Gradient Based Activations for Accurate Bias-Free Learning

Vinod K. Kurmi,[*1] **Rishabh Sharma,**[*2] **Yash Vardhan Sharma,**[*2] **Vinay P Namboodiri**[3]

[1] KU Leuven, Belgium[†], [2] IIT Roorkee, India [3] University of Bath, UK
vinod.kurmi@kuleuven.be, {rsharma1, ysharma}@me.iitr.ac.in, vpn22@bath.ac.uk

Abstract

Bias mitigation in machine learning models is imperative, yet challenging. While several approaches have been proposed, one view towards mitigating bias is through adversarial learning. A discriminator is used to identify the bias attributes such as gender, age or race in question. This discriminator is used adversarially to ensure that it cannot distinguish the bias attributes. The main drawback in such a model is that it directly introduces a trade-off with accuracy as the features that the discriminator deems to be sensitive for discrimination of bias could be correlated with classification. In this work we solve the problem. We show that a biased discriminator can actually be used to improve this bias-accuracy tradeoff. Specifically, this is achieved by using a feature masking approach using the discriminator's gradients. We ensure that the features favoured for the bias discrimination are de-emphasized and the unbiased features are enhanced during classification. We show that this simple approach works well to reduce bias as well as improve accuracy significantly. We evaluate the proposed model on standard benchmarks. We improve the accuracy of the adversarial methods while maintaining or even improving the unbiasness and also outperform several other recent methods.

Introduction

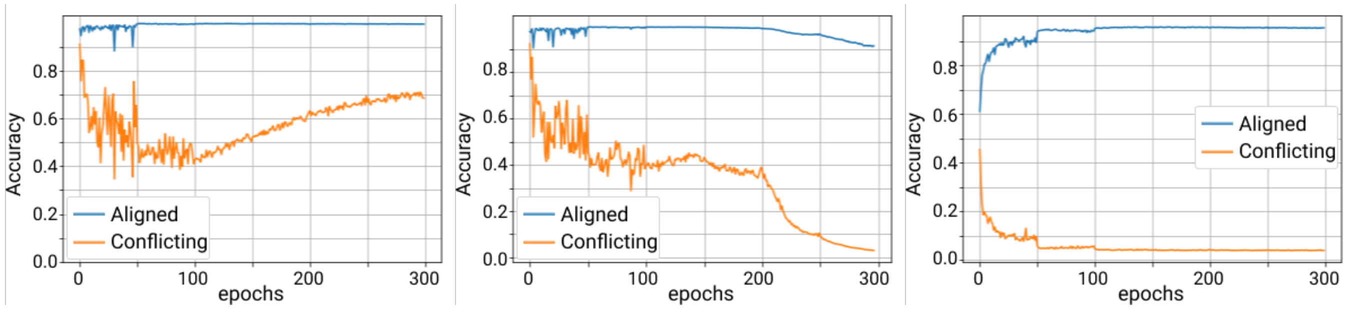
The issue of bias in computer vision has been widely studied where the bias could be in terms of under-represented class samples (Li and Vasconcelos 2019), gender (Wang et al. 2019), demographics (Lahoti et al. 2020) or other cases. The use of computer vision has a variety of practical applications ranging from autonomous driving to medicine. Particularly the use of vision systems in applications such as face recognition (Robinson et al. 2020) and image generation (Ramaswamy, Kim, and Russakovsky 2020) are widespread. The bias in such systems may unduly affect these systems. Practical instances have been observed as that of image super-resolution (Menon et al. 2020) resulting in generating white race images for down-sampled images from other races.

In this paper, we are particularly concerned by such bias in computer vision and would like to address this.

One class of approaches used to solve this problem is to use an adversarial learning technique (Louppe, Kagan, and Cranmer 2017; Madras et al. 2018) where we use a discriminator to distinguish the particular attributes such as race, gender or age that we are concerned about. This discriminator is trained in an adversarial learning framework in a manner reminiscent of the domain adaptation approach (Ganin and Lempitsky 2015). That is, the loss from the discriminator is reversed while training the 'feature extractor'. This is achieved by taking a set of deep learning layers to form a feature extractor and branching from this to two networks a classifier and a discriminator. The classifier thus receives features through which at the end of training, a discriminator is not able to distinguish a bias attribute in question (race/gender). The idea is appealing and can be effective in mitigating bias. However, a drawback that is introduced through this framework is that it directly introduces a trade-off with accuracy. This is particularly the case when there is high skew in terms of the particular attributes, i.e. gender or race, then the accuracy suffers significantly. In our work, we use a simple method to ensure that we obtain both objectives, that is, the bias is minimized and accuracy is high.

The crux of our idea is to ensure explicit de-correlation between the adversarial discriminator and classifier. This is because, when there is substantial skew in the data, for instance very few images of a specific attribute (gender or race for instance) then the discriminator has very few samples to distinguish the set of 'bias' attributes (the set of attributes we are concerned should not be discriminated against). As a result, the discriminator learns more through the set of features that are related to the classifier's set of labels i.e. in the output space of the classifier. For instance, if we consider we are distinguishing digits between 0 to 9 and almost all the samples are grey samples and very few samples are colored samples, (for instance samples of 4 and 6 only) then the discriminator tries to identify the class label in an effort to distinguish whether the sample is colored or not. Thus, in the example knowing whether the sample is 4 or 6 would help the discriminator in predicting whether the digits are colored or not. This is particularly the case as we are using the discriminators loss adversarially and ensuring that the discriminator has a hard time ensuring whether the digits are colored or not. In this paper, we propose to use a biased discriminator to improve the accuracy of adversarial methods

[*] Equal contributions.
[†] Work done at IIT Kanpur



(a) Discriminator on the features of a vanilla model

(b) Discriminator on the features of an adversarially trained model

(c) Discriminator on the features of an adversarially trained model with GBA (ours)

Figure 1: Accuracy plots of discriminator on aligned and conflicting samples on different methods

while debiasing the feature representation. We use gradients of the discriminator and propose a masking scheme for the features, we term as Gradient Based Activation (GBA). Use of GBA provides us a masking rule to drop certain features in order to do unbiased training. The effectiveness of this approach is demonstrated and validated. We further describe the proposed approach in detail in coming sections and provide its extensive analysis to demonstrate the efficacy of the proposed method.

Related Work

Bias in Computer Vision: The issue of bias in computer vision has been considered by several works (Grover et al. 2019; Kim et al. 2019; Quadrianto, Sharmanska, and Thomas 2019; Ganin and Lempitsky 2015). It causes unfair or biased leaning in computer vision tasks such as face recognition (Robinson et al. 2020; Xu et al. 2020), object detection (de Vries et al. 2019) and image generation (Xu et al. 2018). Data imbalance is one of the source of bias learning (Buolamwini and Gebru 2018). Ramaswamy *et al* (Ramaswamy, Kim, and Russakovsky 2020) tackle it by generating realistic samples from the GANs. It has been shown that balanced data also faces bias in feature representations (Wang et al. 2019).

Adversarial debiasing approaches: Some of the works that pursued adversarial learning include the works by (Louppe, Kagan, and Cranmer 2017; Kurmi, Kumar, and Namboodiri 2019) and (Madras et al. 2018). Similarly, to prevent gender bias, Wadsworth *et al.* (Wadsworth, Vera, and Piech 2018) present an adversarially-trained neural network that predicts recidivism and is trained to remove racial bias using a discriminator. (Adel et al. 2019; Zhang, Lemoine, and Mitchell 2018) includes a new hidden layer to enable the concurrent adversarial optimization for fairness and accuracy. Adversarial bias removal methods are also applied in text data. (Elazar and Goldberg 2018). In recent work (Lahoti et al. 2020) train an adversarial reweighting approach for improving fairness.

Other debiasing efforts: A number of other approaches have also been considered for solving this problem. In (Wang et al. 2019), the authors show that even when datasets are balanced (each label co-occurs equally with each gender), learned models do amplify the association between labels and gender. (Kiritchenko and Mohammad 2018; Vig et al. 2020) analyze the gender bias in the case of sentiment analysis. Another gender bias work (Buolamwini and Gebru 2018) evaluates bias present in automated facial analysis algorithms and datasets for phenotypic subgroups. (Leino et al. 2019) demonstrates that bias amplification can arise via an inductive bias in gradient descent methods. A kernel density estimation (Cho, Suh, and Hwang 2020) is applied to tackle the fairness problem. Gat *et al.* (Gat et al. 2020) present a regularization term based on the functional entropy to remove a classifier's bias. In another work, (Nam et al. 2020) show that sample performance-based methods can be used to avoid the bias in the model. A disentanglement approach to obtain the bias invariant representation has been presented in (Sarhan et al. 2020).

In contrast to these other techniques, our work is focused on solving the drawback that we identify in adversarial learning framework for debiasing. Further, our work provides insight into the role of feature representations and feature masking while training a classifier and informs us about the source of bias in a classifier.

Bias in Machine Learning

In any dataset we may have classes that are highly skewed towards a particular sensitive attribute. This skew leads to a classifier correlating class information with these sensitive attributes and hence inducing stereotypes in learning. When the sensitive attributes are easy to learn, there will be no incentive for a model to learn class features. For instance, in a dataset, if images of horse are dominantly in color (RGB channels exist) and that of deer are dominantly grey (only grey channel exists), a grey horse, may be misclassified as a deer based on the number of channels it has, rather than the appearance. This would be because, the classifier would associate 'greyscale' attribute with the label 'deer'. For this example, all examples of 'colored horses' and 'greyscale deer' are aligned with the bias that exists in the dataset and we term these as 'bias aligned' samples. The examples of 'greyscale horses' and 'colored deer' are termed as 'bias conflicting' samples as they are not following the dominant bias in the dataset. In the setting of bias/fairness we want a classifier to be agnostic to these bias attributes. A discriminator model trained on the features of the classifier and its ability to discriminate among the bias attributes aims to give us a measure of bias in the feature representation. For instance, the discriminator for the example would aim to clas-

sify whether a feature representation is of color images or that of greyscale images. This discriminator used adversarially would aim to make the feature representation invariant to the greyscale/color attribute. In general feature space is entangled with bias and class information. The purpose of the debiasing approaches is to remove the entangled bias information from the feature space.
Adversarial approaches build on the above hypothesis and use an adversarial loss in order to debias the feature space. But it has been seen and stated in several works (Wadsworth, Vera, and Piech 2018; Madras et al. 2018) that adversarial approaches for debiasing introduce a trade-off with accuracy.

In this work we identify the reason for this trade-off with empirical justifications and provide a simple and effective method to address this issue.

Class Correlation of Discriminator

In Fig 1a we provide a discriminator's validation accuracy plot for a vanilla model on the CIFAR-10S dataset. We see a discrepancy in the performance between bias-aligned and conflicting samples. In the case of CIFAR-10S, the samples are biased based on the color attribute. This clearly indicates that a discriminator benefits from class correlations in its prediction i.e. a sample of a given bias in its dominant class is identified with higher accuracy by the discriminator due to the presence of class correlation with the bias. This makes the discriminator itself biased. It has high accuracy for the bias aligned samples. For instance for some classes, on the basis of alignment with the dominant color attribute, the discriminator performs well. On the other hand this type of class correlation for predicting the bias attribute harm the discriminator accuracy on bias conflicting samples. For these samples a discriminator perform well if and only if there is certain bias information in the representation. That is, for the example 'greyscale horse', the discriminator performs badly. It perform well for these samples only if the discriminator actually represents the color information accurately. This implies discriminator performance on bias conflicting samples is the indicator of bias rather than on the whole test set.

Adversarial Learning

A class of methods use the adversarial loss of discriminator to debias the feature representation. Adversarial loss is implemented to cause degradation in the discriminator's performance, as we discussed in the previous section discriminator may use class cues for the discrimination and simultaneously there is a classifier which try to learn these class cues. So in a sense we have a conflicting objective here which may harm the class features rather than the bias. We show the discriminator's validation accuracy plot in Fig 1b, being a biased discriminator it correlates with class features and hence the bias aligned accuracy can be maintained high. However, here we observe that in order to debias, the adversarial loss depreciates the bias aligned accuracy along with the bias conflicting accuracy hence degrading the class correlated features in the process of debias causing a bias-accuracy trade-off. This can be observed by considering also the classifier's accuracy for the adversarial training as we show later. We show that through adversarial training, while the discriminator's accuracy reduces, the classifier's accuracy also reduces. This is due to the correlation with bias. This bias-accuracy trade-off is undesirable.

A question may arise that why is bias conflicting accuracy even worse than random? This is because more debias the classifier more biased is the discriminator , i.e. it correlates more with the class cues, hence it predicts the bias attribute by associating with the class of the sample and as the bias conflicting samples do not follow the dominant bias hence it predicts the wrong attribute. It means that correlated features in this case will be bias free. Using these observations we motivate the proposed approach in the next section.

Motivation

In the above sections we have discussed about the behaviour of discriminator on different samples, the possibility of it being biased and how adversarial loss from such a discriminator is responsible for the bias-accuracy trade-off. We started with the problem of bias in classifier and ended up having a biased discriminator. We propose how a biased discriminator can rather be used as an effective tool for debiasing, which can by itself prevent the bias-accuracy trade-off as well as promote an unbiased feature representation.

We analyse this carefully and obtain a method in this paper that shows how a biased discriminator can be used to debias the classifier without compromising on accuracy in an adversarial framework.

The use of biased discriminator is based on the following observation:

- When the prediction of a discriminator is correct, it is attending to the features correlated with the bias attribute. Such features are unwanted in our representation, hence masking them during the classifier training encourage the classifier to learn through the unbiased features. This masking also prevent the adversarial loss from degrading the class features correlated with the bias.
- In the case when the discriminator's prediction is incorrect, it is attending to the features with no bias attribute information and rather is spuriously correlating features with the predicted incorrect bias attribute. For instance, it is predicting 'greyscale' just by observing a deer. In this case, the features are correlated with the class. Hence, we propose to enhance these features to promote unbiased learning in our classifier. In this case, as the discriminator is not able to predict the right bias attribute, it implies that the learning will be neutral with respect to the bias attributes.

In Fig. 1c we show the discriminator test accuracy plot of our approach, we see how using the proposed strategy provides the debias (near zero bias conflicting accuracy). On the bias aligned samples, the discriminator retains its accuracy. This is because, these are aligned with classification and the masking ensures that these features are not degraded during the adversarial training. This approach also ensures that we obtain high classification accuracy for both the bias aligned as well as bias conflicting samples as we show later.

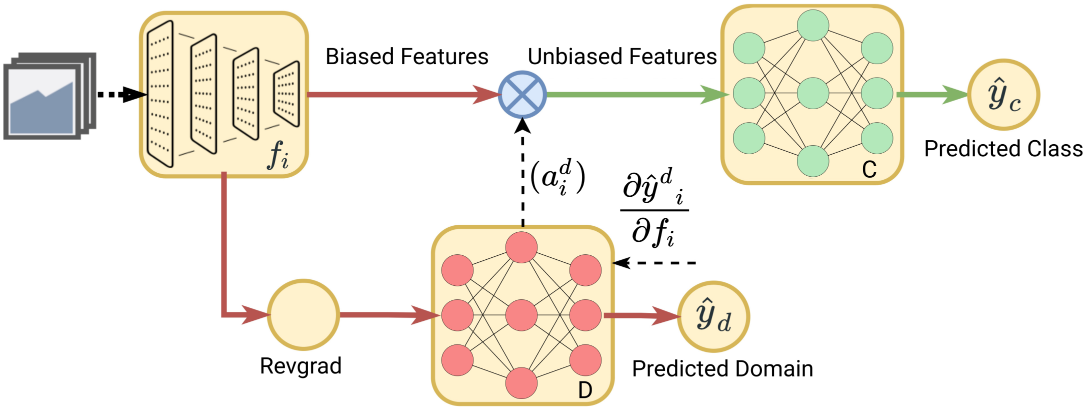


Figure 2: Illustration of training a adversarial model using proposed GBA (Gradient-based Activation) framework for debiasing. The features that are used by the classifier are filtered by discriminator's GBA, which uses discriminator's knowledge to inhibit domain discriminative features. Here the RevGrad is the gradient reversal layer (Ganin and Lempitsky 2015), dashed arrows represent the computation done during the backward propagation.

Proposed Approach

In this section we explain the proposed approach in detail, as illustrated in Fig. 2.

Problem Formulation

We formulate the problem of classification as a supervised learning problem, where the objective is to predict the class label y^c for a given input x. The constraint is that the output variable must be unbiased with respect to some bias attribute variable y^d. Assume that there are total N training samples of distribution $\mathcal{D}$, in form of tuples $\{x_i, y_i^c, y_i^d\}_{i=1}^N$ are available. Where x_i is the input images, y_i^c is the class label and y_i^d refer to the bias attribute, independent of y_i^c. The predictor network $F(x)$ has access to the input variable x_i and the bias attribute y_i^d. We follow the adversarial learning-based framework to debias towards the bias attributes while predicting the class labels. In adversarial learning, a discriminator is trained to predict the bias attribute and a feature extractor is trained adversarially to it using a gradient reversal layer (Ganin and Lempitsky 2015). This is clarified further in the next sub-section.

Adversarial Learning for Bias Mitigation

The adversarial learning framework consists of a feature extractor (F) and a classifier or label predictor network (C). For incorporating adversarial learning, we use a discriminator network (D). Feature extractor, classifier and discriminator are parameterised by parameters θ_f, θ_c and θ_d respectively. The features f_i of the input image (x_i) are encoded using the feature extractor, and these are classified by the classifier. The features are also provided to the discriminator to predict the bias attributes. The corresponding equations are given by:

$$f_i = F(x_i;\theta_f); \quad \hat{y}_i^c = C(f_i, \theta_c); \quad \hat{y}_i^d = D(f_i, \theta_d) \quad (1)$$

$$\mathcal{L}_c = \frac{1}{N}\sum_{x_i \in \mathcal{D}} \mathcal{L}(\hat{y}_i^c, y_i^c) \quad \mathcal{L}_d = \frac{1}{N}\sum_{x_i \in \mathcal{D}} \mathcal{L}(\hat{y}_i^d, y_i^d) \quad (2)$$

where N is the total number of images and $\mathcal{L}$ is the cross-entropy loss. In an adversarial learning setup, the total cost is obtained as:

$$\mathcal{C} = \mathcal{L}_c - \lambda * \mathcal{L}_d \quad (3)$$

$$(\hat{\theta}_f, \hat{\theta}_c) = \arg\min_{\theta_f, \theta_c} \mathcal{C}(\theta_f, \theta_c, \theta_d)$$

$$(\hat{\theta}_d) = \arg\max_{\theta_d} \mathcal{C}(\theta_f, \theta_c, \theta_d)$$

This is the standard setup for adversarial training. In the following subsections we consider our approach in more detail.

Gradient Based Activations to Debias the Features

Our approach towards bias mitigation is based on use of gradient based activations (GBA). The gradient of the output variable with respect to the input feature and its positive activation has been successfully applied to obtain explainability about the prediction (Selvaraju et al. 2017). We follow a similar approach to identify the features used by discriminator in prediction.

The features obtained from Eq.1 are entangled, i.e., contain both class and bias attribute information. As a result, the classifier minimizes the label prediction loss by considering both these features. For unbiased classification, the prediction must be independent of the bias attribute. In order to debias the classifier, we seek the discriminator's knowledge to identify the bias discriminative features and selectively mask the desired features to pass through the classifier.

We obtain the gradients of the prediction of the discriminator $\frac{\partial \hat{y}}{\partial f_i}$ w.r.t the features using the backward propagation to obtain features attended by discriminator. We mask the features with positive gradients when the discriminator is correct, and propagate the features with positive gradient when the discriminator is incorrect.

We define indicator variable ind^d using the following condition in Eq 4. $\hat{y}$ in the Eq 5 is the maximally activated logit, we obtain its gradients with respect to the features in Eq 6.

These gradients are then conditionally used to create the final mask vector a_i^d in Eq 7.

$$ind^d = \begin{cases} 1, & \text{if argmax}(\hat{y}_i^d) = y_i^d \\ \text{-1} & \text{otherwise} \end{cases} \quad (4)$$

$$\hat{y} = max(\hat{y}_i^d) \quad (5)$$

$$g_i^d = \frac{\partial \hat{y}}{\partial f_i} \quad (6)$$

$$a_i^d = \begin{cases} 0, & \text{if } (g_i^d.ind^d) > 0. \\ 1 & \text{if } (g_i^d.ind^d) \leq 0 \end{cases} \quad (7)$$

The effective features for the classifier are obtained as follows: $$f_i^{cls} = f_i * a_i^d \quad (8)$$
'$*$' represents the element-wise multiplications.
f_i^{cls} is now used for the training in the adversarial manner as discussed in previous section.

Experiments and Results

Datasets

We evaluate the proposed model on the following standard datasets : **CIFAR-10S (Wang et al. 2020):** It is a skewed version of CIFAR-10 (Darlow et al. 2018), presented by Wang *et al* (Wang et al. 2020). This data contains tranformational bias. It consists of 50,000 images of size 32×32 of 10 object classes. Each class has a total of 5000 images. CIFAR-10S is divided into two domains color and greyscale domains. In this datasets per class, the 50,000 training images are split 95% to 5% between the two domains; five classes are 95% color, and five classes are 95% greyscale. For testing we evaluate each class on bias aligned and conflicting samples.

CIFAR-I: It is an extension of CIFAR-10S (Wang et al. 2020), where the images of the skewed domain are taken from similar classes of ImageNet (Deng et al. 2009). So in this dataset, there are 10 classes and two domains (bias attributes).

ColoredMNIST: The ColoredMNIST dataset containing colour bias is taken from (Nam et al. 2020). In this dataset, images of greyscale MNIST dataset are injected 10 colors with random perturbation in each class, resulting in a dataset with 10 classes of digits and 10 domains of colors.

CelebA: It is a real world multi-attribute dataset consisting of 40 attributes for each image. Here we have an example of real world bias in form of Gender. We find that Hair Color attribute and Heavy Makeup are the most correlated to the bias attribute (Gender) as done by (Nam et al. 2020), so we perform experiment on two setups, Hair Color as target attribute and Gender as bias attribute and Heavy Makeup as target attribute and Gender as bias attribute.

Training Setup

For the CIFAR-10S, CIFAR-I and CelebA datasets, we use the Resnet-18 (He et al. 2016) model, where the last fully connected layer is replaced with two consecutive fully connected layers. In the Colored MNIST dataset, we use the multi-layered perceptron consisting of three hidden layers as the feature extractor.

Results and Discussion

CIFAR-10S We use a ResNet-18 (He et al. 2016) model trained on CIFAR-10S as the vanilla baseline. The adversarial model is using the gradient reversal layer as in (Ganin and Lempitsky 2015). We evaluate the methods on bias aligned and bias conflicting accuracies along with their mean. For the unbiased model both these accuracies must be close. We measure this using *bias gap* which is the difference between the aligned and conflicting accuracies. In Fig 3 we show the bias aligned and bias conflicting accuracy plots on different methods, bias gap of baseline vanilla model in Fig 3a highlights the problem of bias in standard deep learning models. Fig 3b show how adversarial methods can reduce the bias gap as compared to baseline but degrades the overall class accuracy. Fig 3c show the unbiased learning of the proposed approach, note that it has solved the problem of bias-accuracy trade-off which was there in adversarial methods.

Further we show performance comparison with baselines and recent techniques in Table 1. We observe a increase of **3.5%** and **5.5%** in accuracy as compared to the vanilla baseline and adversarial approach respectively along with greater reduction in bias. Moreover we observe that the proposed approach also outperforms other recent works in the task of bias mitigation.

CIFAR-I The color-greyscale transformation of CIFAR-10S is one difference in terms of data distribution. Another case of bias could be in terms of distribution of data samples. We evaluate our algorithm on such a transformation to simulate real-world data using CIFAR-I with samples from another datasets. The other dataset we use is that of similar classes from ImageNet dataset. This distribution change exhibits a different bias as compared to the color-greyscale transformation. In Table 2 we report and compare the performance of the proposed method with baselines and recent techniques. We can observe that the proposed method achieves a boost of approximately **6%** and **4%** in mean accuracy over the vanilla baseline and adversarial methods respectively and outperforms the state-of-the-art domain independent approach in both mean accuracy and bias removal.

Colored MNIST Another case of biased learning is in Colored MNIST dataset (Nam et al. 2020), where MNIST dataset is injected with colors for each class respectively. In this case, the neural network generally learns to classify them on the basis of color rather than learning about digits. The previous two datasets had only two bias attributes to discriminate; in this dataset we have ten bias attributes, using this dataset we test the scalability of our model to multiple number of attributes. The performance on this dataset has been reported in Table 4 for different level of skews. Here again we see along with outperforming the recent methods the proposed model improves on the adversarial method by a large number. We note the recent domain independent (Wang et al. 2020) network performs poorly in this multiple domains setting.

CelebA This dataset contains gender bias with respect to the heavy makeup and hair color attribute. To evaluate different algoritms on these attribute learning tasks we report

Model Name	**Model**	**Accuracy (↑)**			Bias (GAP)(↓)
		Aligned	Conflicting	Mean	
Baseline	N-way Softmax	94.75 ± 0.25	82.30 ± 0.31	88.43 ± 0.20	12.45 ± 0.40
LfF(Nam et al. 2020)	N-way Softmax	90.33 ± 1.80	68.64 ± 1.71	79.49 ± 1.24	21.69 ± 2.48
Domain Ind(Wang et al. 2020)	N-way Classifier per Domain	92.38 ± 0.20	91.86 ± 0.21	**92.12 ± 0.15**	0.52 ± 0.29
Adversarial	Gradient Reversal	86.98 ± 0.70	86.61 ± 0.37	86.80 ± 0.40	0.37 ± 0.80
Adversarial with GBA	Proposed	91.95 ± 0.22	**92.05 ± 0.23**	92.00 ± 0.15	**0.10 ± 0.31**

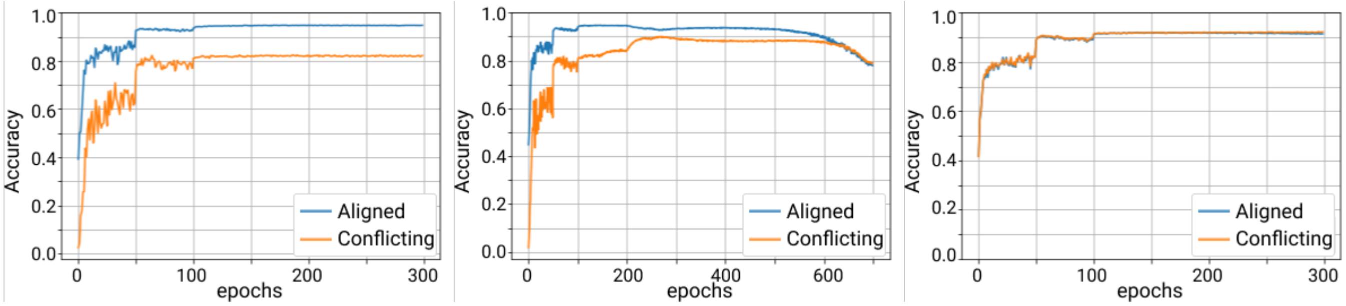
Table 1: Performance comparison of different algorithms on CIFAR-10S, Here we show test accuracy on the bias aligned and bias conflicting samples as a measure of biasness, It can be seen that GBA is the best in terms of debiasing

Model Name	**Model**	**Accuracy (↑)**			Bias (GAP)(↓)
		Aligned	Conflicting	Mean	
Baseline	N-way Softmax	87.94 ± 0.36	69.39 ± 0.42	78.67 ± 0.27	18.55 ± 0.55
LfF(Nam et al. 2020)	N-way Softmax	87.01 ± 0.63	56.87 ± 0.72	71.93 ± 0.47	30.14 ± 0.95
Domain Ind(Wang et al. 2020)	N-way Classifier per Domain	88.39 ± 0.15	78.14 ± 0.13	83.26 ± 0.01	10.25 ± 0.20
Adversarial	Gradient Reversal	85.52 ± 0.65	75.74 ± 0.29	80.63 ± 0.35	9.78 ± 0.71
Adversarial with GBA	Proposed	88.81 ± 0.19	**79.46 ± 0.22**	**84.41 ± 0.14**	**9.35 ± 0.29**

Table 2: Performance comparison on a non-linear transformation, CIFAR-I setting, here also our algorithm outperforms existing algorithms in both bias and accuracy

Model	**Heavy Makeup**				**Hair Color**			
	Aligned	**Conflicting**	**Mean**	**Bias Gap**	**Aligned**	**Conflicting**	**Mean**	**Bias Gap**
Vanilla	92.44±0.74	31.46±2.45	61.95±1.28	60.98±2.56	90.58±0.34	57.35±0.21	73.97±0.20	33.23±0.4
LfF (Nam et al. 2020)	83.85±1.68	45.54±4.28	64.69±2.29	38.31±4.60	88.85±1.27	80.24±2.16	84.55±1.25	8.61±2.5
DomainInd (Wang et al. 2020)	79.88±1.71	43.24±4.33	61.56±2.31	36.64±5.64	90.97±3.71	79.25±3.33	85.11±2.67	7.44±3.2
GrpDRO (Sagawa et al. 2020)	79.28±1.20	46.24±3.61	62.76±2.22	33.04±3.22	89.68±0.65	81.41±1.47	85.55±0.88	8.27±2.0
Adversarial	92.07±2.88	33.79±3.81	62.93±2.38	58.28±4.77	93.4±0.91	62.75±3.47	78.08±1.79	30.65±5.6
Adversarial with GBA	81.49±1.91	**49.79±3.15**	**65.64±1.55**	**31.70±3.10**	90.67±1.01	**83.28±1.83**	**86.98±1.04**	**7.39±2.1**

Table 3: Results on the Heavy Makeup and Hair Color Attributes of the CelebA Dataset, with the bias attribute being the gender, here we show that the Simple Adversarial method was unable to debias the model, the representation. Using GBA, we reduce the bias gap greatly when compared to the Adversarial method, and while maintaining state of the art accuracy.

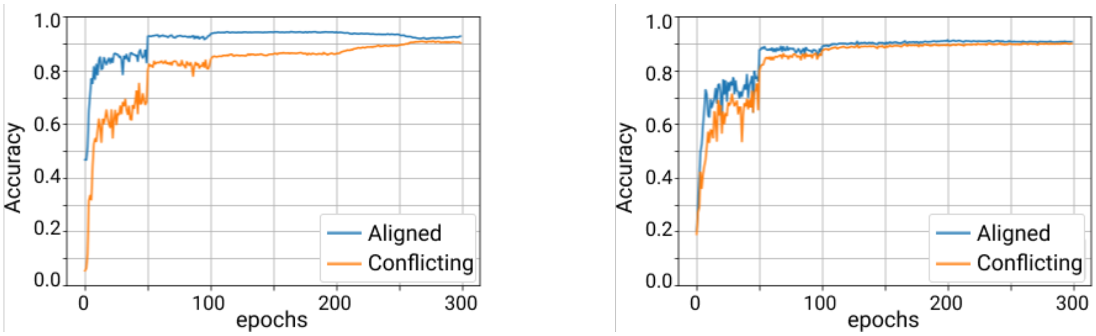


(a) Baseline (b) Adversarial (c) GBA

Figure 3: Figures above show the validation accuracy of bias aligned and bias conflicting samples over the course of training on CIFAR-10S dataset. We observe that the baseline model has poor performance on the bias conflicting samples compared to the bias aligned samples. The adversarial model improves upon the baseline, but the trade-off is evident as to completely debias the model, the class features are harmed. GBA with the adversarial framework makes it completely fair in terms of the bias, there is almost no discrepancy in the aligned and conflicting accuracy, and the average accuracy also improves significantly.

the accuracy on bias aligned and bias conflicting samples, along with mean accuracy and bias gap for particular target attribute on the unbiased test set in Table 3. We observe adversarial method improving by **3%** and **9%** in average accuracy and **7%** and **1.2%** in bias gap on heavy makeup and hair color attributes respectively. We also observe performance of the proposed method outperforms the recent state-of-the-art methods like LfF (Nam et al. 2020).

In this section we discussed the performance of different methods on various datasets and metrics. We observe GBA greatly improves the accuracy of adversarial method, moreover the proposed approach is the best performing method averaged across all the datasets. In the results discussed above adversarial with GBA is on average **5%** better than



(a) GBA-n, Naive masking using GBA

(b) GBA-c, Only correct masking using GBA

Figure 4: Ablation study of masking rule by GBA on CIFAR-10S dataset.

Model	**Accuracy**	
	95%Skew	**98%Skew**
Vanilla	77.63±0.44	62.29±1.47
Domain Ind(Wang et al. 2020)	65.82±0.81	45.39±1.20
Filter-Drop(Nagpal et al. 2020)	78.44±0.58	62.31±1.72
Group-DRO(Sagawa et al. 2020)	84.50±0.46	76.30±1.53
REPAIR(Li and Vasconcelos 2019)	82.51±0.59	72.86±1.47
LfF(Nam et al. 2020)	85.39±0.94	**80.48±0.45**
Adversarial	80.35±0.52	64.83±0.34
Adversarial with GBA	**87.92±0.6**	79.11±1.6

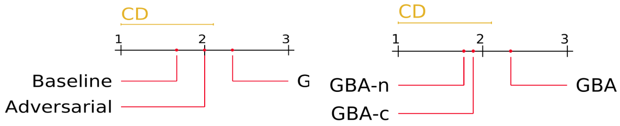
Table 4: Performance comparisons in term of classification accuracy on Colored MNIST dataset.

Model	**Accuracy**			Bias
	Aligned	Conflicting	Mean	(GAP)
GBA-n	92.05	90.60	91.32	1.45
GBA-c	90.69	90.06	90.37	0.63
GBA	92.05	**91.95**	**92.00**	**0.10**

Table 5: Results on the ablation of various activations on CIFAR-10S datasets.

the second best method-LfF in terms of accuracy. The other details and experiments are provided in the project page [1].

Analysis



(a)

(b)

Figure 5: Statistical significant test for CIFAR-10S datasets on (a) baseline, adversarial, and GBA method (b) different variations of proposed model GBA-c, GBA-n and GBA.

Ablation Study of Gradient Based Activations

In the proposed approach of gradient based activations (GBA), masking rule for the features is conditioned on discriminator's prediction, in this section we analyse how this conditioning is crucial for the performance of GBA. Fig 4a represents bias aligned and conflicting accuracy if we naively mask features attended by discriminator (GBA-n) i.e. drop the features that discriminator is using for prediction without considering whether the discriminator is correct or not. The plot when compared to Fig 3c shows highly biased trend highlighting the importance of right conditioning. To improve upon GBA-n, in Fig 4b we see the variation GBA-c where we only attend to the classifier when the discriminator correctly classifies the input and pass the raw unattended features when the discriminator is incorrect. Here we can see better bias removal than GBA-n version which support our hypothesis that discriminator correlates with the class features to predict the incorrect domain. Hence, enhancing these features while training promotes unbiased learning and improves class prediction ability as seen in Fig 3c. In Table 5, we report the performance of different ablations where we see a systematic improvement in bias and accuracy as we apply different components of GBA.

Statistical Significance Analysis

We analyze the statistical significance (Demšar 2006) for the proposed method in bias mitigation for CIFAR-10S dataset. The Critical Difference (CD) is related to the confidence level (0.05 in our case) for the number of tested datasets and average ranks. If the methods' rank difference is outside the CD (1.048 for our case), it implies that these two methods are significantly different. In Fig. 5a and Fig 5b, we provide the statistical test for baselines, adversarial with the proposed method and different variations of the proposed method defined in the previous section. It visualizes the post hoc analysis using the CD diagram for CIFAR-10S dataset. From the figures, it is clear that the proposed method is significantly different from the baseline model and adversarial method.

Conclusion

Through this work, we provide a method to address the crucial problem in the adversarial learning framework to obtain unbiased classification. Through extensive empirical analysis on multiple standard datasets we show that the proposed approach works well. Specifically, we showed that gradient based activation uses a biased discriminator's gradients in order to debias the classifier. Our ablation analysis also justifies the use of the proposed method. Through our work, we also obtain a better understanding of debiasing a classifier, particularly in an adversarial setting.

[1] https://vinodkkurmi.github.io/GBA/

References

Adel, T.; Valera, I.; Ghahramani, Z.; and Weller, A. 2019. One-network adversarial fairness. In *AAAI*, volume 33, 2412–2420.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Cho, J.; Suh, C.; and Hwang, G. 2020. A fair classifier using kernel density estimation. In *NeurIPS 2020*.

Darlow, L. N.; Crowley, E. J.; Antoniou, A.; and Storkey, A. J. 2018. CINIC-10 is not ImageNet or CIFAR-10. *arXiv preprint arXiv:1810.03505*.

de Vries, T.; Misra, I.; Wang, C.; and van der Maaten, L. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 52–59.

Demšar, J. 2006. Statistical comparisons of classifiers over multiple data sets. *JMLR*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 11–21.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *ICML*, 1180–1189. PMLR.

Gat, I.; Schwartz, I.; Schwing, A.; and Hazan, T. 2020. Removing Bias in Multi-modal Classifiers: Regularization by Maximizing Functional Entropies. *NeurIPS*.

Grover, A.; Song, J.; Agarwal, A.; Tran, K.; Kapoor, A.; Horvitz, E.; and Ermon, S. 2019. Bias correction of learned generative models using likelihood-free importance weighting. *arXiv preprint arXiv:1906.09531*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Kim, B.; Kim, H.; Kim, K.; Kim, S.; and Kim, J. 2019. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9012–9020.

Kiritchenko, S.; and Mohammad, S. 2018. Examining Gender and Race Bias in Two Hundred Sentiment Analysis Systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, 43–53.

Kurmi, V. K.; Kumar, S.; and Namboodiri, V. P. 2019. Attending to discriminative certainty for domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 491–500.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. H. 2020. Fairness without demographics through adversarially reweighted learning. *NeurIPS 2020*.

Leino, K.; Black, E.; Fredrikson, M.; Sen, S.; and Datta, A. 2019. Feature-wise bias amplification. *ICLR*.

Li, Y.; and Vasconcelos, N. 2019. Repair: Removing representation bias by dataset resampling. In *CVPR*, 9572–9581.

Louppe, G.; Kagan, M.; and Cranmer, K. 2017. Learning to pivot with adversarial networks. In *NeurIPS*, 981–990.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning Adversarially Fair and Transferable Representations. In *ICML*, 3384–3393.

Menon, S.; Damian, A.; Hu, S.; Ravi, N.; and Rudin, C. 2020. PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. In *CVPR*, 2437–2445.

Nagpal, S.; Singh, M.; Singh, R.; and Vatsa, M. 2020. Attribute aware filter-drop for bias invariant classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 32–33.

Nam, J.; Cha, H.; Ahn, S.; Lee, J.; and Shin, J. 2020. Learning from Failure: Training Debiased Classifier from Biased Classifier. In *NeurIPS*.

Quadrianto, N.; Sharmanska, V.; and Thomas, O. 2019. Discovering fair representations in the data domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8227–8236.

Ramaswamy, V. V.; Kim, S. S.; and Russakovsky, O. 2020. Fair Attribute Classification through Latent Space De-biasing. *arXiv preprint arXiv:2012.01469*.

Robinson, J. P.; Livitz, G.; Henon, Y.; Qin, C.; Fu, Y.; and Timoner, S. 2020. Face recognition: too bias, or not too bias? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 0–1.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2020. Distributionally Robust Neural Networks. In *ICLR*.

Sarhan, M. H.; Navab, N.; Eslami, A.; and Albarqouni, S. 2020. Fairness by Learning Orthogonal Disentangled Representations. *ECCV*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 618–626.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating gender bias in language models using causal mediation analysis. *NeurIPS*, 33.

Wadsworth, C.; Vera, F.; and Piech, C. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *FAT/ML*.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5310–5319.

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*, 8919–8928.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, 570–575. IEEE.

Xu, T.; White, J.; Kalkan, S.; and Gunes, H. 2020. Investigating bias and fairness in facial expression recognition. In *European Conference on Computer Vision*, 506–523. Springer.

Zhang, B. H.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*.