Learning Bias-Invariant Representation by Cross-Sample Mutual Information Minimization

Wei Zhu[1], Haitian Zheng[1], Haofu Liao *[2], Weijian Li[1], and Jiebo Luo[1]

[1]University of Rochester
[2]Amazon Web Services
{wzhu15,hzheng15,wli69}@ur.rochester.edu, liaohaofu@gmail.com, jluo@cs.rochester.edu

Abstract

Deep learning algorithms mine knowledge from the training data and thus would likely inherit the dataset's bias information. As a result, the obtained model would generalize poorly and even mislead the decision process in real-life applications. We propose to remove the bias information misused by the target task with a cross-sample adversarial debiasing (CSAD) method. CSAD explicitly extracts target and bias features disentangled from the latent representation generated by a feature extractor and then learns to discover and remove the correlation between the target and bias features. The correlation measurement plays a critical role in adversarial debiasing and is conducted by a cross-sample neural mutual information estimator. Moreover, we propose joint content and local structural representation learning to boost mutual information estimation for better performance. We conduct thorough experiments on publicly available datasets to validate the advantages of the proposed method over state-of-the-art approaches.

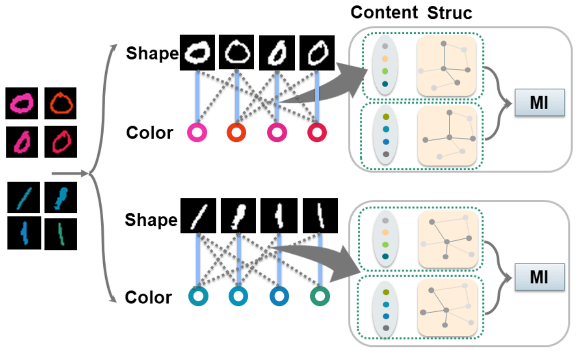

Figure 1. A brief illustration of CSAD for a color-biased binary classification task. Our objective is to obtain a color-invariant digit classifier. Given the i-th training sample x_i as a red "0", most existing methods eliminate the correlation between h_i and red information from the i-th sample (solid line), where h_i is the extracted features. By contrast, CSAD could reduce the correlation between h_i and various red colors extracted from other samples. A cross-sample mutual information estimator measures the correlation with joint content and local structural representation.

1. Introduction

Modern machine learning is built on collected and contributed data. However, real-world data inevitably contains noise and bias and may not be well-distributed. Such flawed datasets may make the learned model unreliable and pose threats to the learned model's generalization capacity to unseen data. This problem is particularly crucial for medical and healthcare-related applications [10]. For example, Parkinson's Disease (PD) is associated with age, and the PD patients are primarily composed of older people in the related datasets [41, 8, 27]. A model learned on these datasets may predict PD by the age of patients instead of the symptoms of the disease. As a result, the age bias renders the learned model hardly useful for real-life disease diagnosis and analysis.

Several methods have been proposed to learn to remove the dataset bias [22, 1, 36, 51, 30, 5]. Among them, some methods regularize the model to not learn bias with additional regularization terms [30, 5], and others learn to eliminate the learned bias information by adversarial learning [22, 36, 51]. Our work follows the

*This work was done when Haofu Liao was at the University of Rochester.

latter and the bias elimination is often conducted by minimizing the correlation between the extracted features and the bias label. The correlation measurement plays a critical role in adversarial debiasing and is often fulfilled by recently proposed neural mutual information estimators [17, 6, 21]. In particular, [22] proposes to adversarially discover and remove the bias information by adding a gradient reversal layer between the feature extractor and the bias branch. [36] adversarially learns to mitigate the bias by minimizing the mutual information between the latent representation and the bias label. To conclude, they essentially learn to remove the bias information from the target classifier by eliminating the dependency between the target and the bias information from same training sample. And as a result, these methods are limited to model and reduce the correlation within each training sample and totally neglect the rich cross-sample information. However, we note that the cross-sample information is important and necessary to be taken into consideration for debiasing. For example, as shown in Fig. (1), given the i-th sample x_i as a red "0", it is not enough to only eliminate the correlation between h_i and the red representation extracted from x_i, as the correlation between h_i and the color representation extracted from other red digits will be preserved, where h_i is the representation of x_i. That is, h_i may still be highly biased and is correlated to pink, rose, ruby, etc. The neglect will pose a grave threat to the reliability of the learned correlation measurement and eventually leads to sub-optimal performance for debiasing. Moreover, although the local structural representation is proven to be helpful for correlation learning [52, 3], it is also hard to be incorporated with existing methods [22, 51, 1].

To address the issues as mentioned above, we propose a cross-sample adversarial debiasing (CSAD) method as shown in Fig. 1. To make it possible to utilize the cross-sample and structural information, inspired by recent progress on domain adaptation [34, 33], CSAD first explicitly disentangles the target and the bias representation. Then, CSAD relies on a cross-sample neural mutual information estimator for correlation measurement, which is conducted on the disentangled bias and target representation. This could also avoid potential problems caused by the domain gap between the latent representation and the bias label used by other methods [22, 36]. With the cross-sample information, CSAD could comprehensively eliminate correlation between target and bias information from different samples. Additionally, the explicit disentanglement makes it possible to consider the local structural representation for mutual information estimation. Specifically, we encourage the bias and target representation to have different topological structures captured by Random Walk with Restart [44], which could avoid the bias information of certain sample to be guessed by its neighbors.

We highlight our main contributions as follows:

1. We propose a flexible and general framework for adversarial debiasing that can explicitly disentangle target and bias representation.

2. Based on the proposed framework, we propose cross-sample adversarial debiasing (CSAD). CSAD eliminates the bias information by a cross-sample mutual information estimator that can jointly exert cross-sample content and structural features.

3. We conduct extensive experiments on benchmark datasets, and our method achieves substantial improvement compared to the current state-of-the-art methods.

2. Related Work

2.1. Debiasing and Fairness

Biases exist across race, gender, and age, and they pose threats to machine learning models in diverse tasks, such as image classification [18, 38, 16, 11, 46], representation learning [25, 24, 29, 13], word embedding [53, 7] and visual question answering [12]. A straightforward way to address the problem is to collect [32] or synthesize more data to balance the training set [18, 42, 37]. However, unbiased data could be expensive to collect and impractical to generate for general tasks.

Other methods alleviate the bias through the learning process. Alvi *et al.* avoid to learn the bias by a maximum cross-entropy term [1]. SenSR adopts a variant of individual fairness as a regularizer so that the learned model could satisfy the individual fairness [48]. Similarly, SenSeI achieves the individual fairness through a transport-based regularizer [49]. Zafar *et al.* develop fairness methods based on the decision boundary fairness [50]. DRO regularizes the model by considering worst group performance [40]. Learned-Mixin encourages the model to focus on different patterns with a ensemble framework [12]. ReBias solves a min-max game to encourage the independence between network and biased prediction [5]. Besides, adversarial learning is also adopted for debiasing, and most methods utilize a discriminator to predict the bias label or to estimate the dependency between latent representation and the bias label. For example, Zhang *et al.* trains the discriminator for bias label with the soft assignment of target [51]. Kim *et al.* unlearns the bias extracted by a bias predictor by adopting a gradient reversal layer [22]. Ruggero *et al.* learns to minimize the mutual information between the latent representation and the bias labels [36]. Few of these methods consider cross-sample and structural information that we discussed above.

Recently, several works focus on learning with weak or even no bias supervision [23, 30, 46, 23]. Learning from Failure (LfF) puts more weight on failure samples [30]. Adversarial Reweighted Learning (ARL) [24] adversarially learn a distribution of hard samples. These models may not be robust as they also tend to overfit the noisy samples [24], and the practical significance still need to be validated.

2.2. Mutual Information Estimation

Mutual information is used to measure the dependency between random variables. As the exact value of mutual information is prohibitive to calculate for large scale data, several papers apply neural networks for efficient mutual information estimation [17, 6]. For random variables X and Z, we denote the product of the margin distribution as $P_X P_Z$ and the joint distribution as P_{XZ}, and the mutual information between X and Z could be estimated by a neural network M by training it to distinguish the samples drawn from the joint distribution P_{XZ} and those drawn from the product of marginal distribution $P_X P_Z$, e.g., MINE [6], Deep InfoMax [20], etc. Since we are not interested in the exact value of the mutual information, a lower bound of the mutual information derived from Jensen Shannon Divergence could be formulated as [20]

$$\begin{aligned} I_{\text{JSD}}(X,Z) = \sup E_{P_{XZ}}&[-\text{sp}(-M(x,z))] \\ - E_{P_X P_Z}&[\text{sp}(M(x,z))], \end{aligned} \tag{1}$$

where $\text{sp}(x) = \log(1+\exp(x))$ is the softplus function and M is a neural network. Information Noise-Contrastive Estimation (Info NCE) [31] is also proposed as a normalized mutual information estimator [35].

3. Problem Statement

Formally, given the i-th sample from the training dataset as (x_i, y_i, b_i), where x_i is the input data, y_i is the groundtruth label for the target task, and b_i is the bias label, we first train a target classifier with a feature extractor F that outputs the latent base representation h_i to maximize the performance on the target task. However, since there are correlations between the target task label y and the bias label b, the target task is likely to rely on the bias information to fulfill its objective, and as a result, the optimized model would generalize poorly on unseen bias-free data. In this paper, we aim to remove the bias information extracted by F while preserving the performance for the target task. As a example, for the colored MNIST dataset shown Fig. 1, different digits will be class-wisely painted with similar color in the training set, *e.g.*, red for zero and green for one. The vanilla digit classifier would likely learn to predict digits based on its color. The objective of debiasing is to make the representation h contain only shape information with no color information. In other words, the digit classifier would rely on shape instead of color to fulfill the task and thus would finally lead to a more practical model.

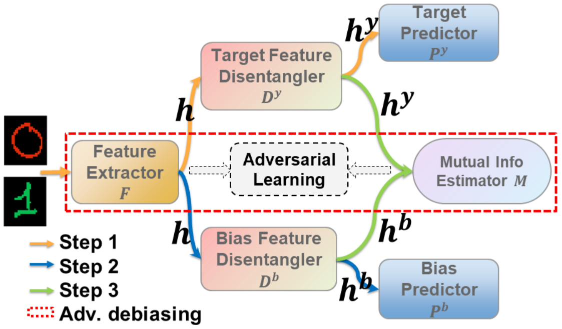

Figure 2. Block diagram of CSAD for the Colored MNIST dataset, and our objective is to obtain color-invariant digit classifier.

4. Method

Our method is illustrated in Fig. (2) and is detailed in Algorithm (1). Our basic idea is to minimize the correlation between the disentangled bias and target representation with a mutual information estimator. We briefly introduce the training procedure as follows: we first pretrain a target classifier composed of F, D^y, and P^y until convergence. The output of F is denoted as the base representation h and the output of D^y is target representation h_y (Step 1); then, we extract the bias representation h_b from h by a bias classifier composed of D^b and P^b. Note we do not update F for bias predication (Step 2); then, we learn the correlation between the bias h^b and the target representations h^y by optimizing the mutual information estimator M (Step 3); at last, we update F to minimize the mutual information estimated by M (Adv. debiasing).

Note that the feature extractor F is only updated to minimize the mutual information estimated by M (Adv.

debiasing) and the target prediction loss (Step 1); it is thus forced to generate bias-invariant features that are still powerful enough for the target task. As a result, after training, the latent representation h extracted by F will contain little information on bias. The bias feature disentangler D^b thus cannot extract useful information h^b that could enable mutual information estimator to associate h_i^y with h_j^b where the i-th and j-th samples share similar bias. The bias branch and mutual information estimator will be discarded at testing time.

4.1. Cross-Sample Adversarial Debiasing

One of the most critical part of our method is to measure the correlation between the target h^y (shape) and bias h^b (color) representations which is fulfilled by a neural mutual information estimator. To conduct the neural estimation, we first need to define the joint and product of marginal distributions between h^y (shape) and h^b (color). As discussed in [45], there is a close relation between mutual information estimation and metric learning. For ease of presentation, we denote the samples drawn from the *joint distribution* $P_{h^y h^b}$ as *positive pairs* and those drawn from the *product of marginal distribution* $P_{h^y} P_{h^b}$ as *negative pairs*. According to recent literature, the pair construction, *i.e.*, the definition of $P_{h^y h^b}$ and $P_{h^y} P_{h^b}$, plays a critical role in mutual information estimation [4]. Usually, the positive pairs could be intuitively constructed by matching different representations of the same sample, while the negative pairs are generated by matching the representation of different samples. For example, Deep InfoMax constructs the positive pairs by matching the global feature and local feature from the same sample [21]; Contrastive Multi-view Coding (CMC) regards different views of the same sample as positive pairs [43]; Contrastive Predictive Coding (CPC) exerts the consecutive of sequential data to construct the positive pairs [31].

Specifically for debiasing, it is common to have the target representation and the bias label of same sample as the positive pairs [22, 36]. However, given the i-*th* sample as a red “0”, these methods only try to extract and reduce the dependency between the target representation h_i^y and the color label “red”, and neglect the dependency between h_i^y and other biases that are similar to the “red” and can be extracted from other samples e.g., pink, rose, light red, and also variants of “red” from other samples. The debiasing models trained without considering these cross-sample information will then omit the potential target-bias correlation, and eventually lead to suboptimal performance. However, it is hard to utilize the cross-sample relation between data especially with same bias label by existing methods [22, 36], since they directly reduce the mutual information between target representation and bias label. By contrast, CSAD conducts debiasing on feature level and explicitly disentangles the target h^y and the bias h^b representation. Then the joint and product of marginal distribution could be easily defined in an across-sample way for debiasing. That is, the positive pairs for the i-th sample are constructed by the target representation h_i^y and bias representation h_b^j, where $(i, j) \in \Omega$ is a set of pairs which share similar bias and i is not necessary to be equal to j. Moreover, as shown in Section 4.2.2, the disentanglement framework also makes it possible to consider the structural representation.

With the definition of the positive and negative pairs, we can conduct neural mutual information estimation by existing estimators e.g., MINE and I_{JSD}. However, these estimators are developed under the assumption that there is only one positive pair per sample, and are thus suboptimal for our case. To consider the multiple positive pairs for each sample, inspired by recent progress on metric learning [47], we propose a cross-sample mutual information estimator for adversarial debiasing as follows

$$\begin{aligned} I_{\text{CS}}(h^y, h^b) = \sup -\log(1 + E_{(i,j)\in\Omega} \exp(-M(h_i^y, h_j^b))) \\ -\log(1 + E_{(i,j)\notin\Omega} \exp(M(h_i^y, h_j^b))), \end{aligned} \tag{2}$$

where $M(h_i^y, h_j^b)$ is implemented by a neural network detailed in Sec. 4.2, and its output denotes the correlation between h_i^y and h_j^b.

We here draw several points for CSAD. First, comparing with existing methods [22, 36], CSAD relies on a disentanglement framework and conducts mutual information estimation on feature level h^y and h^b instead of using the bias label directly. This makes it possible to consider cross-sample correlation and avoid the potential domain gap. Second, the positive pairs (joint distribution) of CSAD are constructed by h_i^y and h_j^b, where $(i, j) \in \Omega$ and i is not necessary to be equal to j. Third, the feature-level debiasing framework also makes it possible to take the local structure of each sample into consideration for mutual information estimation as described later in Section 4.2.2, and jointly adopts content and structural representation yields better performance as shown in our experiments. At last, it is easy to see that the proposed I_{CS} is a lower bound for I_{JSD} shown in Eq. (1). The proof is simple and is based on Jensen’s inequality and the convexity of $-\log(1 + x)$ [9]. Refer to supplemental material for detail. As claimed by [45], maximizing a tighter bound on mutual information does not always lead to better performance. Compared with I_{JSD}, the proposed I_{CS} can automatically reweight training samples [47], and shows superior performance in practice according to our experiments.

4.2. Representation Learning for MI Estimation

This section provides the detail for the implementation of mutual information estimator M with content and local structural representation.

4.2.1 Content Representation Learning

We first demonstrate the proposed I_{CS} with content representation. We implement M with two branches as $M = \{M_y, M_b\}$, and pass h^y and h^b through M_y and M_b respectively to obtain $M_y(h^y)$ and $M_b(h^b)$. Then $M(h_i^y, h_j^b)$ is calculated by the cosine similarity between $M_y(h_i^y)$ and $M_b(h_j^b)$ as

$$M(h_i^y, h_j^b) = \alpha s_c(i,j) = \alpha \frac{M_y^T(h_i^y)M_b(h_j^b)}{\|M_y(h_i^y)\|\|M_b(h_j^b)\|} \quad (3)$$

where α is a learnable scale factor and is initialized as 1 in this paper. We term Eq. (2) with M defined in Eq. (3) as CSAD-Content.

4.2.2 Local Structural Representation Learning

We further introduce local structural representation to enrich the capacity of M. Our intuition is that for the target representation h_i^y of the i-th sample, we would like to have its bias not be guessed by its neighbors. In other words, we would like to encourage h^y and h^b to be different in terms of topological structure, and intuitively, such constraint can provide the clue for learning a stronger M for cross-sample mutual information estimation. For simplicity, we only elaborate on how to learn the local structural feature for h^y, and a similar procedure applies to h^b.

First, we construct an undirected graph for h^y as $G^y =< V^y, E^y >$ over the training samples, here $V^y \in \mathbb{R}^N$ represents the nodes, $E^y \in \mathbb{R}^{N\times N}$ denotes edges, and N is the number of nodes. To obtain the weights for edges E^y, we first calculate the pair-wise cosine similarity among samples in h^y as

$$s_{i,j}^y = \frac{M_y^T(h_i^y)M_y(h_j^y)}{\|M_y(h_i^y)\|\|M_y(h_j^y)\|}, \quad (4)$$

and then obtain a normalized adjacency matrix by softmax. That is, $e_{i,j}^y$, which is the (i,j)-th element of E^y, could then be calculated as

$$e_{i,j}^y = \frac{\exp(\tau s_{i,j}^y)}{\sum_j \exp(\tau s_{i,j}^y)}, \quad (5)$$

where τ is a learnable scale factor and is initialized as 10 in this paper. With the obtained graph G^y, we then apply Random Walk with Restart (RWR) to capture the local structure of each sample [52, 3]. Formally, RWR is conducted as:

$$r_i^{t+1,y} = cE^y r_i^{t,y} + (1-c)a_i, \quad (6)$$

where the $r_i^{t,y} \in \mathbb{R}^N$ is the proximity between the i-th node to all other nodes at the t-th propagation, E^y is the edge and defines the transition probability of the propagation, c is the restart probability and is set to 0.5 in this paper, and $a_i \in \mathbb{R}^N$ is the starting vector with the i-th element set to 1 and 0 for others. We start with $r_i^{0,y} = a_i$ and recursively perform Eq. (6) until convergence. The closed-from converged solution is [44]

$$r_i^y = (1-c)(I - cE_i)^{-1}a_i. \quad (7)$$

At last, we normalize r_i^y as $r_{i,j}^y = \frac{r_{i,j}^y}{\sum_j r_{i,j}^y}$ to have a categorical instance-wise assignment vector. The obtained r_i^y captures the local structure of h_i^y. Likewise, the local structural representation r_i^b is computed in the same fashion for h_i^b.

To obtain the structural similarity $s_s(i,j)$ between h_i^y and h_j^b, since r_i^y and r_j^b are categorical vectors, we calculate the negative symmetric Cross Entropy between them as

$$s_s(i,j) = \frac{1}{2}((r_i^y)^T \log r_j^b) + (r_j^b)^T \log r_i^y)). \quad (8)$$

We term Eq. (2) with M defined in Eq. (8), *i.e.*, $M(h_i^y, h_j^b) = \alpha s_s(i,j)$, as CSAD-Struc. Note that it is infeasible to directly calculate the graph with all training data and we construct the graph G with the samples in the mini-batch.

4.2.3 Joint Content and Structural Representation Learning

We jointly use the content and local structural representation for our mutual information estimator M as

$$M(h_i^y, h_j^b) = \alpha(s_c(h_i^y, s_j^b) + s_s(h_i^y, s_j^b)). \quad (9)$$

By jointly adopting the content and local structural representations, we could provide a more comprehensive estimation of the mutual information. We term Eq. (2) with M defined in Eq. (9) as CSAD. Experimental results show that the joint representation outperforms either CSAD-Content or CSAD-Struc.

4.3. Training Strategy

We present pseudo code of CSAD in Algorithm 1 and please refer to Fig. (2) for an illustration. We omit the pretraining stages for all components and the algorithm to train mutual information estimator, and please refer to supplemental material for more detail. The number of inner-loop iteration K is set as 10 throughout the paper. The target and bias predictors are trained with cross-entropy loss in this paper but can be directly replaced with other loss functions for general tasks. We note that, from Algorithm 1, the feature extractor F is updated only to minimize the target prediction loss (Line 4) and to minimize the mutual information Eq. (2) (Line 14), and will never be updated to minimize the bias prediction loss and maximize the Eq. (2). Therefore, after optimized, the representation learned by the feature extractor will only be able to fulfill the target task and have little information on the bias. In our implementation, we propose a hyper-parameter λ for Step 4 to achieve a balance between fairness and accuracy. As the bias branch and mutual information estimator will be discarded at testing time, our method introduces no extra cost for inference.

Algorithm 1 Learning Bias-Invariant Representation

```
Input: Training data $x = \{(x_i, y_i, b_i)\}$;
1:  while not converge do
2:      Sample a minibatch of data;
3:      # STEP 1: Update Feature Extractor and Target Branch
4:      Update $F$, $D^y$, and $P^y$ to minimize the target predication loss;
5:      # STEP 2: Update Bias Branch
6:      for k = 1, ..., K do
7:          Update $D^b$ and $P^b$ to minimize the bias predication loss;
8:      end for
9:      # STEP 3: Update MI Estimator
10:     for k = 1, ..., K do
11:         Update $M$ to maximize Eq. (2);
12:     end for
13:     # STEP 4: Adversarial Debiasing
14:     Update $F$ to minimize Eq. (2);
15: end while
```

5. Experiments

In this section, we conduct experiments on various datasets to fully demonstrate the effectiveness of the proposed method. Experimental settings and datasets adopted by debiasing and fairness papers are different from each other, and we mainly follow three different settings as [22, 36, 1], [30] and [48, 49]. We conduct experiments on Colored MNIST [22], IMDB face [39], CelebA [28], mPower [8], and Adult [2]. For Colored MNIST, IMDB face, and mPower, we follow the debiasing setting adopted by [22, 36, 1], for CelebA, we follow the setting adopted by [30], and for Adult, we follow the fairness setting adopted by [48, 49]. Among these datasets, Colored MNIST, IMDB Face, and CelebA are image datasets, mPower is a time series dataset, and Adult is a tabular dataset. We run all experiments three times and report the mean accuracy [36]. We implement our method with Pytorch and all experiments are run on a Linux machine with a Nvidia GTX 1080 Ti graphic card.

5.1. Colored MNIST

The Colored MNIST dataset [22] introduces color bias to the standard MNIST dataset [26], and the digits are class-wisely colored for the training set following [22]. Smaller σ means more severely biased training data. We compare CSAD with other debaising methods, including [1], [22] and [36]. The results of other methods are retrieved from their papers. In addition, we compare our approach with three ablation models of our methods, namely AD-JSD, CSAD-Content and CSAD-Struc. For AD-JSD, we adopt the disentanglement framework of CSAD but with I_{JSD} for mutual information estimation Eq.(1) instead of the proposed I_{CS} Eq.(2) with the content representation Eq.(3). For CSAD-content and CSAD-struc, we use content Eq.(3) or structural Eq.(8) features only for mutual information estimation. For CSAD, we adopt the proposed I_{CS} with the joint features Eq. (9). For this dataset, $(i, j) \in \Omega$ if the difference between their color is equal to or smaller than 1 for each channel.

According to Table 1, all variants of CSAD outperform the existing approaches with different σ^2. Notably, our model achieves even more significant improvement for severely biased datasets (smaller σ^2), demonstrating the effectiveness of the disentanglement framework and the necessity to consider cross-sample information. Moreover, the proposed I_{CS} outperforms I_{JSD} by comparing AD-JSD and CSAD-Content with AD-JSD, showing the superiority of the proposed mutual information estimator Eq. (2). Additionally, the proposed CSAD-Struc performs slightly better than CSAD-Content, and CSAD with the joint representation performs favourably against either CSAD-Content or CSAD-Struc, showing that 1) the structural representation could benefit the learning process for adversar-

Table 1. Evaluation results on the Colored MNIST test set. Smaller variance σ^2 on training set represents severer bias.

Method	Color vairance σ^2						
	0.020	0.025	0.030	0.035	0.040	0.045	0.050
Baseline	0.476	0.542	0.664	0.720	0.785	0.838	0.870
Alvi et al. [1]	0.676	0.713	0.794	0.825	0.868	0.890	0.917
Kim et al. [22]	0.818	0.882	0.911	0.929	0.936	0.954	0.955
Ruggero et al. [36]	0.864	0.925	0.959	0.973	0.975	0.980	0.982
AD-JSD (ours)	0.896	0.937	0.959	0.974	0.975	0.980	0.980
CSAD-Content (ours)	0.933	0.959	0.963	0.976	0.978	0.980	0.983
CSAD-Struc (ours)	0.928	0.955	0.967	0.973	0.980	0.981	0.985
CSAD (ours)	**0.943**	**0.961**	**0.970**	**0.980**	**0.981**	**0.982**	**0.985**

Table 2. Evaluation results on the IMDB face dataset for gender prediction. All networks were evaluated with test set and the other training set.

Method	Trained on EB1		Trained on EB2	
	EB2	Test	EB1	Test
Baseline	0.5986	0.8442	0.5784	0.6975
Alvi et al. [1]	0.6374	0.8556	0.5733	0.6990
Kim et al. [22]	0.6800	0.8666	0.6418	0.7450
Ruggero et al. [36]	0.6840	**0.8720**	0.6310	0.7450
CSAD (ours)	**0.7038**	0.8696	**0.6811**	**0.7865**

Table 3. Evaluation results on CelebA.

Mehtod	Unbaised	Bias-conflicting
Target attribute: *BlondHair*		
Baseline	0.7025	0.5252
Group DRO [40]	0.8424	0.8124
LfF [30]	0.8543	0.8340
CSAD (ours)	**0.8936**	**0.8753**
Target attribute: *HeavyMakeup*		
Baseline	0.6200	0.3375
Group DRO [40]	0.6488	0.5024
LfF [30]	0.6620	0.4548
CSAD (ours)	**0.6788**	**0.5344**

Table 4. Evaluation results on the mPower tapping dataset.

Method	AUC	AP	F1
Baseline	0.735	0.419	0.553
Kim et al. [22]	0.759	0.424	0.572
CSAD (ours)	**0.772**	**0.434**	**0.581**

ial debiasing and 2) jointly considering the content and structural features could lead to better performance.

5.2. IMDB face

The IMDB face dataset [39] is a face image dataset. Following [22, 36], the images are divided into three subsets, namely: Extreme bias 1 (EB1): women aged 0-29, men aged 40+; Extreme bias 2 (EB2): women aged 40+, men aged 0-29; Test set: 20% of the cleaned images aged 0-29 or 40+. As a result, EB1 and EB2 are biased towards the age, since EB1 consists of younger females and older males and EB2 consists of younger males and older females. We adopt a pretrained ResNet18 [19] on ImageNet [15] following [36, 22] as the feature extractor. Moreover, we freeze the BN layer to stabilize the training process. The $(i, j) \in \Omega$ if they share same bias label. We train our models on EB1 (EB2) and evaluate the trained model on EB2 (EB1) and the testing set. Table 2 shows the prediction results. The biased training samples pose a serious threat to the baseline method and make the obtained model generalize poorly on unseen data. By contrast, the models obtained by debiasing methods are much more robust with a age-invariant representation, and the proposed CSAD achieves a much better performance.

5.3. CelebA

CelebA dataset contains 40 attributes of face images. We follow Nam *et al.* to conduct experiments on the official training (162770 samples) and validation (19867 samples) set to respectively predict *BlondHair* and *HeavyMakeup* with the bias attribute as *Male* [30]. To evaluate the performance, we construct unbiased set and bias-conflicting set from original validation set. Following [30], the unbiased set is constructed with all validation data, and we report a weighted average accuracy based on the target-bias pairs. The bias-conflicting set is constructed by the data which have same target and bias values, e.g., *BlondHair-Male* as there are few male with *BlondHair* in the training set. $(i, j) \in \Omega$ if they share same bias label. The results are shown in Table 3. We adopt a pretrained Resnet-18 as the feature extractor with frozen BN layer. We compare our method with Group DRO [40] and LfF [30]. Based on the results,

Table 5. Income prediction of Adult dataset. * denote the results are reproduced by ourselves with a three-layer MLP.

	BA↑	S-Con↑	GR-Con↑	$\text{Gap}_G^{\text{RMS}}$ ↓	$\text{Gap}_R^{\text{RMS}}$ ↓	$\text{Gap}_G^{\text{max}}$ ↓	$\text{Gap}_R^{\text{max}}$ ↓
Baseline	**82.9**	.848	.865	.179	.089	.216	.105
Baseline*	82.7	.844	.831	.182	.087	.212	.110
Project [48]	82.7	.868	**1.00**	.145	.064	.192	.086
Adv. debiasing [51]	81.5	.807	.841	.082	.070	.110	.078
CoCL [14]	79.0	-	-	.163	.080	.201	.109
SenSR [48]	78.9	.934	.984	.068	.055	.087	.067
SenSeI [49]	76.8	**.945**	.963	**.043**	.054	**.053**	.064
CSAD (ours)	80.4	.938	.975	.065	**.042**	.073	**.058**

our method outperforms other approaches on different scenarios.

5.4. mPower

The mPower is collected to develop a smartphone-based remote diagnosis system for PD's patients. Subjects are required to conduct well-designed activities which could reveal the PD's symptoms. Here we conduct adversarial debiasing for the finger tapping task, where patients will tap their phones alternatively with two fingers. The mPower data has a clear bias on age, and detail statistics are provided in supplemental material. We compare our method against baseline and [22]. $(i, j) \in \Omega$ if they share same bias label. As shown in the Table 4, our model improves over the results of the baseline and [22], suggesting that the learned representation of our method is more robust to the age bias.

5.5. Adult

The Adult dataset is a commonly used benchmark in the algorithmic fairness literature [48]. The task is to predict if a subject earns more than $50k per year with the attributes including education, gender, race, etc. We aim to learning an income predictor that is invariant to gender and race which are protected attributes [48]. We first preprocess the dataset following [48], and split the data into 80% training and 20% testing. We compare our method with advanced fairness methods including Project [48], CoCL [14], Adversarial Debiasing [51], SenSR [48], and SenSeI [49]. The results of compared methods are duplicated from [49], and are obtained with a two-layer MLP for the target task. However, recall that CSAD contains three different modules for target task including F, D^y, and P^y, which make the two-layer MLP not applicable for our method, we thus construct a three-layer MLP which has similar performance as shown in Table 5. $(i, j) \in \Omega$ if they share same bias label. The results reported in Table 5 is averaged over 10 different train/val splits following [48]. We use seven evaluation metrics, and all of them are for fairness evaluation except Balanced Accuracy (BA) [48]. Refer to [48] or supplemental material for detail. Overall, according to Table 5, the proposed CSAD outperforms the state-of-the-art fairness methods that are specifically trained with the fairness metric. We note that although SenSeI [49] seems to achieve a better performance in terms of the fairness, this is at the expense of a significant balanced accuracy drop (6.1%) and thus may be less impractical for real life applications. By contrast, CSAD obtains a state-of-the-art performance in terms of both individual and group fairness metrics with a relatively small balanced accuracy drop (2.3%).

6. Conclusions

This paper presents an adversarial debiasing method named CSAD. CSAD is built on a novel disentanglement framework composed of six learnable modules that can respectively extract target and bias features form the input. Then, we adversarially learn to mine and remove the correlation between the target and bias features, and the correlation is measurement by a cross-sample mutual information estimator. We further boost CSAD with a joint structural and content representation. At last, we show a carefully designed training strategy to obtain the debiased model. To validate the effectiveness of the proposed method, we conduct experiments on five datasets with three debiasing benchmark settings.The results show the superior performance of CSAD for various tasks. In the future, we will extend CSAD to deal with incomplete or noisy labels and investigate the explainability and the fairness-accuracy trade-off.

7. Acknowledgements

This research was supported in part by NIH grants R21DE030251-01, P50NS108676, and RF1AG063811-01S2.

References

[1] Mohsan Alvi, Andrew Zisserman, and Christoffer Nellåker. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1, 2, 6, 7

[2] Arthur Asuncion and David Newman. Uci machine learning repository, 2007. 6

[3] Nicolas Aziere and Sinisa Todorovic. Ensemble deep manifold similarity learning using hard proxies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7299–7307, 2019. 2, 5

[4] Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. In *Advances in Neural Information Processing Systems*, pages 15535–15545, 2019. 4

[5] Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pages 528–539. PMLR, 2020. 1, 2

[6] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *International Conference on Machine Learning*, pages 531–540, 2018. 2, 3

[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357, 2016. 2

[8] Brian M Bot, Christine Suver, Elias Chaibub Neto, Michael Kellen, Arno Klein, Christopher Bare, Megan Doerr, Abhishek Pratap, John Wilbanks, E Ray Dorsey, et al. The mpower study, parkinson disease mobile data collected using researchkit. *Scientific data*, 3(1):1–9, 2016. 1, 6

[9] Stephen Boyd, Stephen P Boyd, and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004. 4

[10] Richard Chen, Filip Jankovic, Nikki Marinsek, Luca Foschini, Lampros Kourtis, Alessio Signorini, Melissa Pugh, Jie Shen, Roy Yaari, Vera Maljkovic, et al. Developing measures of cognitive impairment in the real world from consumer-grade multimodal sensor streams. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2145–2155, 2019. 1

[11] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pages 853–865, 2019. 2

[12] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*, 2019. 2

[13] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. *arXiv preprint arXiv:1906.02589*, 2019. 2

[14] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 120–128, 2019. 8

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7

[16] Prithviraj Dhar, Joshua Gleason, Hossein Souri, Carlos D Castillo, and Rama Chellappa. An adversarial learning algorithm for mitigating gender bias in face recognition. *arXiv preprint arXiv:2006.07845*, 2020. 2

[17] Yixiao Ge, Dapeng Chen, and Hongsheng Li. Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. *arXiv preprint arXiv:2001.01526*, 2020. 2, 3

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. *arXiv preprint arXiv:1811.12231*, 2018. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[20] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *arXiv preprint arXiv:1808.06670*, 2018. 3

[21] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 2, 4

[22] Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9012–9020, 2019. 1, 2, 3, 4, 6, 7, 8

[23] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed

Chi. Fairness without demographics through adversarially reweighted learning. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[24] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. ifair: Learning individually fair data representations for algorithmic decision making. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 1334–1345. IEEE, 2019. 2, 3

[25] Preethi Lahoti, Krishna P Gummadi, and Gerhard Weikum. Operationalizing individual fairness with pairwise fair representations. *arXiv preprint arXiv:1907.01439*, 2019. 2

[26] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2, 2010. 6

[27] Weijian Li, Wei Zhu, E Ray Dorsey, and Jiebo Luo. Predicting parkinson's disease with multimodal irregularly collected longitudinal smartphone data. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1106–1111. IEEE, 2020. 1

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 6

[29] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Aram Galstyan, and Greg Ver Steeg. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, pages 9084–9093, 2018. 2

[30] Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020. 1, 3, 6, 7

[31] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[32] Rameswar Panda, Jianming Zhang, Haoxiang Li, Joon-Young Lee, Xin Lu, and Amit K Roy-Chowdhury. Contemplating visual emotions: Understanding and overcoming dataset bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 579–595, 2018. 2

[33] Xingchao Peng, Zijun Huang, Ximeng Sun, and Kate Saenko. Domain agnostic learning with disentangled representations. *arXiv preprint arXiv:1904.12347*, 2019. 2

[34] Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. *arXiv preprint arXiv:1911.02054*, 2019. 2

[35] Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A Alemi, and George Tucker. On variational bounds of mutual information. *arXiv preprint arXiv:1905.06922*, 2019. 3

[36] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning unbiased representations via mutual information backpropagation. *CoRR*, abs/2003.06430, 2020. 1, 2, 3, 4, 6, 7

[37] Arijit Ray, Karan Sikka, Ajay Divakaran, Stefan Lee, and Giedrius Burachas. Sunny and dark outside?! improving answer consistency in vqa through entailed question generation. *arXiv preprint arXiv:1909.04696*, 2019. 2

[38] Joseph P Robinson, Gennady Livitz, Yann Henon, Can Qin, Yun Fu, and Samson Timoner. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–1, 2020. 2

[39] Rasmus Rothe, Radu Timofte, and Luc Van Gool. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 10–15, 2015. 6, 7

[40] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 2, 7

[41] Patrick Schwab and Walter Karlen. Phonemd: Learning to diagnose parkinson's disease from smartphone data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1118–1125, 2019. 1

[42] Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk–quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8218–8226, 2019. 2

[43] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*, 2019. 4

[44] Hanghang Tong, Christos Faloutsos, and Jia-Yu Pan. Fast random walk with restart and its applications. In *Sixth international conference on data mining (ICDM'06)*, pages 613–622. IEEE, 2006. 2, 5

[45] Michael Tschannen, Josip Djolonga, Paul K Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. *arXiv preprint arXiv:1907.13625*, 2019. 4

[46] Haohan Wang, Zexue He, Zachary C Lipton, and Eric P Xing. Learning robust representations by projecting superficial statistics out. *arXiv preprint arXiv:1903.06256*, 2019. 2, 3

[47] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5022–5030, 2019. 4, 5

[48] Mikhail Yurochkin, Amanda Bower, and Yuekai Sun. Training individually fair ml models with sensitive subspace robustness. In *International Conference on Learning Representations*, 2019. 2, 6, 8

[49] Mikhail Yurochkin and Yuekai Sun. Sensei: Sensitive set invariance for enforcing individual fairness. *arXiv preprint arXiv:2006.14168*, 2020. 2, 6, 8

[50] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017. 2

[51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018. 1, 2, 3, 8

[52] Kai Zhang, Yaokang Zhu, Jun Wang, and Jie Zhang. Adaptive structural fingerprints for graph attention networks. In *International Conference on Learning Representations*, 2019. 2, 5

[53] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019. 2