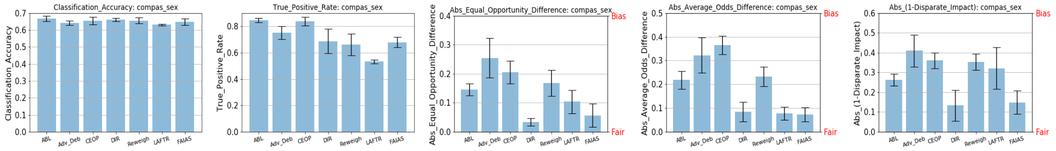
To complement the results in our main paper, we further show results using *race* as a sensitive feature in Adult data, and *sex* as a sensitive feature in Compas data in Figure 1b and 1c. In Figure 1a, we plot



(a) VGG16 (*sex*)



(b) Adult (*race*)



(c) Compas (*sex*)

Figure 1: Comparison of model performance via classification accuracy and true positive rate on Adult and Compas datasets (with sensitive feature shown in the parenthesis). Higher accuracy and true positive rate indicates better performance of prediction. Comparison of fairness via absolute equal opportunity difference, absolute average odds difference, and disparate impact on Adult and Compas datasets (sensitive feature shown in the parenthesis). Lower values for all three metrics indicates better fairness.

the result of VGG16 on CelebA data similar to the ones in the Figure 2 in the main paper. The results are consistent with the reported outcome that we reported in the main paper.

REFERENCES

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

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

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *NIPS*, pp. 971–980, 2017.

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

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.

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*, pp. 8227–8236, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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