Information-Theoretic Bias Reduction via Causal View of Spurious Correlation

Seonguk Seo[1] **Joon-Young Lee**[2] **Bohyung Han**[1]

[1]ECE & ASRI & GSAI, Seoul National University
[2]Adobe Research
{seonguk, bhhan}@snu.ac.kr jolee@adobe.com

Abstract

We propose an information-theoretic bias measurement technique through a causal interpretation of spurious correlation, which is effective to identify the feature-level algorithmic bias by taking advantage of conditional mutual information. Although several bias measurement methods have been proposed and widely investigated to achieve algorithmic fairness in various tasks such as face recognition, their accuracy- or logit-based metrics are susceptible to leading to trivial prediction score adjustment rather than fundamental bias reduction. Hence, we design a novel debiasing framework against the algorithmic bias, which incorporates a bias regularization loss derived by the proposed information-theoretic bias measurement approach. In addition, we present a simple yet effective unsupervised debiasing technique based on stochastic label noise, which does not require the explicit supervision of bias information. The proposed bias measurement and debiasing approaches are validated in diverse realistic scenarios through extensive experiments on multiple standard benchmarks.

Introduction

Various recognition algorithms based on deep neural networks have achieved remarkable performance improvement by learning useful patterns from a large number of training examples, and have started to be deployed in many real-world applications. However, the objective of the optimization problem is mainly concerned about the final accuracy, which makes trained models vulnerable to unexpected decision rules affected by spurious correlation. Although such decision rules work well on most of training examples, they often lead to poor worst-case generalization performance and make learned models unfair to the examples in under-represented groups. For instance, a few existing works (Kim 2016; Buolamwini and Gebru 2018) have discovered the performance gap across demographic subgroups in real-world face recognition tasks; the accuracy of a model on *darker-skinned women* is often much lower than that on *lighter-skinned men*. This tendency becomes a major weakness in achieving algorithmic fairness and generalizing on unseen test environments with domain or distribution shifts.

Mitigating spurious correlation has recently emerged as an important issue in learning debiased models (Hardt, Price, and Srebro 2016; Zhao et al. 2017; Zhang, Lemoine, and Mitchell 2018; Li and Vasconcelos 2019; Wang et al. 2019; Gong, Liu, and Jain 2020; Sagawa et al. 2020b,a). Given the target and bias variables, they typically suppress spurious correlation by making the algorithm independent of the bias variables while maintaining the predictions to the target variables. To this end, various measurements about algorithmic fairness or bias have been introduced, *e.g.*, demographic parity (Calders, Kamiran, and Pechenizkiy 2009), equality of odds (Hardt, Price, and Srebro 2016), and group-fairness accuracy (Sagawa et al. 2020a; Zhang et al. 2020). Although they have been widely used to maintain fairness, their accuracy- or logit-based bias measurement schemes may lead to inaccurate algorithmic bias measurement. For example, we observe that fine-tuning the last linear classification layer of a model is sufficient to achieve high unbiased accuracy and worst-group accuracy, even without updating the biased representations.

To better quantify algorithmic bias, we first formulate the spurious correlation from a causal point of view. Following the strict definition of fairness, we derive the condition of independence between a bias variable and a predicted target variable, which is evaluated by their mutual information. However, in our causal view, it gives a biased estimation of spurious correlation. To handle the issue, we propose a new bias measurement technique based on *conditional* mutual information using feature representations, which is claimed to quantify the algorithmic bias more accurately while maintaining the original objective, fairness.

Based on the new bias measurment, we propose a debiasing framework to mitigate the algorithmic bias by augmenting a loss term for bias regularization, which is derived by the conditional mutual information. We also introduce a simple yet effective unsupervised debiasing technique by exploiting stochastic label noise, which also prevents a model from capturing spurious correlation even without the prior knowledge of bias information. Our experiments verify that both approaches are helpful for alleviating the algorithmic bias while preserving model accuracy.

The main contributions of our work are summarized as follows.

- We propose an information-theoretic measurement tech-

nique for the amount of algorithmic bias via conditional mutual information on learned feature representations, which is derived from the causal view of spurious correlation.

- Based on the new bias measurement approach, we propose two novel debiasing frameworks; one employs a information-theoretic bias regularization loss and the other is with stochastic label noise even without the supervision of bias information.
- We evaluate our bias measurement scheme and debiasing techniques in various realistic scenarios and achieve promising results on multiple standard benchmarks.

Setup and Preliminaries

Let $(X, Y, Z) \in \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}$ be a triplet representing a joint distribution over the space $(\mathcal{X}, \mathcal{Y}, \mathcal{Z})$, where X is an input variable, Y is a target variable, and Z is a bias variable. We denote a learned model by a function $c \in \mathcal{C}$ mapping $\mathcal{X}$ to $\mathcal{Y}$. The classification function $c : X \rightarrow \hat{Y}$ consists of the feature extractor $g : X \rightarrow F$ and the linear classifier $h : F \rightarrow \hat{Y}$, *i.e.*, $c(\cdot) = h(g(\cdot))$, where $F \in \mathbb{R}^d$ is a feature representation and $\hat{Y} \in \mathcal{Y}$ is the predicted target variable. Following the conventional notations in probability theory, we will use the comma (,) to denote the joint distribution and the semicolon (;) to separate the input arguments of mutual information. For example, $I(X; Y, Z)$ indicates the mutual information between X and the joint distribution of Y and Z.

Group-Fairness Accuracy

In classification tasks, unbiased accuracy, worst-group accuracy, or accuracy disparity are employed to address non-uniform accuracy and evaluate algorithm fairness and robustness in the presence of dataset bias (Sagawa et al. 2020a; Zhang et al. 2020; Koh et al. 2021; Zhang and Sang 2020). Formally, if $Y_i \in \{1, ..., A\}$ and $Z_i \in \{1, ..., B\}$ denote the target and bias values of X_i, respectively, then the unbiased accuracy is given by

$$\frac{1}{AB} \sum_{a,b} \frac{\sum_i \mathbb{1}(c(X_i) = Y_i = a, Z_i = b)}{\sum_i \mathbb{1}(Y_i = a, Z_i = b)}, \tag{1}$$

which indicates the average accuracy over all groups defined by a pair of target and bias values. Other metrics, the worst-group accuracy and the accuracy disparity, are obtained by the worst accuracy of all groups and the discrepancy between the best and the worst group, respectively.

However, the unbiased accuracy may not be a good metric for evaluating the amount of bias in data by itself. Given a baseline model converged sufficiently with *hair color* classification on the CelebA dataset, we fix its feature extractor $f(\cdot)$ and fine-tune its classification layer $h(\cdot)$ with a simple resampling technique to reduce class imbalance problem. We select *gender* as a bias variable for evaluation. Although the features in this model remain unchanged even after the fine-tuning phase, it attains a significant performance gain, 8% points in the unbiased accuracy and 19% points in the worst-group accuracy. This result implies that the accuracy- or logit-based bias measurement methods may not be able to capture the innate bias residing in the features, and that high fairness scores can be achieved by adjusting prediction scores in linear classification layers.

Estimating Mutual Information

To identify the feature-level bias, we first introduce the mutual information, which measures the co-dependence between two variables. For random variables X and Y over the space $\mathcal{X} \times \mathcal{Y}$, the mutual information is defined as

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X), \tag{2}$$

where $H(\cdot)$ denotes the Shannon entropy. The mutual information is also defined by the Kullback-Leibler (KL) divergence between the joint distribution of two random variables and the products of their marginal distributions:

$$I(X; Y) = D_{KL}(P_{(X,Y)} || P_X \otimes P_Y), \tag{3}$$

which implies that the larger the divergence between the joint and the product of the marginals is, the stronger the dependence between X and Y is.

However, its exact computation is tractable only for discrete variables or continuous variables under some constraints. Recently, Belghazi *et al*. (Belghazi et al. 2018) propose a neural estimator of mutual information (MINE) between continuous high-dimensional random variables, by offering a lower bound of the mutual information based on the Donsker-Varadhan representation[1] (Donsker and Varadhan 1975),

$$\begin{aligned} I(X; Y) \geq I_\phi(X; Y) = \sup_{\phi \in \Phi} \mathbb{E}_{P_{(X,Y)}}[f_\phi(x, y)] \\ - \log(\mathbb{E}_{P_X \otimes P_Y}[\exp(f_\phi(x, y))]), \end{aligned} \tag{4}$$

where $f_\phi(\cdot, \cdot)$ is a statistical neural network parameterized by $\phi \in \Phi$. The expectations $\mathbb{E}_{P_{(X,Y)}}$ and $\mathbb{E}_{P_X \otimes P_Y}$ are approximated using empirical sampling from the joint distributions and the products of the marginal distributions, respectively. By maximizing the right-hand side of (5) with respect to ϕ, we can obtain a tighter lower bound, which leads to a more accurate estimation of the mutual information.

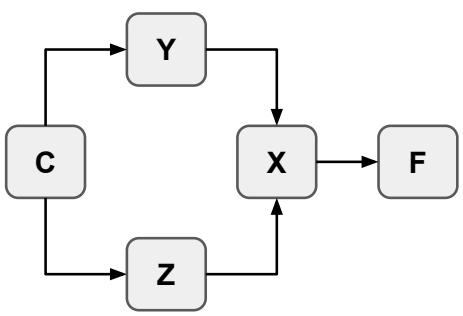
Cobias: Bias Measurement with Conditional Mutual Information

We interpret the spurious correlation in a causal view, and present the bias measurement technique based on the conditional independence derived from the causal view.

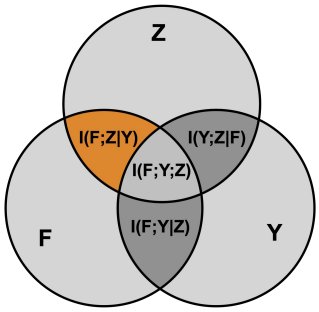
Causal Model of Spurious Correlation

We formulate the causal relationship among input image X, target label Y, bias variable Z, context prior C, and feature representation F using a structural causal model (Spirtes et al. 2000). Figure 1a illustrates the relationship, where the direct link $A \rightarrow B$ indicates that A is the cause of B. Note

[1]The Donsker-Varadhan representation provides the KL-divergence as a supremum over all functions T: $D_{\mathrm{KL}}(P \parallel Q) = \sup_{T:\Omega \rightarrow \mathbb{R}} \mathbb{E}_P[T] - \log(\mathbb{E}_Q[\exp(T)])$.



(a) The diagram of structural causal model. We aim to measure the co-dependence between Z and F via $Z \rightarrow X \rightarrow F$, but it is tricky because there exists a backdoor path $Z \leftarrow C \rightarrow Y \rightarrow X \rightarrow F$.



(b) Venn diagram of information-theoretic measures for F, Y, and Z. The region in orange denotes the mutual information between F and Z, excluding the amount of information that explains Y.

Figure 1: The causal and information-theoretic diagrams of our problem setting.

that we adopt the feature vector F, instead of logits or predicted target values as discussed before.

From the viewpoint of information leakage (Dwork et al. 2012), it is possible to quantify the algorithmic bias by measuring the co-dependence between feature and bias variables via mutual information, *e.g.*, $I(Z;F)$. However, as depicted in Figure 1a, the path between Z and F contains not only the direct path $Z \rightarrow X \rightarrow F$, but also a backdoor path $Z \leftarrow C \rightarrow Y \rightarrow X \rightarrow F$. Because the model is trained to maximize the mutual information between Y and F via $Y \rightarrow X \rightarrow F$, the backdoor path is constructed naturally if there exists the correlation between Y and Z. To focus on the direct path, we should block the backdoor path, which is easily done by conditioning on Y. Since the path $C \rightarrow Y \rightarrow X$ is a serial connection, conditioning on Y blocks the backdoor path and we can now measure the mutual information between Z and F via the direct path only.

Cobias: Conditional Mutual Information as Bias Measurement

Generalizing the mutual information to multivariate cases, we derive the conditional mutual information with three variables X, Y, and Z, which is given by

$$I(X;Y|Z) = \mathbb{E}_{\mathcal{Z}}[D_{KL}(P_{(X|Z,Y|Z)}||P_{X|Z} \otimes P_{Y|Z})]. \quad (5)$$

Note that this equation is an extension of (3) to a conditional setting. The conditional mutual information measures the conditional independence between the relevant variables, *i.e.*, $I(X;Y|Z) = 0 \Leftrightarrow X \perp Y|Z$. Compared to the standard setting, this is particularly important if co-dependence between X and Y changes conditioned on variable Z.

We also consider three variables in our problem setting, which include feature representation vector F, target variable Y, and bias variable Z. As discussed earlier, we concentrate on the mutual information between F and Z conditioned on Y and derive a bias measurement as follows:

$$\text{Cobias} := I(F;Z|Y) = I(F;Z) - I(F;Z;Y) \quad (6)$$
$$= I(F;Z,Y) - I(F;Y), \quad (7)$$

where $I(F;Z;Y)$ is an interaction information among three variables and $(Z,Y) \in \mathbb{R}^2$ denotes the joint variables of Z and Y. We estimate the conditional mutual information by computing the difference between two mutual information $I(F;Z,Y)$ and $I(F;Y)$ as expressed in (7), which is derived from the chain rule of mutual information.

Figure 1b presents a Venn diagram of the information-theoretic measures for those variables. Our bias measurement $I(F;Z|Y)$ quantifies the mutual information between feature and bias variable $I(F;Z)$ excluding the amount of the information that also explains the target variable $I(F;Y;Z)$, which corresponds to the region in orange.

Debiasing Frameworks

This section introduces the proposed two debiasing techniques, which are given by the bias-supervised debiasing regularization and the unsupervised stochastic label noise.

Cobias as Debiasing Regularizer

To mitigate the algorithmic bias, we directly incorporate the proposed bias measurement $I(F;Z|Y)$ as an additional loss term to learn our model. Then, given the extra supervision about bias attributes Z, the final objective function is formulated as

$$\min_{\theta,\psi} \ell(h(F;\psi),Y) + \beta I(F;Z|Y)$$
$$= \min_{\theta,\psi} \ell(h(f(X;\theta);\psi),Y) + \beta I_{\phi}(f(X;\theta);Z|Y), \quad (8)$$

where $\ell(\cdot,\cdot)$ is a cross-entropy loss, θ and ψ denote the parameters of the feature extractor $f(\cdot)$ and the classification layer $h(\cdot)$, respectively, ϕ is the parameters of the mutual information estimator network. The second term in the objective function plays a role as a regularizer to prevent the feature representation from being biased, where β is the hyperparameter of its weight. Note that this regularization term aims to reduce inherent bias within feature representations and has complementary characteristics to the task-specific loss term. Since the loss does not depend on model architecture or algorithm, it can be easily adopted in other existing debiasing frameworks. Because the feature representation $F = f(X;\theta)$ is updated during training, the conditional mutual information $I(F;Z|Y)$ also needs to be re-estimated

each time, which results in the revision of the objective function as

$$\min_{\theta,\psi} \max_{\phi} \ell(h(f(X;\theta);\psi), Y) + \beta I_{\phi}(f(X;\theta); Z|Y). \quad (9)$$

Because the objective function is a minimax problem, we alternate to train and update the parameters (θ, ψ) and ϕ in every epoch.

To compute $I(F; Z|Y)$, we should measure the difference between two estimated mutual information values, $I_{\phi_1}(F; Z, Y)$ and $I_{\phi_2}(F; Y)$ as shown in (7). However, minimizing the difference of the two values obtained from the two different estimators parameterized by ϕ_1 and ϕ_2 may hamper training stability. We sidestep this issue by minimizing a surrogate mutual information, $I(F, Y; Z)$, with respect to θ, instead of $I(F; Z|Y)$. Because $I(f(X;\theta), Y; Z) = I(f(X;\theta); Z|Y) + I(Y; Z)$ and $I(Y; Z)$ is a constant independent of θ, minimizing $I(f(X;\theta); Z|Y)$ with respect to θ is equivalent to minimizing $I(f(X;\theta), Y; Z)$. This enables us to train the network based on (9) using a single estimator network and consequently facilitates stable training.

Stochastic Label Noise

We now introduce a simple yet effective unsupervised debiasing approach exploiting synthetic label noise. Stochastic label noise perturbs the target label from its true class to any other class with a noise rate ρ on each mini-batch independently. Let Y and $\widetilde{Y}$ be the true and noisy target labels, respectively. Then, $p(\widetilde{Y} = k|Y = y) = \frac{\rho}{K-1}$ for all $k \neq y$ and $1 - \rho$ for $k = y$, where K is the number of classes. The objective function is given by

$$\min_{\theta,\psi} \ell(h(f(X;\theta);\psi), \widetilde{Y}). \quad (10)$$

Stochastic label noise, even without any explicit supervision of bias information, mitigates the algorithmic bias while minimizing the performance degradation in the classification for target variable. Intuitively, this is because spurious correlation is more susceptible to label noise than true causation. Label noise gives an implicit ensemble effect (Xie et al. 2016), which helps to find invariant properties between a feature representation and its target label. Because spurious correlation does not appear to be stable properties (Arjovsky et al. 2019; Woodward 2005), label noise helps to absorb spurious correlation and mitigate algorithmic bias.

From the perspective of information theory, it is natural that injecting label noise reduces the mutual information between target variable Y and bias variable Z, *i.e.*, $I(Z; \widetilde{Y}) < I(Z; Y)$. In addition, by adding stochastic label noise, the reduction of mutual information between Y and Z turns out to be more significant than that of Y's entropy. According to our experiment, the ratio $R := I(Z; \widetilde{Y})/I(Y; \widetilde{Y})$ is 0.187 in the absence of label noise with $Y =$ "*hair color*" and $Z =$ "*gender*" in the CelebA dataset. However, it decreases when we add several different levels of label noise, and the tendency is consistent in other (Y, Z) pairs if they are correlated. This observation implies that injecting label noise is helpful for alleviating spurious correlation without affecting classification performance. Note that this label perturbation method is orthogonal to existing debiasing frameworks and can be applied to them with no modification.

Experiments

Experimental Setup

Datasets We conduct experiments on the three standard benchmarks: CelebA (Liu et al. 2015), Waterbirds (Sagawa et al. 2020a), and FairFace (Kärkkäinen and Joo 2021). CelebA is a large-scale face dataset composed of 202,599 celebrity images with 40 attributes. This dataset is available for non-commercial research purposes. We follow the original train-val-test split (Liu et al. 2015) throughout the experiments. Waterbirds (Sagawa et al. 2020a) is a synthesized dataset with 4,795 training examples, which are created by combining bird images in the CUB dataset (Wah et al. 2011) and background images from the Places dataset (Zhou et al. 2017). Following the setup in (Sagawa et al. 2020a), each image has two attributes; one is the type of bird, {waterbird, landbird} and the other is the background place, {water, land}. FairFace (Kärkkäinen and Joo 2021) is a recently proposed face image dataset containing 108,501 images, which are collected from the YFCC-100M Flickr dataset (Thomee et al. 2016). The dataset has seven race groups[2], nine age groups, and two gender groups.

Implementation details We use the ResNet-18 (He et al. 2016) pretrained on ImageNet (Deng et al. 2009) as our backbone network for all experiments. We train our models using the stochastic gradient descent method with the Adam optimizer for 50 epoch. The learning rate is 1×10^{-4}, and the batch size is 256. We set a weight decay to 1×10^{-2} for Waterbirds and 1×10^{-4} for the other two datasets. The weight of the bias regularizer β is fixed to 5. The noise rate ρ is set to 0.1 for Waterbirds and 0.2 for the rest. For all methods, we leverage a simple resampling technique based on the class size to alleviate the class imbalance issue. Our algorithms are implemented in the Pytorch (Paszke et al. 2019) framework and all experiments are conducted on a single unit of NVIDIA Titan XP GPU.

Evaluation metrics In addition to average accuracy, we evaluate all the compared algorithms with three main metrics, Cobias, unbiased accuracy, and worst-group accuracy, to provide a comprehensive view of the algorithmic bias. We also adopt other fairness metrics such as bias amplification (BA) (Zhao et al. 2017), equalized opportunity difference (EO), and disparate impact (DI) (Hardt, Price, and Srebro 2016), where low values are preferred for the extra metrics.

Results

CelebA Table 1 presents the experiment results on the test split of the CelebA dataset. Among all attributes, we choose *blond hair*, *pale skin*, and *smiling* as target variables while *gender* is used as the bias variable, setting up three different classification tasks. We employ two baseline algorithms,

[2] *White*, *Black*, *Indian*, *East Asian*, *Southeast Asian*, *Middle Eastern*, and *Latino*.

Table 1: Experimental results of our debiasing frameworks on the test split of the CelebA dataset. We set *gender* attribute to the bias variable.

	Target	Cobias	Unbiased Acc.	Worst-group Acc.	BA	EO	DI	Average Acc.
ERM	Hair Color	0.435	84.1	53.2	0.014	0.48	0.113	**95.1**
ERM + label noise	Hair Color	0.286	87.8	66.9	**-0.002**	0.35	**0.036**	93.8
ERM + bias regularizer	Hair Color	**0.111**	**88.1**	**67.4**	0.001	**0.31**	0.039	94.5
Group DRO (Sagawa et al. 2020a)	Hair Color	0.409	89.4	77.6	0.005	0.17	0.044	**94.2**
Group DRO + label noise	Hair Color	0.270	90.6	84.0	**-0.004**	**0.07**	0.030	93.5
Group DRO + bias regularizer	Hair Color	**0.045**	**90.7**	**85.4**	0.002	0.09	**0.018**	94.0
ERM	Pale Skin	0.387	82.7	59.6	0.012	0.38	0.273	**95.5**
ERM + label noise	Pale Skin	0.192	**87.6**	**73.8**	**0.002**	**0.19**	**0.049**	94.7
ERM + bias regularizer	Pale Skin	**0.066**	86.1	68.7	0.006	0.28	0.153	95.2
Group DRO (Sagawa et al. 2020a)	Pale Skin	0.346	88.8	84.6	0.005	0.11	0.109	93.8
Group DRO + label noise	Pale Skin	0.189	**90.3**	**86.4**	**0.001**	**0.06**	**0.002**	**94.2**
Group DRO + bias regularizer	Pale Skin	**0.114**	89.1	85.5	0.003	0.10	0.063	93.9
ERM	Smiling	0.126	92.1	88.6	0.002	**0.01**	**0.008**	92.5
ERM + label noise	Smiling	0.017	**92.4**	88.7	**-0.015**	0.10	0.057	**92.8**
ERM + bias regularizer	Smiling	**0.011**	**92.4**	**89.0**	-0.011	0.08	0.038	**92.8**
Group DRO (Sagawa et al. 2020a)	Smiling	0.130	92.1	89.6	0.003	0.02	0.012	92.2
Group DRO + label noise	Smiling	**0.019**	92.5	**90.1**	-0.003	**0.01**	**0.011**	92.6
Group DRO + bias regularizer	Smiling	0.024	**92.7**	89.9	**-0.005**	**0.01**	0.019	**92.8**

Table 2: Experimental results on the test split of the Waterbirds dataset, where bird type and background place are set to target and bias variables, respectively.

	Cobias	Unbiased	Worst	Average
ERM	0.482	81.4	52.3	83.6
ERM + label noise	0.413	83.0	58.2	83.1
ERM + bias regularizer	**0.107**	**84.3**	**72.2**	**88.4**
Group DRO	0.376	86.3	72.4	85.7
Group DRO + label noise	0.348	**86.7**	70.6	84.7
Group DRO + bias regularizer	**0.098**	**86.7**	**76.9**	**87.4**

Table 3: Experimental results on the test split of the Fairface dataset, where *age* and *race* are set to target and bias variables, respectively.

	Cobias	Unbiased	Worst	Average
ERM	0.019	47.6	16.9	52.1
ERM + label noise	0.017	48.7	**18.6**	53.2
ERM + bias regularizer	**0.004**	**49.6**	18.5	**53.4**
Group DRO	0.015	48.4	18.8	51.5
Group DRO + label noise	0.004	49.2	16.8	52.5
Group DRO + bias regularizer	**0.002**	**49.4**	**19.3**	**53.6**

including empirical risk minimization (ERM) and group distributionally robust optimization (Group DRO) (Sagawa et al. 2020a), to which the proposed two debiasing methods are applied. As shown in Table 1, incorporating the bias regularization loss significantly reduces Cobias and other fairness scores, while preserving the average classification accuracy. This implies that the proposed bias measurement is meaningful for identifying the algorithmic bias. We also observe that applying stochastic label noise is effective to mitigate the algorithmic bias even without extra supervision, which validates our claim that spurious correlation is more sensitive to label noise than true causation. Note that Group DRO yields high bias scores in spite of its high group-fairness accuracy, which implies that it may adjust the parameters for bias-related features in the classification layer instead of learning debiased representations.

Waterbirds The results on the Waterbirds dataset are presented in Table 2, where the bird type and the background place are set to target and bias variables, respectively. According to our experiments, the proposed debiasing frameworks consistently produce promising results in this small-scale synthesized dataset. Compared to the CelebA dataset, injecting label noise is less effective to reduce Cobias in this dataset, which is partly due to its small size.

FairFace We also evaluate our frameworks on the FairFace dataset and demonstrate the results in Table 3, where the target is *age* and the bias is *race*. Our frameworks still improves both Cobias and group-fairness accuracy when combined with both ERM and Group DRO algorithms. Note that the Cobias scores are particularly small because the FairFace dataset is constructed with an emphasis of balanced race composition, as stated in (Kärkkäinen and Joo 2021).

Analysis

Domain generalization scenario We extend our task to domain generalization scenario, where training and test sets belong to different domains. This scenario is realistic because domain and distribution shifts may occur simultaneously. To build this setup, we introduce another variable, other than for target or bias attributes, which is for domain attribute. The examples in the training and test datasets are not supposed to have the same value for the domain variable. For example, if we have a domain attribute, *Young*, then training examples may be composed of the images with the old (*Young* = *false*) while test dataset may contain the images of young people only (*Young* = *true*). Since the domain attribute, *Young* in this case, is correlated to both the target and bias variables, it affects the group distributions in the training and test sets significantly and they are not con-

Table 4: Experimental results in domain generalization setting on the CelebA dataset, where the bias variable is fixed to *gender* for all settings.

	Target	Domain (train)	Domain (test)	Cobias	Unbiased Acc.	Worst-group Acc.	Average Acc.
ERM	Hair Color	Old	Young	0.432	81.5	43.5	93.8
ERM + label noise	Hair Color	Old	Young	0.367	88.1	68.3	**93.9**
ERM + bias regularizer	Hair Color	Old	Young	**0.272**	**90.4**	**79.7**	93.0
Group DRO (Sagawa et al. 2020a)	Hair Color	Old	Young	0.387	86.2	68.1	**94.2**
Group DRO + label noise	Hair Color	Old	Young	0.171	88.6	77.4	93.9
Group DRO + bias regularizer	Hair Color	Old	Young	**0.061**	**89.5**	**78.6**	94.1
ERM	Hair Color	Slim	Chubby	0.234	75.6	29.5	**98.1**
ERM + label noise	Hair Color	Slim	Chubby	0.107	82.7	**54.5**	97.9
ERM + bias regularizer	Hair Color	Slim	Chubby	**0.073**	**84.5**	54.2	97.9
Group DRO (Sagawa et al. 2020a)	Hair Color	Slim	Chubby	0.192	84.5	67.6	96.5
Group DRO + label noise	Hair Color	Slim	Chubby	0.062	**87.2**	64.2	96.8
Group DRO + bias regularizer	Hair Color	Slim	Chubby	**0.029**	87.1	**69.7**	**96.9**

Table 5: Ablative results for the weight parameter β of the bias regularization loss on the CelebA dataset.

Bias weight	Cobias	Unbiased Acc.	Worst-group Acc.
0	0.435	84.1	53.2
1	0.202	84.7	58.3
2	0.183	87.2	62.1
5	**0.111**	**88.1**	**67.4**

Table 6: Ablative results for the noise rate ρ in the stochastic label noise on the CelebA dataset.

Noise rate	Cobias	Unbiased Acc.	Worst-group Acc.
0.0	0.435	84.1	53.2
0.1	0.362	87.3	65.7
0.2	0.286	**87.8**	**66.9**
0.4	**0.209**	86.1	60.6

sistent no longer. Table 4 presents the domain generalization results with two domain attributes, *Young* and *Chubby*, where our frameworks outperform the baselines consistently in the presence of domain shift. We also observe that ERM with the bias regularizer often gives better results than the original Group DRO method. This results validate that our frameworks are particularly effective when domain and distribution shifts exist at the same time.

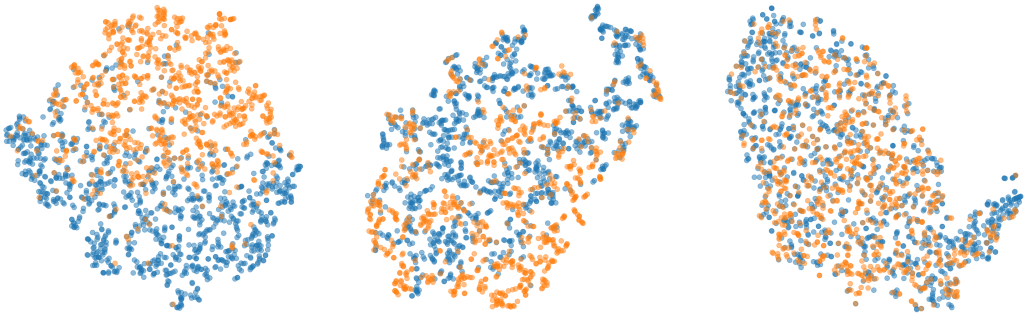
Ablation study To validate the effectiveness of the bias regularization loss and the stochastic label noise scheme, we perform the ablation study on the weight parameter β and the noise rate ρ. Table 5 and 6 show the ablative results on the CelebA dataset with *hair color* and *gender* for target and bias variables, respectively. The results show that increasing the weight of the bias regularization loss and the noise rate improves both Cobias and group-fairness accuracy within sufficiently wide ranges.

Feature visualization Figure 2 illustrates the t-SNE visualizations (Van der Maaten and Hinton 2008) of the feature embeddings of the samples in the CelebA test split given by *hair color* classification. For simplicity, we visualize only the examples with blond hair. Blue and orange colors denote female and male, values of the *gender* bias variable, respectively. Figure 2a, even without using the bias information (*gender*) for training, shows that the examples drawn from a bias group are distinguishable from those from the other bias group. Compared to vanilla ERM, as illustrated in Figure 2b and 2c, ERMs with bias regularization and stochastic label noise successfully make the examples with different bias attributes confused on the feature embedding space, which is desirable for learning debiased representations.

Comparison to logit-based bias regularizer To analyze the effectiveness of our feature-level bias measurement, we compare the feature-based bias regularizer $I(F; Z|Y)$ in (9) with logit-based bias regularizer $I(\hat{Y}; Z|Y)$, where $\hat{Y} = h(F) \in \mathcal{Y}$ is a predicted target variable from a classifier based on a fully connected layer, $h(\cdot)$. Table 7 presents the results from the two regularizers, where the feature-based approach outperforms its logit-based counterpart significantly in terms of both Cobias and group-fairness accuracy. This means that exploiting feature representation would be more effective to identify and mitigate the algorithmic bias than logits.

Related Work

Facial recognition datasets often contain inevitable biases due to their insufficiently controlled data collection process, and consequently, recognition algorithms tend to inherit and even amplify the dataset bias. To address this issue, numerous debiasing frameworks have been proposed for identifying and mitigating the potential risks posed by dataset or algorithmic bias. These frameworks can be categorized in pre-processing (Li and Vasconcelos 2019; Sagawa et al. 2020b; Kamiran and Calders 2012), in-processing (Sagawa et al. 2020a; Sohoni et al. 2020; Wang et al. 2019; Zhang, Lemoine, and Mitchell 2018; Gong, Liu, and Jain 2020; Seo, Lee, and Han 2021; Ragonesi et al. 2021; Wang et al. 2020; Guo et al. 2020), and post-processing (Hardt, Price, and Srebro 2016; Zhao et al. 2017) ones. Pre-processing techniques transform the data distribution to keep the training data balanced across groups, where dataset resampling or reweighting methods (Li and Vasconcelos 2019; Sagawa et al. 2020b; Kamiran and Calders 2012) are usually exploited to balance the distribution by under-sampling the



(a) ERM (b) ERM + label noise (c) ERM + bias regularizer

Figure 2: The t-SNE plots of feature representations from the ResNet-18 models trained with ERM and our debiasing frameworks on the CelebA dataset using the *hair color* classification. We visualize the distribution of samples who have the same target value (blond hair). Blue and orange colors denote different gender values (female and male, respectively).

Table 7: Comparison to the logit-based bias regularization loss on the CelebA dataset with *gender* bias.

Bias regularizer type	Target	Cobias	Unbiased Acc.	Worst-group Acc.	Average Acc.
logit-based	Hair Color	0.219	84.9	62.7	**94.6**
feature-based (Ours)	Hair Color	**0.111**	**88.1**	**67.4**	94.5
logit-based	Pale Skin	0.182	84.0	60.8	**95.7**
feature-based (Ours)	Pale Skin	**0.066**	**86.1**	**68.7**	95.2

majority or over-sampling the minority classes. Debiasing through in-processing aims to build algorithms that can learn fair representation, by taking advantage of adversarial training (Wang et al. 2019; Zhang, Lemoine, and Mitchell 2018), representation disentanglement (Gong, Liu, and Jain 2020; Ragonesi et al. 2021), or robust optimization (Sagawa et al. 2020a; Sohoni et al. 2020; Seo, Lee, and Han 2021). Post-processing methods modify the predicted outputs to meet fairness criterion, mainly by calibrating the outputs (Hardt, Price, and Srebro 2016; Zhao et al. 2017). Our debiasing frameworks belong to in-processing methods, which adopts a bias regularization loss or injecting label noise during training to reduce the algorithmic bias. The bias regularization loss is somewhat similar to (Ragonesi et al. 2021), but the major difference is that we introduce *conditional* mutual information from the structural causal model. The use of unconditional mutual information as a bias regularizer makes it difficult to achieve our goal, learning the relationship between features and target variables.

Mutual information is an information-theoretic quantity to measure the relationship between two variables. Because it can capture non-linear dependencies between the variables, the mutual information is widely used in various areas, including unsupervised representation learning (Comon 1994; Tishby, Pereira, and Bialek 2000; Oord, Li, and Vinyals 2018; Hjelm et al. 2019; Sun et al. 2020), generative models (Chen et al. 2016; Qian and Cheung 2019), reinforcement learning (Oord, Li, and Vinyals 2018), and fair supervised learning (Kamishima et al. 2012; Fukuchi, Kamishima, and Sakuma 2015; Ragonesi et al. 2021). However, the exact computation of mutual information between continuous variables is basically not tractable. There exists some non-parametric estimators based on kernel density estimation (Kwak and Choi 2002; Suzuki et al. 2008) to deal with continuous variables, but those are not scalable with high-dimensional or large-scale data. To overcome this limitation, recent works on mutual information estimation focus on training neural network to represent its variational lower bound (Belghazi et al. 2018; Lin et al. 2019; Poole et al. 2019). Our bias measurement takes also advantage of the variational lower bound (Donsker and Varadhan 1975; Belghazi et al. 2018) of conditional mutual information.

Conclusion

We proposed an information-theoretic bias measurement which can identify the feature-level algorithmic bias from the causal view of spurious correlation. Based on the new measurement approach, we presented two types of debiasing frameworks with bias regularizer or label noise, each of which can be utilized with or without explicit knowledge of bias information, respectively. We demonstrated the effectiveness and versatility of proposed frameworks on multiple standard benchmarks. We also conducted a detailed analysis of our measurement and frameworks via extensive ablation studies with more realistic scenarios.

Acknowledgement

This work was partly supported by Samsung Advanced Institute of Technology, the Bio & Medical Technology Development Program of the National Research Foundation (NRF) funded by the Korea government (MSIT) [No. 2021M3A9E4080782], and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [No.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)].

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Belghazi, M. I.; Baratin, A.; Rajeshwar, S.; Ozair, S.; Bengio, Y.; Courville, A.; and Hjelm, D. 2018. Mutual information neural estimation. In *ICML*.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM FAccT*.

Cadene, R.; Dancette, C.; Cord, M.; Parikh, D.; et al. 2019. Rubi: Reducing unimodal biases for visual question answering. In *NeurIPS*.

Calders, T.; Kamiran, F.; and Pechenizkiy, M. 2009. Building classifiers with independency constraints. In *ICDM*.

Chen, X.; Duan, Y.; Houthooft, R.; Schulman, J.; Sutskever, I.; and Abbeel, P. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *NIPS*.

Comon, P. 1994. Independent component analysis, a new concept? *Signal processing*, 36(3): 287–314.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.

Donsker, M. D.; and Varadhan, S. S. 1975. Asymptotic evaluation of certain Markov process expectations for large time, I. *Communications on Pure and Applied Mathematics*, 28(1): 1–47.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *ITCS*.

Fukuchi, K.; Kamishima, T.; and Sakuma, J. 2015. Prediction with model-based neutrality. *IEICE TRANSACTIONS on Information and Systems*, 98(8): 1503–1516.

Gong, S.; Liu, X.; and Jain, A. K. 2020. Jointly de-biasing face recognition and demographic attribute estimation. In *ECCV*.

Guo, J.; Zhu, X.; Zhao, C.; Cao, D.; Lei, Z.; and Li, S. Z. 2020. Learning meta face recognition in unseen domains. In *CVPR*.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In *NIPS*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Kamiran, F.; and Calders, T. 2012. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer. In *ECML PKDD*.

Kärkkäinen, K.; and Joo, J. 2021. Fairface: Face attribute dataset for balanced race, gender, and age. In *WACV*.

Kim, P. T. 2016. Data-driven discrimination at work. *Wm. & Mary L. Rev.*, 58: 857.

Koh, P. W.; Sagawa, S.; Marklund, H.; Xie, S. M.; Zhang, M.; Balsubramani, A.; Hu, W.; Yasunaga, M.; Phillips, R. L.; Gao, I.; Lee, T.; David, E.; Stavness, I.; Guo, W.; Earnshaw, B. A.; Haque, I. S.; Beery, S.; Leskovec, J.; Kundaje, A.; Pierson, E.; Levine, S.; Finn, C.; and Liang, P. 2021. WILDS: A Benchmark of in-the-Wild Distribution Shifts. *arXiv*.

Kwak, N.; and Choi, C.-H. 2002. Input feature selection by mutual information based on Parzen window. *TPAMI*, 24(12): 1667–1671.

Li, Y.; and Vasconcelos, N. 2019. Repair: Removing representation bias by dataset resampling. In *CVPR*.

Lin, X.; Sur, I.; Nastase, S. A.; Divakaran, A.; Hasson, U.; and Amer, M. R. 2019. Data-efficient mutual information neural estimator. *arXiv preprint arXiv:1905.03319*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep learning face attributes in the wild. In *ICCV*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Poole, B.; Ozair, S.; Van Den Oord, A.; Alemi, A.; and Tucker, G. 2019. On variational bounds of mutual information. In *ICML*.

Qian, D.; and Cheung, W. K. 2019. Enhancing variational autoencoders with mutual information neural estimation for text generation. In *EMNLP*.

Ragonesi, R.; Volpi, R.; Cavazza, J.; and Murino, V. 2021. Learning unbiased representations via mutual information backpropagation. In *CVPR Workshop*.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020a. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*.

Sagawa, S.; Raghunathan, A.; Koh, P. W.; and Liang, P. 2020b. An Investigation of Why Overparameterization Exacerbates Spurious Correlations. In *ICML*.

Seo, S.; Lee, J.-Y.; and Han, B. 2021. Unsupervised Learning of Debiased Representations with Pseudo-Attributes. *arXiv preprint arXiv:2108.02943*.

Sohoni, N.; Dunnmon, J.; Angus, G.; Gu, A.; and Ré, C. 2020. No Subclass Left Behind: Fine-Grained Robustness in Coarse-Grained Classification Problems. In *NeurIPS*.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, prediction, and search*. MIT press.

Sun, F.-Y.; Hoffmann, J.; Verma, V.; and Tang, J. 2020. Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization. In *ICLR*.

Suzuki, T.; Sugiyama, M.; Sese, J.; and Kanamori, T. 2008. Approximating mutual information by maximum likelihood density ratio estimation. In *PMLR*.

Thomee, B.; Shamma, D. A.; Friedland, G.; Elizalde, B.; Ni, K.; Poland, D.; Borth, D.; and Li, L.-J. 2016. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2): 64–73.

Tishby, N.; Pereira, F. C.; and Bialek, W. 2000. The information bottleneck method. *arXiv preprint physics/0004057*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset.

Wang, T.; Zhao, J.; Yatskar, M.; Chang, K.-W.; and Ordonez, V. 2019. Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. In *ICCV*.

Wang, Z.; Qinami, K.; Karakozis, I. C.; Genova, K.; Nair, P.; Hata, K.; and Russakovsky, O. 2020. Towards fairness in visual recognition: Effective strategies for bias mitigation. In *CVPR*.

Woodward, J. 2005. *Making things happen: A theory of causal explanation*. Oxford university press.

Xie, L.; Wang, J.; Wei, Z.; Wang, M.; and Tian, Q. 2016. Disturblabel: Regularizing cnn on the loss layer. In *CVPR*.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *AAAI*.

Zhang, M.; Marklund, H.; Gupta, A.; Levine, S.; and Finn, C. 2020. Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift. *arXiv preprint arXiv:2007.02931*.

Zhang, Y.; and Sang, J. 2020. Towards Accuracy-Fairness Paradox: Adversarial Example-based Data Augmentation for Visual Debiasing. *arXiv preprint arXiv:2007.13632*.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457*.

Zhou, B.; Lapedriza, A.; Khosla, A.; Oliva, A.; and Torralba, A. 2017. Places: A 10 million image database for scene recognition. *TPAMI*, 40(6): 1452–1464.