Fairness via Group Contribution Matching

Tianlin Li[1] , **Zhiming Li**[1] , **Anran Li**[1,*] ,
Mengnan Du[2] , **Aishan Liu**[3] , **Qing Guo**[4,5] , **Guozhu Meng** [6] and **Yang Liu** [1,*]
[1]Nanyang Technological University, Singapore,
[2]New Jersey Institute of Technology, USA,
[3]Beihang University, China,
[4]Institute of High Performance Computing (IHPC), A*STAR, Singapore,
[5]Centre for Frontier AI Research (CFAR), A*STAR, Singapore,
[6]SKLOIS, Institute of Information Engineering, Chinese Academy of Sciences, China,
{tianlin001, zhiming001, anran.li, yangliu}@ntu.edu.sg, mengnan.du@njit.edu, liuaishan@buaa.edu.cn,
tsingqguo@ieee.org, mengguozhu@iie.ac.cn

Abstract

Fairness issues in Deep Learning models have recently received increasing attention due to their significant societal impact. Although methods for mitigating unfairness are constantly proposed, little research has been conducted to understand how discrimination and bias develop during the standard training process. In this study, we propose analyzing the contribution of each subgroup (*i.e.*, a group of data with the same sensitive attribute) in the training process to understand the cause of such bias development process. We propose a gradient-based metric to assess training subgroup contribution disparity, showing that unequal contributions from different subgroups are one source of such unfairness. One way to balance the contribution of each subgroup is through oversampling, which ensures that an equal number of samples are drawn from each subgroup during each training iteration. However, we find that even with a balanced number of samples, the contribution of each group remains unequal, resulting in unfairness under such a strategy. To address the above issues, we propose an easy but effective group contribution matching (GCM) method to match the contribution of each subgroup. Our experiments show that our GCM effectively improves fairness and outperforms other methods significantly.

1 Introduction

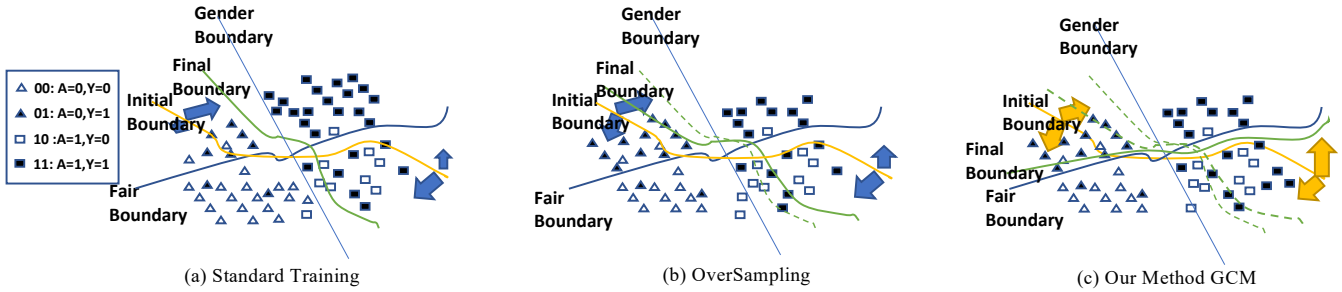
Deep learning has been increasingly adopted in more and more social applications, such as image classification [Deng *et al.*, 2009], speech recognition [Deng *et al.*, 2013], and natural language processing [Goldberg, 2016]. However, deep learning models often exhibit discriminatory behaviors (*e.g.*, distinct accuracy differences) towards certain groups (*e.g.*, African Americans and females), which are against people's desperate desire for social fairness. For example, when constructing a recidivism predictor using the dataset COMPAS, it is likely to predict African-American offenders with higher risk scores compared with Caucasians [ProPublica, 2016], which might threaten social stability and cause harm to individuals.

To alleviate model discrimination, various mitigation methods have been proposed [Donini *et al.*, 2018; Wu *et al.*, 2019; Bahng *et al.*, 2020; Du *et al.*, 2021]. However, those methods might build sub-optimal models which fail to produce fair solutions [Lohaus *et al.*, 2020]. In addition, it appears to any practitioner that the unfairness is dynamically changing during training. Existing methods neglect the dynamic and fail to provide insights into how unfairness develops in the training process. Understanding the unfairness development is vital as it can help prevent unfair growth in training iterations and enable better fairness guarantees. To this end, our work aims to interpret *how unfairness develops gradually during the training process* and thereby mitigate the unfairness.

We investigate the training process from the perspective of *data contribution* which estimates the contribution of training examples to the prediction, based on the following observations. First, when the training dataset is highly imbalanced, the model is prone to catch certain spurious correlations between target labels and protected attributes, and exhibits unfair behaviors. Second, the contributions of highly imbalanced training subgroups should be disparate intuitively, which has also been showcased in our experimental results in Sec. 3.1. To measure data contributions, existing methods evaluating individual sample contribution through techniques such as the Shapley value or influence functions incur large computational overhead. To this end, we propose a gradient-based method to investigate the contribution of each subgroup rather than each individual sample to avoid the heavy calculation. Specifically, we quantify the *group contribution disparity* metric through the gradient-based method, and use it to measure the contribution discrepancy between different subgroups.

Thus, the discrimination problem can be alleviated by balancing the contributions of different subgroups. Oversampling appears to be an effective strategy for equalizing the contribution of different subgroups in the training process. However, our preliminary experiments show that even when the sample

*Corresponding Authors.



(a) Standard Training (b) OverSampling (c) Our Method GCM

Figure 1: The decision boundary under different training strategies. (a) Under the standard training, the final decision boundary is trained to get close to the gender boundary due to the severely imbalanced data which generates a highly unfair model. (b) With the oversampling strategy, the skewed training process is mitigated. However, the actual contributions are still not matched. Compared with the standard training, the final decision boundary is less correlated to the gender boundary and gets closer to the ground-truth fair boundary, which means the fairness could be improved to some extent. (c) Our method matches the actual contributions of different subgroups, which leads to a fairer decision boundary (*i.e.*, the final boundary is close to the ground-truth fair boundary and almost orthogonal to the gender boundary).

sizes of different subgroups are the same during the training process, the contributions of different sub-groups remain disparate. Due to the disparity, fairness cannot be achieved even with an oversampling strategy. Furthermore, calculating the contribution of a sample is computationally prohibitive, so it is difficult to equalize the contribution of each subgroup by adjusting the subgroup size. To address these challenges, we propose an easy but effective Group Contribution Matching (GCM) method for aligning the contribution of each subgroup through a gradient-based strategy as shown in Fig. 1. Specifically, we integrate the gradient matching method into a gradient reweighing framework, which could adaptively adjust the contribution of each subgroup to achieve group contribution matching. Moreover, our layer-wise analysis shows that the cause of unfairness might mainly lie in the "bias" parameters in the model rather than the "weight" parameters.

In summary, we make the following contributions:

- We propose the *group contribution disparity* metric to evaluate the discrepancy among training subgroup contributions. With this metric, we analyze the training process and reveal that the unequal contributions of different subgroups are a source of unfairness.
- Based on the metric, we propose an easy but effective group contribution matching (GCM) method to improve fairness. Specifically, we design a gradient reweighing strategy to adaptively adjust the subgroup contribution.
- Extensive experiments on three public datasets show that our GCM method could effectively improve fairness and outperform other baseline methods significantly.

2 Preliminaries

In this section, we will first give the notations used in this work, followed by a detailed introduction to the fairness metrics.

2.1 Notation

We consider the task of learning a predictive model parameterized by a weight vector $\theta \in \mathbb{R}^p$ with p parameter elements, that maps an input space $\mathcal{X}$ to an output space $\mathcal{Y}$. Specifically, given a dataset $\mathcal{D} = \{z_1, z_2, \cdots, z_n\}$ with n training samples, where $z = (x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$, for a sample $z = (x, y)$ and parameters θ, let $l(z, \theta)$ denote the loss function and $\hat{y} = F(x)$ be the prediction result. The standard model training aims to select parameters in order to minimize an empirical risk $\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} l(z_i, \theta)$. Training samples can be divided into subgroups based on some sensitive or protected attributes $A \in \mathcal{A}$ such as gender, age, and race. Without loss of generality, we consider the binary classification task, *i.e.*, $Y \in \{0, 1\}$ and binary protected attribute settings, *i.e.*, $A \in \{0, 1\}$, where $A = 0$ represents unprivileged groups, while $A = 1$ represents privileged groups, respectively.

2.2 Fairness Evaluation Metrics

In this work, we follow the existing work [Wang *et al.*, 2022a] to consider two metrics to evaluate fairness: Equality of Opportunity [Verma and Rubin, 2018; Hardt *et al.*, 2016] and Equalized Odds [Romano *et al.*, 2020; Verma and Rubin, 2018]. The measures of the two metrics are based on the true positive rate $TPR_{A=a} = P(\hat{Y} = 1|A = a, Y = 1)$ and the false positive rate $FPR_{A=a} = P(\hat{Y} = 1|A = a, Y = 0)$ for $a \in \mathcal{A}$.

Equality of Opportunity expects both the privileged group ($A = 1$) and unprivileged group ($A = 0$) to have an equal probability of assigning a positive outcome to an instance from the positive class, which can be formulated as $P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$. Here we apply EOP given as follows to evaluate Equality of Opportunity,

$$EOP = \frac{TPR_{A=0}}{TPR_{A=1}} = \frac{P(\hat{Y} = 1|A = 0, Y = 1)}{P(\hat{Y} = 1|A = 1, Y = 1)}. \quad (1)$$

Equalized Odds considers the ground truth label y, and requires favorable outcomes to be conditionally independent of the sensitive attributes, which can be defined as $P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y)$ for $y \in \mathcal{Y}$. To evaluate Equalized Odds, ΔEO combines the difference of TPR and FPR across two sensitive groups as where $\Delta TPR = |TPR_{A=0} - TPR_{A=1}|$ and $\Delta FPR = |FPR_{A=0} - FPR_{A=1}|$.

$$\Delta EO = \Delta TPR + \Delta FPR. \quad (2)$$

Under the above definitions, EOP close to 1 and ΔEO close to 0 indicate fair classification results.

3 Methodology

In this section, we first introduce our proposed fairness metric, named *group contribution disparity*, for evaluating the contribution discrepancy during the training process, and illustrate its utility through various data-driven analyses (Sec. 3.1). Then, we propose our group contribution matching method based on the group contribution disparity metric to efficiently and adaptively improve fairness (Sec. 3.2).

3.1 Group Contribution Disparity

Intuitively, the building of spurious correlations between predictions and sensitive attributes indicates that the various training subgroups do not contribute equally to the training process. For example, if males in the test data are predicted with higher scores, the cause might be that the males with higher scores in the training data contribute more to the model training. To explore the contributions of various subgroups, we propose to estimate the subgroup-wise contribution through tracing gradient descent inspired by [Pruthi *et al.*, 2020; Paul *et al.*, 2021], which calculates individual data contribution via tracing the training process. It helps to understand the connection between the disparate contributions among different subgroups and the unfair prediction in the training process.

Formally, we define the idealized contribution of the training batch Z as the total reduction in loss on the test data Z', that is induced in the training process iteration t, *i.e.*, $C(Z, Z') = \sum_t \mathcal{L}(Z', \theta^t) - \mathcal{L}(Z', \theta^{t+1})$. Since the step sizes used in updating the parameters in the training process are typically quite small, we can approximate the change in the loss of the test data at t via a simple first-order approximation:

$$\mathcal{L}(Z', \theta^{t+1}) = \mathcal{L}(Z', \theta^t) + \nabla\mathcal{L}(Z', \theta^t) \cdot (\theta^{t+1} - \theta^t) + O(||\theta^{t+1} - \theta^t||^2). \tag{3}$$

As we utilize the gradient descent method to optimize, the change in parameters can be calculated as $\theta^{t+1} - \theta^t = -\eta_t \nabla\mathcal{L}(Z, \theta^t)$. Substituting the item $\theta^{t+1} - \theta^t$ in the first-order approximation, and ignoring the higher-order term, we obtain the approximation for the contribution of the training batch Z regarding the test data Z':

$$\begin{aligned} C(Z, Z') &= \sum_t \mathcal{L}(Z', \theta^t) - \mathcal{L}(Z', \theta^{t+1}) \\ &\approx \sum_t \eta_t \nabla\mathcal{L}(Z', \theta^t) \cdot \nabla\mathcal{L}(Z, \theta^t) \end{aligned} \tag{4}$$

Noted that we here follow previous works to calculate a scalar to evaluate the contribution and focus on comparing the contribution magnitudes of different subgroups. We first think about the contribution of each subgroup $Z_{A=a}$ (*i.e.*, the subgroup with attribute $A = a$). The contribution of $Z_{A=a}$ in the training batch Z to the test data Z' could be approximated as:

$$C(Z_{A=a}, Z') \approx \sum_t \eta_t \nabla\mathcal{L}(Z', \theta^t) \cdot \nabla\mathcal{L}(Z_{A=a}, \theta^t), a \in \mathcal{A}. \tag{5}$$

We then consider the subgroup $Z_{A=0}$ and its counterpart $Z_{A=1}$ with a greedy strategy in which we examine each individual parameter w in the model parameters θ one by one. The equal contributions in terms of w made by $Z_{A=0}$ and $Z_{A=1}$ are highly related to the difference between $|\nabla\mathcal{L}(Z_{A=0}, w)|^2$ and $|\nabla\mathcal{L}(Z_{A=1}, w)|^2$ under the independent and identically distributed training and test setting [1]. Inspired by this, we here separately consider when $Y = 0$ and $Y = 1$, and define the group contribution disparity for w as:

$$\hat{r}_w^{t,Y=y} = \frac{|g_w^{A=0,Y=y}|}{|g_w^{A=1,Y=y}|} = \frac{|Z_{0y}| \cdot |\overline{g}_w^{A=0,Y=y}|}{|Z_{1y}| \cdot |\overline{g}_w^{A=1,Y=y}|}, \tag{6}$$

$$g_w^{A=a,Y=y} = \frac{\partial \mathcal{L}^{A=a,Y=y}}{\partial w}, \forall a \in \mathcal{A}, \forall y \in \mathcal{Y}, \forall w \in \theta, \tag{7}$$

where $|Z_{ay}|$ represents the average group size of Z_{ay} for $a \in \{0, 1\}$, $y \in \{0, 1\}$, $\mathcal{L}^{A=a,Y=y}$ is the cross entropy regarding the training group Z_{ay}, and $g_w^{A=a,Y=y}$ denotes the average gradient derived by individual example $z_{ay} \in Z_{ay}$ in terms of the parameter w. We see that the group contribution disparity score $\hat{r}_w^{t,Y=y}$ close to 1 indicates equal contribution, and the score is largely influenced by the group size disparity $\frac{|Z_{0y}|}{|Z_{1y}|}$ in the training batch Z. Then the group contribution disparity score of $Y = y$ for all p parameters in θ could be defined as:

$$\hat{r}_\theta^{t,Y=y} = \frac{1}{p} \sum_{w \in \theta} \hat{r}_w^{t,Y=y} = \frac{1}{p} \sum_{w \in \theta} \frac{|Z_{0y}| \cdot |\overline{g}_w^{A=0,Y=y}|}{|Z_{1y}| \cdot |\overline{g}_w^{A=1,Y=y}|}. \tag{8}$$

In this way, our metric has the following advantages: ❶ The calculation of $\hat{r}_\theta^{t,Y=y}$ avoids the heavy gradient computation regarding the loss on the test data, which makes the estimation of the contribution disparity score in each training iteration possible. ❷ Noted that, compared with previous methods computationally prohibitive to calculate each training sample's importance, our calculation process splits the training data into several subgroups and calculates the gradient for each subgroup, which largely lessens the computation.

Moreover, we can also calculate the group contribution disparity for a particular layer l at the iteration t as follows:

$$\hat{r}_{\theta_l}^{t,Y=y} = \frac{1}{p_l} \sum_{w \in \theta_l} \frac{|Z_{0y}| \cdot |\overline{g}_w^{A=0,Y=y}|}{|Z_{1y}| \cdot |\overline{g}_w^{A=1,Y=y}|}, \forall y \in \mathcal{Y}, \forall l \in L. \tag{9}$$

where θ_l is the model parameters in layer l, p_l is the parameter size in layer l, and L contains all parameter layers in the model.

In the following sections, we empirically analyze the connection between group contribution disparity and fairness, which helps us to understand how unfairness develops.

Analyses and Observations

We conduct analyses to explore the relationship between fairness and the group contribution disparity score. We used the

[1] When we assume the identical distribution between Z and Z', and $\nabla\mathcal{L}(Z', w) \approx \nabla\mathcal{L}(Z, w) \approx \nabla\mathcal{L}(Z_{A=0}, w) + \nabla\mathcal{L}(Z_{A=1}, w)$ [Goodfellow *et al.*, 2016], if we expect $C(Z_{A=0}, Z') = C(Z_{A=1}, Z')$, we can have $(\nabla\mathcal{L}(Z_{A=0}, w) + \nabla\mathcal{L}(Z_{A=1}, w)) \cdot \nabla\mathcal{L}(Z_{A=0}, w) \approx (\nabla\mathcal{L}(Z_{A=0}, w) + \nabla\mathcal{L}(Z_{A=1}, w)) \cdot \nabla\mathcal{L}(Z_{A=1}, w)$, *i.e.*, $|\nabla\mathcal{L}(Z_{A=0}, w))|^2 \approx |\nabla\mathcal{L}(Z_{A=1}, w))|^2$.

CelebA	$\lvert Z_{00}\rvert$	$\lvert Z_{01}\rvert$	$\lvert Z_{10}\rvert$	$\lvert Z_{11}\rvert$	Adult	$\lvert Z_{00}\rvert$	$\lvert Z_{01}\rvert$	$\lvert Z_{10}\rvert$	$\lvert Z_{11}\rvert$
Standard Training	34	51	35	16	Standard Training	78	10	125	57
OverSampling	32	32	32	32	OverSampling	100	100	100	100

Table 1: Averaged data sizes of different subgroups in one iteration.

well-known tabular dataset Adult [Dua and Graff, 2017] and ResNet-18 [Liu *et al.*, 2015]. We train an MLP [Bishop, 1996] and a residual network [He *et al.*, 2016] as the classification models, respectively. For the CelebA dataset, we train the ResNet-18 for classifying the *attractive* attribute as adopted in [Chuang and Mroueh, 2021]. We mainly conduct our analyses on the $-\Delta$EO metric. More details about the datasets, the models, the metric, and the training methods are introduced in Sec. 4.1. We train models both under the standard training setting and the oversampling setting. The final fairness scores of trained models are reported as follows. On the CelebA dataset, the fairness score ($-\Delta$EO) is escalated from -0.480 to -0.049 with the oversampling strategy (*i.e.*, fairness is improved significantly). On the Adult dataset, the fairness score ($-\Delta$EO) changes from -0.096 to -0.140 after oversampling. The sizes of training subgroups in each training iteration are shown in Table 1. We can see that the size of Z_{01} (*i.e.*, 51) is far larger than the size of Z_{11} (*i.e.*, 16) under the standard training on the CelebA dataset. And the training subgroup size discrepancy is even larger on the Adult dataset.

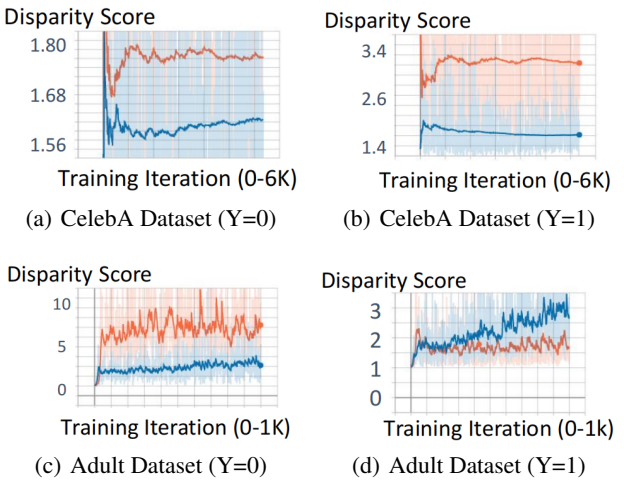


Figure 2: The Group Contribution Disparity score on the CelebA ((a), (b)) and Adult dataset ((c), (d)). The red line represents the standard training process. The blue line represents the training process when oversampling is used. The plot is smoothed for better observation.

Fig. 2 shows the group disparity score varying in the training process of the CelebA dataset and the Adult dataset. We can make the following key observations: ❶ From subfigures (a) and (b), we can see that the oversampling strategy reduces the group contribution disparity both when Y=0 and Y=1 on the CelebA dataset. The fairness improvement is salient with an 89.8% improvement ($-\Delta$EO score from -0.480 to -0.049). There is a positive correlation between the reduction in the contribution disparity score and fairness. ❷ From subfigures

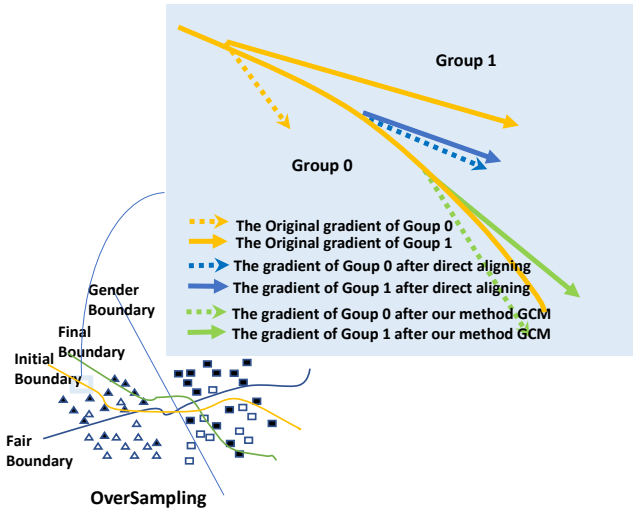


Figure 3: The relation of different gradient strategies. We here only consider two subgroups with the same targets and different genders to demonstrate our method.

(a) and (b), we can see that the disparity score when Y=0 is smaller than that of Y=1 (1.76 *vs.* 3.0) in the standard training process. We think this is because the data sizes of Z_{00} and Z_{10} are closer (34 *vs.* 35) while the data size of Z_{01} and Z_{11} are disparate (51 *vs.* 16). ❸ From From subfigures (c) and (d), we can see that on the Adult dataset, the oversampling strategy only reduces the group contribution disparity when Y=0 and the contribution disparity score increases when Y=1. The $-\Delta$EO score even decreases to -0.140 compared with the standard training method (-0.096), which indicated the fairness even dropped with the oversampling strategy. The results show that oversampling cannot consistently reduce the disparity score and improve fairness. ❹ The contribution disparity still exists after oversampling even on the CelebA dataset (*i.e.*, the group contribution disparity score is around 1.5 for both $Y = 0$ and $Y = 1$).

The experimental results show that fairness and group contribution disparity scores are highly related: higher group contribution disparity means worse fairness. We conclude that the unfairness gradually develops in the training process due to the existing training group contribution disparity. Moreover, these observations motivate us to explore further reducing group contribution disparity to improve fairness.

3.2 Proposed Group Contribution Matching

Motivated by the phenomenon that unequal contributions brought by different subgroups in the learning process cause unfairness, we propose the Group Contribution Matching (GCM) framework to mitigate the bias via equalizing the contribution of different sub-groups. As the sample reweighing paradigm could not easily match the group contribution due to the large computation overhead to calculate the contribution of each individual sample, here we consider equalizing the contributions through an in-processing method (*i.e.*, designing a new loss item). One straightforward way to optimize Eq. (6) close to 1 is to directly align the gradient of Z_{0y} and Z_{1y}

(*i.e.*, minimize $\nabla\mathcal{L}(Z_{0y},\theta) - \nabla\mathcal{L}(Z_{1y},\theta)$. This method is referred as *direct gradient aligning*). However, it might still result in a biased optimization as shown in Fig. 3 (*i.e.*, the two blue lines are aligned but the directions are still biased to group 1). Under such an optimization, the unprivileged group also be optimized to make contributions toward the privileged group, which will still lead to an unfair model. Experiments to showcase the biased optimization are deferred to Sec. 4.3.

To address this issue, we propose to minimize the magnitude variance in the gradients of two subgroups. In this way, we can keep the original direction of the unprivileged group and match the gradient magnitudes of the unprivileged and privileged groups. We propose the gradient matching method as follows.

Gradient Matching

To achieve equal contributions, we design a new loss item:

$$\mathcal{L}_{fair} = \sum_{w\in\theta}\sum_{y\in\{0,1\}}(|g_w^{A=0,Y=y}| - |g_w^{A=1,Y=y}|). \quad (10)$$

Then the final loss item can be revised as

$$\mathcal{L} = \mathrm{E}_{(\mathbf{x},y)\sim P}(\mathcal{L}_{\text{cls}}(\mathrm{F}(\mathbf{x}),y)) + \lambda\mathcal{L}_{fair}. \quad (11)$$

Furthermore, from the training process in Fig. 2, we can observe that the group contribution disparity and unfairness exist in each single training iteration. We expect that the unprivileged group could contribute more to the model in the following training iterations instead of keeping an equal contribution with its counterpart subgroup.

Gradient Reweighing

To distinguish which group is privileged in each training iteration, normally we have to evaluate the model via the fairness metrics introduced in Sec. 2.2. However, the evaluation in each training iteration is time-consuming. To address such an issue, we design a group confidence score as the privilege indicator. Specifically, we get inspired from [Madras *et al.*, 2018] and regard a subgroup with a higher confidence score as the privileged group. For example, when $Y=1$, males are privileged if they get higher prediction scores compared with females. The confidence score of the subgroup $(A=a, Y=y)$ can be calculated by the average output of the data group which is denoted as $\overline{y}^{A=a,Y=y}$:

$$c^{A=a,Y=y} = (\overline{y}^{A=a,Y=y} * y + (1-\overline{y}^{A=a,Y=y}) * (1-y)). \quad (12)$$

As we expect the unprivileged group (of lower confidence score) to be optimized to contribute more (*i.e.*, own higher gradient magnitude), we assign the unprivileged group a lower weight. Thus, we here set the weight as $W^{A=a,Y=y} = (c^{A=a,Y=y})^T$, where T is the temperature scale. Specifically, the loss item is as follows:

$$\mathcal{L}_{fair} = \sum_{w\in\theta}\sum_{Y\in\{0,1\}}(W^{A=0,Y=y} * |g_w^{A=0,Y=y}| - W^{A=1,Y=y} * |g_w^{A=1,Y=y}|). \quad (13)$$

In this way, Eq. (10) can be reformulated as a case of the gradient reweighing loss item when $T=0$. In our paper, we mainly set the T value in the range $\{0,2\}$.

4 Experiments

4.1 Experimental Settings

Datasets. In our experiments, we use two tabular benchmarks (**Adult** and **COMPAS**) and one image dataset (**CelebA**) that are all for binary classification tasks: ❶ Adult [Dua and Graff, 2017]. The original aim of the dataset Adult is to determine whether a person makes salaries over 50K a year. We consider *gender* as the sensitive attribute, and the vanilla training will lead the model to predict females to earn fewer salaries. ❷ CelebA [Liu *et al.*, 2015]. The CelebFaces Attributes dataset is to predict the attributes of faces. We split into two subgroups according to the attribute *gender*. Here we consider two attributes classification tasks. For the task to predict whether the hair in an image is *wavy* or not, the standard training will show discrimination towards the male group; when predicting whether the face is *attractive*, the standard training will result in a model prone to predict males as less attractive. ❸ COMPAS [Mele and many others, 2017 2021]. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a well-known commercial algorithm that judges and parole authorities use to determine whether a criminal defendant is likely to commit another crime (recidivism). It has been demonstrated that the algorithm is biased against black inmates and in favor of white defendants (*i.e.*, who actually committed crimes or violent crimes after 2 years).

Metrics. For fairness evaluation, we take two group fairness metrics ΔEO and EOP as we introduced in the Sec. 2.2 and define $-\Delta$EO and EOP as fairness scores since higher $-\Delta$EO and EOP mean better fairness. We use the average precision (AP) for classification accuracy evaluation. Our method could also be extended to more fairness metrics.

Models. For tabular benchmarks, we use the MLP (multilayer perception) as the classification model, which is commonly adopted in classifying tabular data. For the CelebA dataset, we use AlexNet [Krizhevsky *et al.*, 2012] and ResNet-18 [He *et al.*, 2016], both of which are popular in classifying image data [Alom *et al.*, 2018]. We mainly show the experimental results of predicting *wavy hair* using AlexNet.

Mitigation Baselines. Following the common setups in [Chuang and Mroueh, 2021], we compare our method with several baselines: ❶ Standard training (*i.e.*, Vanilla). The training is based on the empirical risk minimization (ERM) principle. DNNs are trained only with the cross entropy loss. ❷ Oversample (*i.e.*, OverSampling) [Wang *et al.*, 2022a]. This method samples from the subgroup with rare examples more often, making a balanced sampling in each epoch. ❸ Equalized Odds Regularization (*i.e.*,EOR) [Madras *et al.*, 2018]. This method is to directly regularize the fairness metrics. ❹ Adversarial debiasing. (*i.e.*, Adversarial) [Zhang *et al.*, 2018]. This method minimizes the adversary's ability to predict sensitive attributes.

Implementation Details. For the adult dataset, we follow the settings in [Chuang and Mroueh, 2021] for data preprocessing. The hidden size of MLP is 200. We use Adam as the learning optimizer and the batch size is set as 2000 following the setting in [Chuang and Mroueh, 2021]. The learning rate

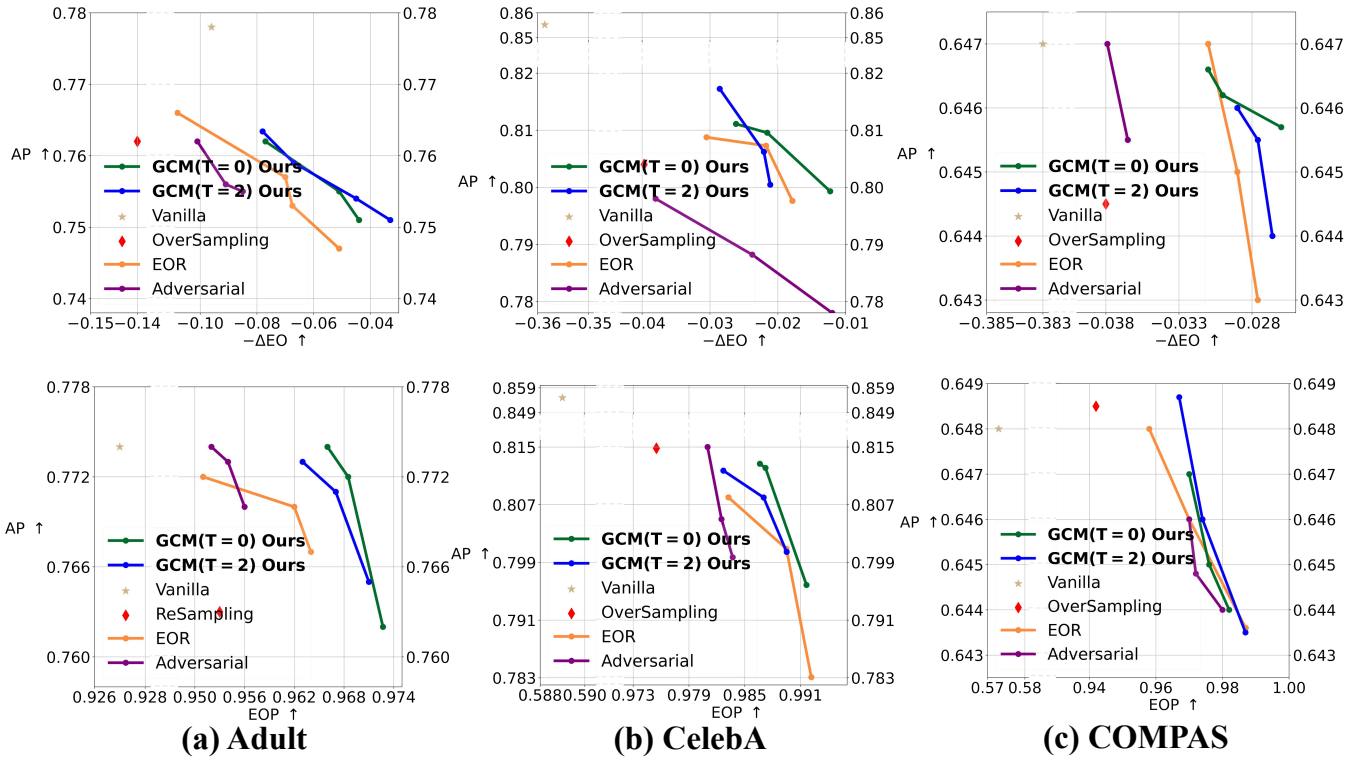


Figure 4: Comparing different methods on AP vs. (-ΔEO/EOP). We train networks with the compared methods for 10 times and the averaging results are reported. We here show the results of our method GCM when T is set as 0 and 2.

is set as 0.001. For the CelebA dataset, We follow the settings in [Chuang and Mroueh, 2021] for data preprocessing. We use Adam as the learning optimizer and the batch size is set as 128. The learning rate is set as 0.0001. For the COMPAS dataset, we use Adam as the learning optimizer, and the batch size is set as 2000. The learning rate is set as 0.001.

4.2 Experimental Results

As shown in Fig. 4, we have the following observations: ❶ Although fairness improvement methods could generate fairer networks with high fairness scores, the corresponding accuracies of all methods decrease significantly. From the figure, we can see that when using the EOR method on the Adult dataset, the $-\Delta$EO score increases from -0.108 to -0.051 and the accuracy decreases from 0.766 to 0.747. Our method (*i.e.*, GCM (T=2)) also increases the $-\Delta$EO score from -0.078 to -0.033, and the accuracy is reduced from 0.763 to 0.751. ❷ Oversampling could improve fairness under most settings. For example, on the CelebA dataset, the $-\Delta$EO score is escalated from -0.359 to -0.040. Also, on the COMPAS dataset, the $-\Delta$EO score is improved 90.1% (-0.383 to -0.038). However, as we introduced in Sec. 3.1, oversampling fails to improve fairness due to the large group contribution disparity on the Adult dataset. Furthermore, we can see that oversampling is less effective than some other methods. For example, on the CelebA dataset, EOR method could achieve -0.031 $-\Delta$EO score and maintain the AP score as 0.809, both of the metrics are better than those of oversampling. ❸ Our method GCM (T=0) and GCM (T=2) achieve higher fairness (i.e., higher -ΔEO and EOP scores) than all baseline methods when they have similar accuracy. In particular, on the Adult dataset, GCM (T=2) arrives at -0.033 $-\Delta$EO score when the accuracy decreases to 0.751. The best method among baselines (EOR) arrives at -0.051 $-\Delta$EO score when the accuracy decreases to 0.747. GCM (T=2) gains a relative 35.3% improvement compared with the best baseline method while maintaining a higher accuracy score. Moreover, on the Adult dataset, GCM (T=2) outperforms GCM (T=0) when using $-\Delta$EO metric, which shows the superiority of the gradient reweighing strategy. Overall, our method can enhance fairness significantly with much less accuracy drop.

4.3 Further Discussion

Layer-wise Analysis

We here further analyze which part of the parameters should be more responsible for the unfairness on the Adult dataset. Here we separately consider the "weight" parameters and the "bias" parameters and do the analysis layer by layer. The fairness loss item to optimize layer l is designed as follows:

$$\mathcal{L}^{l}_{fair} = \sum_{w \in \theta_l} \sum_{Y \in \{0,1\}} (W^{A=0,Y=y} * |g_w^{A=0,Y=y}| - W^{A=1,Y=y} * g_w^{A=1,Y=y}|). \quad (14)$$

The experimental results are shown as follows:

From Table 2, we can see that optimizing "$bias_0$" achieves better fairness compared with "$weight_0$" (the APs are the

	OverSampling	$weight_0$	$bias_0$	$weight_1$	$bias_1$	$weight_{0+1}$
λ		0.5	5.0	50.0	50.0	1.0
AP score	0.762	0.759	0.759	0.747	0.755	0.749
Fairness score ($-\Delta$EO)	-0.140	-0.118	-0.058	-0.079	-0.070	-0.090

Table 2: The experimental results of the layer-wise analysis. "$weight_0$", "$bias_0$", "$weight_1$", "$bias_1$" and "$weight_{0+1}$" represent when we only align the group contribution for the "weight" in layer 0, the "bias" in layer 0, the "weight" in layer 1, the "bias" in layer 1, and the "weight" in layer 0 and layer1, respectively.

same and the $-\Delta$EO value -0.058 is far better than -0.118). The "bias" parameter is more likely to be the cause of the unfairness because optimizing the "bias" parameters tends to achieve better fairness. Moreover, we optimize the "weight" parameters in both layers ("$weight_{0+1}$"), the fairness improvement of which is also limited. Recently some work is constantly proposed to understand the model behaviors at the neuron level. Such work usually bases their analysis on the neuron output (*i.e.*, the output of the "weight" parameters), while neglecting the "bias" parameters. However, our experiments show that the "bias" parameters are also likely to be responsible for the unfairness.

Other Gradient Matching Choices

Fairness improvement could also be regarded as requiring different domains (*i.e.*, domains of different subgroups) to own the same generalization performance, the essence of which is similar to domain generalization. Recently, some gradient matching methods have emerged in the domain generalization field. One typical gradient matching method Inter Gradient Alignment (IGA) is proposed by [Koyama and Yamaguchi, 2020] to improve the invariance of inter-domain gradients to learn invariant features. Specifically, IGA learns invariant features by minimizing the variance of inter-domain gradients. The original optimization objective is as follows:

$$\text{argmin}_\theta E[L_\epsilon(\theta)] + \lambda trace(Var(\Delta_\theta L_\epsilon(\theta))), \tag{15}$$

where L_ϵ is the loss of the θ-paramterized prediction model computed on the environment ϵ [2]. Following the spirit of IGA in the fairness area, we adapt the optimization objective as:

$$\mathcal{L}_{fair} = \sum_{w\in\theta}\sum_{a\in\{0,1\}} (g_w^{A=a} - g_w), \tag{16}$$

where g_w is the gradient regarding the whole training batch and $g_w^{A=a}$ is the parameters gradients regarding the data group $A = a$. The experimental results are shown in Table 3.

	Van	λ = 0	λ=1.0	λ=2.0	λ=10.0
AP score	0.778	0.762	0.765	0.771	0.755
Fairness score ($-\Delta$EO)	-0.096	-0.140	-0.095	-0.121	-0.103

Table 3: The experimental results of the IGA method. The fairness is restrictedly improved by the IGA method.

We can find that the IGA method cannot be directly used to improve fairness. Even if the gradients of the two domains (*i.e.* males and females) are aligned, we believe that optimizing IGA will result in a biased gradient and an imbalanced contribution. We could revise the IGA loss function as follows:

$$\mathcal{L}_{fair} = \sum_{w\in\theta}\sum_{y\in\{0,1\}} (g_w^{A=0,Y=y} - g_w^{A=1,Y=y}). \tag{17}$$

Different from the initial IGA method, this method splits the data into four groups rather than two domains. This method is identical to the algorithm to align the gradients of different subgroups directly as we mentioned in Sec. 3.2 which is referred as *direct gradient aligning*. Our experimental results are shown in Table 4. We can see that such a method still barely improves the fairness of the target model when under satisfying accuracy which shows the superiority of our gradient reweighing method.

	λ=0.0	λ=0.3	λ=0.5	λ=0.8	λ=1.0
AP score	0.762	0.750	0.748	0.738	0.732
Fairness score (-ΔEO)	-0.140	-0.130	-0.106	-0.071	-0.061

Table 4: Experimental results of the *direct gradient aligning* method.

Experiments on Other Datasets

We further extend experiments on two datasets, Colored MNIST [Arjovsky *et al.*, 2019], and CIFAR-10S [Wang *et al.*, 2020] to show the performance of GCM. Specifically, for Colored MNIST, we color each image with two colors spuriously with the label to assign a preliminary binary label $\widetilde{y}$: $\widetilde{y}$ = 0 for digits 0-4 and $\widetilde{y}$ = 1 for 5-9 and the sensitive attribute is color [Arjovsky *et al.*, 2019]. For CIFAR-10S, it has 10 object classes and we employ it to validate the multi-classification scenarios. The sensitive attribute is grayscale or color. The fairness metrics, ΔEO and EOP, are primarily designed and widely adopted by binary classification tasks [Chuang and Mroueh, 2021; Du *et al.*, 2021]. We use the Bias metric [Wang *et al.*, 2020] for the multi-class task. The results below show that GCM outperforms all baselines as presented in Table 5 and 6.

Colored MNIST	Vanilla	OverSampling	Adv	EOR	GCM (T=0)	GCM (T=2)
AP ↑	0.995	0.906	0.896	0.907	0.912	0.908
-ΔEO ↑	-0.750	-0.020	-0.031	-0.016	-0.015	**-0.013**

Table 5: Experimental results on the Colored MNIST dataset.

CIFAR-10S	Vanilla	OverSampling	Adv	GCM
Acc of Color/Gray/Mean↑	89.0/88.0/88.5	89.2/89.1/89.1	84.6/83.5/84.1	89.6/88.2/88.8
Fairness: Metric Bias ↑	-0.074	-0.066	-0.094	**-0.028**

Table 6: Experimental results on the CIFAR-10S dataset.

Experiments on Other Fairness Metrics

We mainly present two group fairness measures following [Wang *et al.*, 2022a; Jung *et al.*, 2022] and we can also potentially extend GCM to other measures. For example, considering the metric demographic parity DP, we take the resampling method in [Baniecki *et al.*, 2021] to sample Z_{00}, Z_{01}, Z_{10}, and

[2] Please refer to the Domainbed benchmark [Gulrajani and Lopez-Paz, 2020] for more details.

Z_{11}, and then align $\overline{g}_w^{A=0,Y=y}$ and $\overline{g}_w^{A=1,Y=y}$ for $y \in \{0,1\}$, $w \in \theta$. We conduct experiments on the COMPAS dataset and find that compared with the vanilla method with AP/-ΔDP scores 0.649/-0.245, our GCM method can achieve an AP/-ΔDP 0.633/-0.0027. The -ΔDP is 1/10 of that of resampling (0.636/-0.029) and outperforms other baselines largely (EOR: 0.634/-0.015). Our method GCM is compatible with the resampling strategies designed for different fairness measures and improves fairness for various fairness measures.

5 Related Work

5.1 Fairness Mitigation

Deep learning models easily exhibit some undesirable behaviors on concerns such as robustness, privacy, and other trustworthiness issues [Goodfellow *et al.*, 2014; Madry *et al.*, 2017; Kurakin *et al.*, 2018; Liu *et al.*, 2019; Liu *et al.*, 2020b; Liu *et al.*, 2020a; Hu *et al.*, 2021; Li *et al.*, 2021d; Liu *et al.*, 2021; Li *et al.*, 2022; Xie *et al.*, 2022; Liu *et al.*, 2023; Guo *et al.*, 2023; Xiao *et al.*, 2023; Huang *et al.*, 2023], in which discrimination could be extremely socially influential. There is a line of work dedicated to alleviating unfairness in DNNs. For example, several mitigation methods including oversampling, adversarial training, and other domain-independent methods are compared in [Wang *et al.*, 2020]. [Bahng *et al.*, 2020; Sarhan *et al.*, 2020] propose to disentangle unbiased representations to ensure fair DNNs. Moreover, [Du *et al.*, 2021] directly repair the classifier head even though the middle representations are still biased. [Madras *et al.*, 2018] propose to relax the fairness metrics for optimization. [Roh *et al.*, 2020] propose a batch gradient descent approach that can be used to learn fair models, which induces accurate models with multiple groups but is sub-optimal in terms of fairness [Maheshwari and Perrot, 2022]. Recently, [Li *et al.*, 2023] propose an approach to enhance fairness by emphasizing decision rationale alignment, while this approach requires a more complex hyper-parameter setting. However, these methods neglect the unfairness development process. In our work, the analysis to *how unfairness develops gradually during the training process* provides insights into the blind spots of models, which in turn offers us clues to build fairer models.

5.2 Data Contribution

Some works [Schaul *et al.*, 2015; Koh and Liang, 2017; Katharopoulos and Fleuret, 2018; Pruthi *et al.*, 2020; Li *et al.*, 2021b; Wang *et al.*, 2022b] have been proposed to evaluate the instance contribution to better control the training data input in the real world. These works either use the gradients [Li *et al.*, 2021a; Li *et al.*, 2021c] or the loss to compute each sample's importance. However, the former is computationally expensive and the latter is not a particularly good approximation of the gradient norm [Katharopoulos and Fleuret, 2018].

In our work, we investigate the contribution of each training subgroup rather than each individual sample, which avoids the heavy calculation. Moreover, different from previous methods, our definition to group contribution disparity could bypass the calculation regarding the test data, which is also time-saving.

6 Conclusions and Future Work

In this paper, we examine the training process to better understand how unfairness dynamically develops. We propose the group contribution disparity metric and observe that unequal contributions of different sub-groups are the source of unfairness. We further illustrate that the oversampling strategy fails to match the contribution of each sub-group. To effectively match the group contribution, we then propose the gradient reweighing method, which significantly improves fairness. Although promising, our method necessitates the computation of second-order derivatives. In subsequent research, we aim to enhance our method through more efficient calculations.

Ethical Statement

Our work would not produce potential negative societal consequences and has no ethical concerns.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08-022T). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore. It is also supported by A*STAR Centre for Frontier AI Research, the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019), National Satellite of Excellence in Trustworthy Software System No. NRF2018NCR-NSOE003-0001, NRF Investigatorship No. NRF-NRFI06-2020-0001, and the National Natural Science Foundation of China 62206009. We gratefully acknowledge the support of NVIDIA AI Tech Center (NVAITC). IIE authors are supported in part by Beijing Nova Program.

Contribution Statement

Tianlin Li and Zhiming Li made equal contributions. All the authors participated in designing research, performing research, analyzing data, and writing the paper.

References

[Alom *et al.*, 2018] Md Zahangir Alom, Tarek M Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S Awwal, and Vijayan K Asari. The history began from alexnet: A comprehensive survey on deep learning approaches. *arXiv preprint arXiv:1803.01164*, 2018.

[Arjovsky *et al.*, 2019] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[Bahng *et al.*, 2020] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *ICML*, 2020.

[Baniecki *et al.*, 2021] Hubert Baniecki, Wojciech Kretowicz, Piotr Piatyszek, Jakub Wisniewski, and Przemyslaw Biecek. dalex: Responsible machine learning with interactive explainability and fairness in python. *Journal of Machine Learning Research*, 22(214):1–7, 2021.

[Bishop, 1996] Christopher M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., USA, 1996.

[Chuang and Mroueh, 2021] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2021.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[Deng *et al.*, 2013] Li Deng, Geoffrey Hinton, and Brian Kingsbury. New types of deep neural network learning for speech recognition and related applications: an overview. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8599–8603, 2013.

[Donini *et al.*, 2018] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *NeurIPS*, 31, 2018.

[Du *et al.*, 2021] Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *NeurIPS*, 2021.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2017. Accessed: 2023-03-11.

[Goldberg, 2016] Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Goodfellow *et al.*, 2016] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[Gulrajani and Lopez-Paz, 2020] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. *CoRR*, abs/2007.01434, 2020.

[Guo *et al.*, 2023] Jun Guo, Wei Bao, Jiakai Wang, Yuqing Ma, Xinghai Gao, Gang Xiao, Aishan Liu, Jian Dong, Xianglong Liu, and Wenjun Wu. A comprehensive evaluation framework for deep model robustness. *Pattern Recognition*, 2023.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *NeurIPS*, 29, 2016.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[Hu *et al.*, 2021] Ming Hu, Jiepin Ding, Min Zhang, Frédéric Mallet, and Mingsong Chen. Enumeration and deduction driven co-synthesis of ccsl specifications using reinforcement learning. In *RTSS*, 2021.

[Huang *et al.*, 2023] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W. Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything, 2023.

[Jung *et al.*, 2022] Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. In *CVPR*, 2022.

[Katharopoulos and Fleuret, 2018] Angelos Katharopoulos and François Fleuret. Not all samples are created equal: Deep learning with importance sampling. In *ICML*, 2018.

[Koh and Liang, 2017] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *ICML*, 2017.

[Koyama and Yamaguchi, 2020] Masanori Koyama and Shoichiro Yamaguchi. When is invariance useful in an out-of-distribution generalization problem? *arXiv preprint arXiv:2008.01883*, 2020.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Li *et al.*, 2021a] Anran Li, Lan Zhang, Juntao Tan, Yaxuan Qin, Junhao Wang, and Xiang-Yang Li. Sample-level data selection for federated learning. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*, 2021.

[Li *et al.*, 2021b] Anran Li, Lan Zhang, Junhao Wang, Feng Han, and Xiang-Yang Li. Privacy-preserving efficient federated-learning model debugging. *IEEE Transactions on Parallel and Distributed Systems*, 2021.

[Li *et al.*, 2021c] Anran Li, Lan Zhang, Junhao Wang, Juntao Tan, Feng Han, Yaxuan Qin, Nikolaos M Freris, and Xiang-Yang Li. Efficient federated-learning model debugging. In *ICDE*, pages 372–383, 2021.

[Li *et al.*, 2021d] Tianlin Li, Aishan Liu, Xianglong Liu, Yitao Xu, Chongzhi Zhang, and Xiaofei Xie. Understanding adversarial robustness via critical attacking route. *Information Sciences*, 547:568–578, 2021.

[Li *et al.*, 2022] Zhiming Li, Yanzhou Li, Tianlin Li, Mengnan Du, Bozhi Wu, Yushi Cao, Xiaofei Xie, Yi Li, and Yang Liu. Unveiling project-specific bias in neural code models. *arXiv preprint arXiv:2201.07381*, 2022.

[Li *et al.*, 2023] Tianlin Li, Qing Guo, Aishan Liu, Mengnan Du, Zhiming Li, and Yang Liu. FAIRER: Fairness as decision rationale alignment. In *ICML*, 2023.

[Liu *et al.*, 2015] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[Liu *et al.*, 2019] Aishan Liu, Xianglong Liu, Jiaxin Fan, Yuqing Ma, Anlan Zhang, Huiyuan Xie, and Dacheng Tao. Perceptual-sensitive gan for generating adversarial patches. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 1028–1035, 2019.

[Liu *et al.*, 2020a] Aishan Liu, Tairan Huang, Xianglong Liu, Yitao Xu, Yuqing Ma, Xinyun Chen, Stephen Maybank, and Dacheng Tao. Spatiotemporal attacks for embodied agents. In *ECCV*, 2020.

[Liu *et al.*, 2020b] Aishan Liu, Jiakai Wang, Xianglong Liu, bowen Cao, Chongzhi Zhang, and Hang Yu. Bias-based universal adversarial patch attack for automatic check-out. In *ECCV*, 2020.

[Liu *et al.*, 2021] Aishan Liu, Xianglong Liu, Hang Yu, Chongzhi Zhang, Qiang Liu, and Dacheng Tao. Training robust deep neural networks via adversarial noise propagation. *IEEE TIP*, 2021.

[Liu *et al.*, 2023] Aishan Liu, Jun Guo, Jiakai Wang, Siyuan Liang, Renshuai Tao, Wenbo Zhou, Cong Liu, Xianglong Liu, and Dacheng Tao. X-adv: Physical adversarial object attacks against x-ray prohibited item detection. In *USENIX Security*, 2023.

[Lohaus *et al.*, 2020] Michael Lohaus, Michael Perrot, and Ulrike Von Luxburg. Too relaxed to be fair. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 6360–6369. PMLR, 13–18 Jul 2020.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *ICML*, 2018.

[Madry *et al.*, 2017] 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.

[Maheshwari and Perrot, 2022] Gaurav Maheshwari and Michaël Perrot. Fairgrad: Fairness aware gradient descent. *arXiv preprint arXiv:2206.10923*, 2022.

[Mele and many others, 2017 2021] Tom Van Mele and many others. COMPAS: A framework for computational research in architecture and structures. https://doi.org/10.5281/zenodo.2594510, 2017-2021. Accessed: 2023-03-11.

[Paul *et al.*, 2021] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *NeurIPS*, 34:20596–20607, 2021.

[ProPublica, 2016] ProPublica. There's software used across the country to predict future criminals. and it's biased against blacks. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing#disqus_thread, 2016. Accessed: 2023-03-11.

[Pruthi *et al.*, 2020] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *NeurIPS*, 2020.

[Roh *et al.*, 2020] Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

[Romano *et al.*, 2020] Yaniv Romano, Stephen Bates, and Emmanuel J. Candès. Achieving equalized odds by resampling sensitive attributes. In *NeurIPS*, 2020.

[Sarhan *et al.*, 2020] Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *ECCV*, 2020.

[Schaul *et al.*, 2015] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. *arXiv preprint arXiv:1511.05952*, 2015.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[Wang *et al.*, 2020] Zeyu Wang, Klint Qinami, Ioannis Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*, 2020.

[Wang *et al.*, 2022a] Guanchu Wang, Mengnan Du, Ninghao Liu, Na Zou, and Xia Hu. Mitigating algorithmic bias with limited annotations. *arXiv preprint arXiv:2207.10018*, 2022.

[Wang *et al.*, 2022b] Junhao Wang, Lan Zhang, Anran Li, Xuanke You, and Haoran Cheng. Efficient participant contribution evaluation for horizontal and vertical federated learning. In *ICDE*, 2022.

[Wu *et al.*, 2019] Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *The World Wide Web Conference*, pages 3356–3362, 2019.

[Xiao *et al.*, 2023] Yisong Xiao, Aishan Liu, Tianlin Li, and Xianglong Liu. Latent imitator: Generating natural individual discriminatory instances for black-box fairness testing. In *ISSTA*, 2023.

[Xie *et al.*, 2022] Xiaofei Xie, Tianlin Li, Jian Wang, Lei Ma, Qing Guo, Felix Juefei-Xu, and Yang Liu. Npc: N euron p ath c overage via characterizing decision logic of deep neural networks. *TOSEM*, 31(3):1–27, 2022.

[Zhang *et al.*, 2018] 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.