Counterfactual Interpolation Augmentation (CIA): A Unified Approach to Enhance Fairness and Explainability of DNN

Yao Qiang , **Chengyin Li** , **Marco Brocanelli** and **Dongxiao Zhu***
Department of Computer Science, Wayne State University, USA
{yao, cyli, brok, dzhu}@wayne.edu

Abstract

Bias in the training data can jeopardize fairness and explainability of deep neural network prediction on test data. We propose a novel bias-tailored data augmentation approach, Counterfactual Interpolation Augmentation (CIA), attempting to debias the training data by d-separating the spurious correlation between the target variable and the sensitive attribute. CIA generates counterfactual interpolations along a path simulating the distribution transitions between the input and its counterfactual example. CIA as a pre-processing approach enjoys two advantages: First, it couples with either plain training or debiasing training to markedly increase fairness over the sensitive attribute. Second, it enhances the explainability of deep neural networks by generating attribution maps via integrating counterfactual gradients. We demonstrate the superior performance of the CIA-trained deep neural network models using qualitative and quantitative experimental results. Our code is available at: https://github.com/qiangyao1988/CIA

1 Introduction

Deep neural network (DNN) trained with biased data is known to learn and exploit the spurious correlation between the target variable and the sensitive attribute (e.g., color, gender, and race) as a shortcut for prediction [Kim *et al.*, 2019; Geirhos *et al.*, 2020]. However, the spurious correlation may only reflect dataset-specific biases or sampling artifacts rather than the causal mechanism between the intended feature and target variable. As a result, the DNN's output may be biased against the protected groups defined by the sensitive attribute. For example, a facial recognition model performs poorly for female with darker skin compared to other gender/race groups [Buolamwini and Gebru, 2018]. Developing bias mitigation techniques to alleviate the adverse effect has attracted increasing attention in recent years.

Extensive approaches have been developed to mitigate bias in DNN's prediction. Many methods attempt to remove sensitive information from the learned features during the training process [Madras *et al.*, 2018; Kim *et al.*, 2019; Li *et al.*, 2020]. However, the adversarial training and disentangled representation learning approaches are limited because they potentially remove some useful information related to the sensitive attribute, thus compromising the model performance on the target task. [Kim *et al.*, 2021] aim to debias and increase the quality of the training set via data augmentation. Despite its initial success, they augment data through linearly interpolating the latent features from the discriminative models, limiting their capability to generate a set of legitimate and manifold data augmentations. Clearly, generative models that learn the distribution of features provide a promising solution.

While many existing approaches ensure fairness, explainability arises as another salient challenge. Besides selecting appropriate metrics (e.g., demographic parity, equality-of-odds) for fairness evaluation, researchers attempt to apply model explanation techniques to help understand whether a DNN model makes fair decisions [Qiang *et al.*, 2020; Pan *et al.*, 2020; Tong and Kagal, 2020]. Among others, feature attribution methods (e.g., IG [Sundararajan *et al.*, 2017]) calculating the attribution of each input feature as its importance have gained great success. Nevertheless, the computing process may be misled by the sensitive attribute, resulting in incorrect explanations as shown in Figure 1(d), due to the arbitrary choices of the baseline and integral path.

To address the above problems, we design a bias-tailored counterfactual interpolation augmentation (CIA) approach to 1) mitigate bias in the training set, and 2) develop fair and explainable DNN models using the counterfactual interpolations generated from CIA. Our unified approach is illustrated in Figure 1. Here we mitigate bias in the training set through the lens of counterfactual fairness [Kusner *et al.*, 2017; Pfohl *et al.*, 2019]. The counterfactual causal inference is modeled using a conditional variational auto-encoder (CVAE) [Sohn *et al.*, 2015], which generates the counterfactual interpolations by interpolating the sensitive attribute along a constructed path simulating the distribution transitions between the sensitive groups. We then inject the bias-tailored counterfactual interpolations into the biased training set to intervene the spurious causal effect. Therefore, DNN models trained with CIA tend to learn the features that are truly causal to the target variables, resulting in fair outputs.

Similar to the attribution methods, the counterfactual ex-

*Corresponding Author

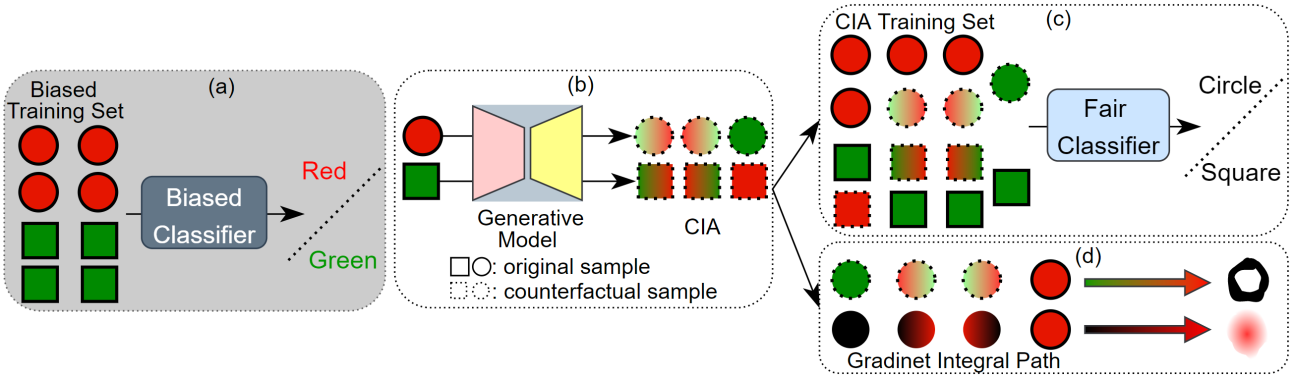


Figure 1: An illustrative example. (a) The target variable (shape) is spuriously correlated with the sensitive attribute (color) in the biased training set. A biased classifier undesirably learns and leverages the spurious correlations for prediction. (b) Our CIA generates bias-tailored counterfactual interpolation augmentation to mitigate bias in the training set and to enhance fair explanation. (c) CIA enables training a fair classifier to learn discriminative features for shape classification. (d) In the first row, CIA generates a meaningful explanation for classifying the target (shape). In the second row, a baseline interpolation generates explanation of the target (shape) confounded by the sensitive attribute (color). Best viewed in color.

planation can give powerful insights into what is important to the underlying decision process leveraging the counterfactual examples, which are in contrast with the original input by making some artificial modifications on the features of interest [Kusner *et al.*, 2017; Wachter *et al.*, 2017]. Here we develop a new DNN model explanation method that integrates gradients along the interpolated path simulating the distribution transitions from the counterfactual example to the input. Since the gradient integration focuses on the intended attributes and does not get distracted by the sensitive attribute, our method can generate more meaningful explanations by dissolving the negative impacts from the sensitive attribute.

We summarize our contributions as follows:

- First, we propose CIA, a novel data augmentation strategy to increase DNN fairness via de-correlating the target from the sensitive attribute in training data.
- Second, we design an DNN model explanation method that leverages the generated counterfactual interpolations from CIA for gradients integration. We demonstrate this work as a unified approach to enhance both fairness and explainability of the DNN.
- Third, we experimentally show that CIA minimizes the detrimental effects of bias using two benchmark datasets. The experiment results demonstrate several quantitative and qualitative benefits of our DNN model explanation approach.

2 Related Works

2.1 Debiasing Learning

Fairness has attracted increasing attention since DNN often exhibits bias towards/against certain protected groups, e.g., as defined by sensitive attributes, such as gender and race [Madras *et al.*, 2018]. The existing fairness-aware prediction methods can be categorized into pre-processing, in-processing, and post-processing approaches.

Pre-processing. Pre-processing approaches attempt to debias and increase the quality of a training set through data augmentation. [Zhang and Sang, 2020] propose to balance data distribution for visual debiasing by adding supplementary adversarial examples. This method relies on selecting adversarial attacks and an auxiliary task classifier to generate adversarial examples. [Kim *et al.*, 2021] first divide the training images into bias-guiding and bias-contrary samples based on the assumption that the bias attributes are easy-to-learn. Then, they generate the bias-swapped image augmentations containing the bias attributes from the bias-contrary images while preserving bias-irrelevant ones in the bias-guiding images. [Chuang and Mroueh, 2021] present fair mixup as a new data augmentation method to generate interpolated samples between the sensitive groups. Fairness is achieved by regularizing the trained DNN model on the path of the generated interpolations with fairness constraints. However, their approach only performs data augmentation by leveraging the latent features from the discriminative models limiting their ability to generate a set of legitimate and manifold instances. Differently, generative models are designed to characterize the probability density of observations in the latent space, leading to a better description of the training dataset. Consequently, they can generate the manifold data augmentations via sampling from the learned latent feature distributions.

In-processing. In-processing approaches aim to remove the sensitive information from the learned features during the training process. Some approaches enforce constraints for specific fairness metrics (e.g., demographic parity and equality-of-odds) via an auxiliary regularization term, either adding constraints to disentangle the association between model predictions and sensitive attributes [Nam *et al.*, 2020] or updating objective function to minimize the performance difference between certain groups [Sagawa *et al.*, 2019]. The problem is that the models may behave differently at inference time even though such fairness constraints are satisfied during training. Adversarial training is enabled through the min-max objective: maximizing the classifier's ability to pre-

dict the target variable while minimizing the adversary's capability to predict the sensitive attribute [Madras *et al.*, 2018; Kim *et al.*, 2019]. Nevertheless, this process can compromise the model performance on the main classification task. [Hong and Yang, 2021] design a novel fairness constraint loss, Bias-Contrastive, utilizing the constractive learning to encourage the proximity between the training examples with the same target class but different bias class in the feature space. Performance of this constractive learning method heavily relies on the choice of positive and negative samples.

Post-processing. Post-processing approaches calibrate or modify the predictions according to the sensitive attribute at inference time [Hardt *et al.*, 2016]. These methods require access to the sensitive attribute, they are not feasible for real-world applications due to the salient security and privacy concerns.

Here we advocate for the pre-processing approach, which is agnostic to the choice of fairness algorithms and DNN model architectures. Moreover, it generates debiased data that can be used for training DNNs to further increase their fairness. Our training data debiasing strategy falls under the umbrella of causal fairness [Kusner *et al.*, 2017] aiming to enforce the model to concentrate more on task-relevant causal features while getting rid of the superficial correlations. Moreover, our strategy is expected to enhance the quality of the existing model explanation as described below.

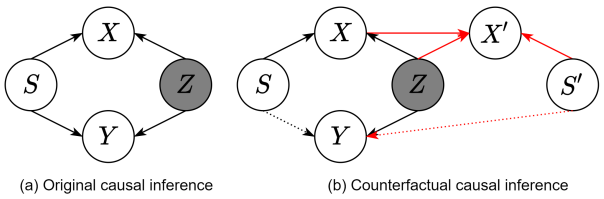
2.2 Integrated Gradients and Variants

Gradient-based feature attribution techniques interpret DNN in terms of the gradient, i.e., the partial derivative of the output with respect to the input, as a sensitivity measurement of the network for each input element [Sundararajan *et al.*, 2017; Erion *et al.*, 2021]. Integrated Gradients (IG) [Sundararajan *et al.*, 2017] applies integrated gradients along a linear path from a baseline to the input avoiding the well-known problem of gradient saturation, i.e., the gradients may not reflect feature importance [Miglani *et al.*, 2020]. The formulation of IG is:

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_0^1 \frac{\partial F(x_i' + \alpha \cdot (x_i - x_i'))}{\partial x_i} d\alpha, \quad (1)$$

where x is the input, x' is the baseline representing a missing or neutral input (e.g., a black or random noise image). $F(\cdot)$ denotes the prediction output. The linear integral path is denoted as: $\gamma(\alpha) = x' + \alpha \times (x - x')$, where $\alpha \in [0, 1]$. To compute this integral efficiently, authors propose a Riemann summation approximation.

Recent studies demonstrate that the choice of baseline heavily impacts the quality of feature attributions [Haug *et al.*, 2021]. [Sturmfels *et al.*, 2020] present several alternatives: (1) Maximum distance baseline; (2) Blurred baseline [Xu *et al.*, 2020]; (3) Gaussian baseline; and (4) Uniform baseline. [Izzo *et al.*, 2020] propose the neutral baseline lying on the decision boundary of the predictive model. More recently, [Pan *et al.*, 2021] develop Adversarial Gradient Integration, which releases the choice of the baseline by integrating the gradients from adversarial examples to the target input. [Kapishnikov *et al.*, 2021] introduce Guided IG, which aligns the model's prediction and the input to explicitly reduce the noise in resulting attributions with a condition path.



(a) Original causal inference (b) Counterfactual causal inference

Figure 2: Structure of the hypothesized causal graphs. (a) Unobserved latent variables Z and sensitive attribute S are two confounders that jointly generate the observed data X and the outcome Y. (b) Add another confounder S' to generate the counterfactual example X'.

We point out a key issue that the existing choices of the baseline critically impact attribution quality, leading to unfair explanations in the debiasing learning scenario. To overcome this, we develop a new attribution based technique, which integrates gradients along the counterfactual interpolated path to achieve a higher explanation quality.

3 Counterfactual Interpolation Augmentation

3.1 Notations

Let $\mathcal{X} = \{x_i, y_i, s_i\}, i \in 1, ..., N$ be the training set, where x_i is the input, y_i denotes the target label, and s_i represents the sensitive attribute. For ease of notation, we consider binary sensitive attributes in the following sections. z is the latent space feature. x' and s' denote the counterfactual samples of x and s, respectively. We use capital letters to denote the random variables.

3.2 Counterfactual Causal Inference

Counterfactual fairness [Kusner *et al.*, 2017] requires the same distribution of predictions for each sample in the factual world where $S = s$ and in counterfactual world where $S = s'$, for all $s' \neq s \in \mathcal{S}$. It refrains the sensitive attribute from being the cause of a change in the model prediction.

Definition 1. (Counterfactual Fairness) [Kusner *et al.*, 2017] *A classifier $\hat{Y}$ is counterfactually fair if under any context $X = x$ and $S = s$,*

$$\begin{aligned} &p(\hat{Y}_{S \leftarrow s} = y | X = x, S = s) \\ =&p(\hat{Y}_{S \leftarrow s'} = y | X = x, S = s'), \end{aligned} \quad (2)$$

for all y and for any value s' attainable by S.

However, the counterfactual fairness only requires the predictions to be the same across factual-counterfactual pairs, regardless of whether those pairs share the same value of the target y. Following [Pfohl *et al.*, 2019], we further require the model to be counterfactually fair, conditioning on the factual target y, formally:

$$\begin{aligned} &p(\hat{Y}_{S \leftarrow s} = y | X = x, Y = y, S = s) \\ =&p(\hat{Y}_{S \leftarrow s'} = y | X = x, Y = y, S = s'). \end{aligned} \quad (3)$$

We seek to address the training data bias problem through the lens of causal inference motivated by Definition 1 and Eq.

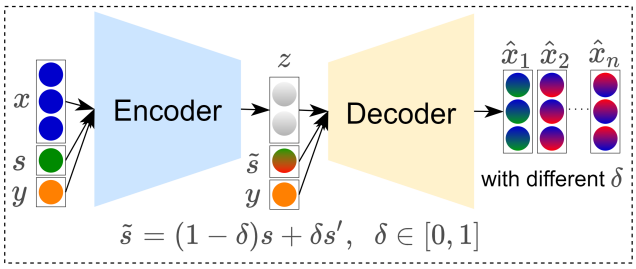


Figure 3: CIA employs a pre-trained CVAE to generate a set of counterfactual interpolations $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$ of x conditioned on interpolated sensitive attributes $\tilde{s}$ and y, where s' contrasts with s.

3. However, it is hard to identify the causal mechanisms from limited observational data that may be sampled from a single biased training distribution. It would be a natural decision to help identify the counterfactual causal mechanisms with additional hand-designed counterfactual examples.

Figure 2(a) illustrates the causal graph, modeling the generative process of the original biased dataset $\mathcal{X}$, in which z is drawn from an isotropic Gaussian prior: $z \sim p(Z) = \mathcal{N}(0, I)$, s is drawn form a multinomial distribution with marginals π: $s \sim p(S) = \text{Categorical}(S|\pi)$, and x and y are drawn independently given s and z: $x, y = p(X|Z, S)p(Y|Z, S)$. The data bias problem is caused by the distribution of sensitive attribute $p(S)$, e.g., s is randomly drawn from a multinomial distribution. We model the counterfactual causal inference to generate counterfactual interpolation augmentations illustrated in Figure 2(b). A counterfactual generative process is $x', y = p(X'|Z, S')p(Y|Z, S')$, and here S' is a new confounding variable in contrast with S.

3.3 Generating Counterfactual Interpolations

It is generally impossible to infer the causal structure of the underlying data generating process directly from the observable properties. Therefore, we employ a generative model to capture the causal structure in the presence of an unobserved confounder with observable proxies [Madras *et al.*, 2019].

We first pre-train a generative model (e.g., CVAE) in which the encoder and decoder inputs are conditioned on the sensitive attribute and target variable. Concretely, the encoder learns $q_\phi(z|x, y, s)$, which is equivalent to learning latent feature z of data x with condition s and y. The decoder learns $p_\theta(x|z, y, s)$ decoding the latent feature z with condition s and y to input space. The generative model is trained to minimize the following objective function:

$$\begin{aligned}\mathcal{L}_{\text{CVAE}}(\theta, \phi) = &- \mathbb{E}_{q_\phi(z|x,y,s)} \log p_\theta(x|z, y, s) \\ &+ \text{KL}(q_\phi(z|x, y, s)||p_\theta(z)).\end{aligned} \tag{4}$$

The first term denotes a reconstruction loss encouraging the encoder to map the observed data (x, y, s) into latent feature z and the decoder to reconstruct x from (z, y, s). The second term indicates a regularization making the distribution $q_\phi(z|x, y, s)$ similar to a prior Gaussian distribution $p(z)$ by Kullback–Leibler (KL) divergence

CVAE can generate non-existent manipulated samples as interpolations for real samples along any arbitrary axis. We design an interpolated path moving linearly along the sensitive attribute s as:

$$\tilde{s} = (1 - \delta) \cdot s + \delta \cdot s', \delta \in [0, 1], \tag{5}$$

and inject $\tilde{s}$ into the decoder of the pre-trained CVAE as shown in Figure 3. We generate a set of counterfactual interpolations $(\hat{x}_1, \hat{x}_2, \cdots, \hat{x}_n)$ transiting from the factual example x to its counterfactual example x' along the interpolated path defined in Eq. 5. The variation of δ determines the number of generated interpolations. This interpolated process is applicable regardless of a single sensitive attribute (e.g., color in BiasedMNIST dataset) or multiple sensitive attributes (e.g., gender and age in CelebA dataset).

4 Training and Interpreting Fair DNN

4.1 Training Fair DNN with CIA

By adding the generated counterfactual interpolations $\mathcal{X}_{\text{CIA}}$, we obtain our augmented training dataset $\mathcal{X}_{\text{AUG}} = \mathcal{X} \cup \mathcal{X}_{\text{CIA}}$. A reasonable amount of counterfactual interpolations in $\mathcal{X}_{\text{CIA}}$ alleviate the dataset bias caused by the sensitive attribute in $\mathcal{X}$, thus preventing the model from learning biased representation. Finally, we train our debiased model F_{debias} on $\mathcal{X}_{\text{AUG}}$ with the cross-entropy objective:

$$\mathcal{L}_{\text{class}} = \mathbb{E}_{x \sim \mathcal{X}_{\text{AUG}}} \left[- \sum_c y_c \log F_{\text{debias}}(x) \right], \tag{6}$$

where c is the index of the classes.

4.2 Counterfactual Gradients Integration

IG sums gradients over gradual modifications from a baseline to the original input, essentially distributing the total change in model output across gradual input changes. IG's performance heavily relies on the choice of baseline. An arbitrary choice could negatively impact the explanatory power and lead to meaningless explanations. The explanation generated from IG using a black image baseline without the sensitive attribution cannot correctly reflect feature importance in the debiasing learning scenario as illustrated in Figure 1(d).

We propose a gradient-based feature attribution technique, Counterfactual Gradients Integration (CGI), which leverages the counterfactual interpolations generated from CIA to artificially induce a procedure on how the model attention moves across the gradual changes on the sensitive attribute of the input while computing the final prediction score. Thus, CGI can generate explanations regardless of bias while querying a fair DNN model for gradients.

4.3 Path Integral of CGI

IG pre-defines a straight line as the path integral from the baseline x' to the original input x as $\gamma(\alpha) = x' + \alpha(x - x')$, where $\alpha \in [0, 1]$, i.e., $\gamma(0) = x'$ and $\gamma(1) = x$. The baseline x' represents the absence of features. In CGI, we design the path integral as the interpolated path, transiting from the counterfactual sample x' to the input x for generating counterfactual interpolations in CIA, formally: $\gamma(\delta) = g(x, (1-\delta) \cdot s + \delta \cdot s')$, where $g(\cdot)$ denotes the pre-trained generative model and $\delta \in [0, 1]$. We formulate $\text{CGI}_i(x)$ along the

i-th dimension for an input x and its counterfactual example x' as:

$$\mathrm{CGI}_i(x) = (x_i - x'_i) \int_{\delta=0}^{1} \frac{\partial F(\gamma(\delta))}{\partial \gamma_i(\delta)} \frac{\partial \gamma_i(\delta)}{\partial \delta} d\delta. \quad (7)$$

CGI is obtained by accumulating the gradients along the integration path $\gamma(\delta)$ by varying the δ parameter. The model will encounter interpolations on the sensitive attribute from s' to s during the CGI process.

5 Experiments and Results

5.1 Datasets

BiasedMNIST. Following [Arjovsky *et al.*, 2019], we modify MNIST by introducing color (i.e., red and green) as the sensitive attribute correlating strongly (but spuriously) with the target labels in the training set. A fairness-indifferent DNN model can easily achieve high accuracy by only learning the superficial properties (colors) instead of the inherent properties (shapes) for digit recognition. However, such a biased model can fail at inference time when the spurious correlation between the sensitive attribute and the target shifts or vanishes, for example, randomly coloring the digits.

CelebA. The CelebA is a multi-attribute dataset for face recognition with 40 binary attribute annotations for each image. Following [Nam *et al.*, 2020], we select $HeavyMakeup$ and $HairColor$ as target attributes (y) and $Gender$ as the sensitive attribute (s). There is a significant spurious correlation between the target and the sensitive attributes (i.e., most women have blond hair or wear heavy makeup in this dataset). [Nam *et al.*, 2020] compiled two test datasets: unbiased, by selecting the same number of images for every possible value of the pair (y, s), and bias-conflict, by removing all the samples where y and s have the same values from the unbiased set.

5.2 Implementation Details

Architecture details. We employ the LeNet-5 and a pre-trained VGG-16 as the feature extractor along with two fully connected layers as the classification models for BiasedMNIST and CelebA, respectively. The encoder and decoder in CVAE for BiasedMNIST are multi-layered perceptrons consisting of three hidden layers where the latent feature dimension is set to be 2. For CelebA, the encoder of CVAE has 4 $\times$ Conv2D layers with a 3×3 kernel. The decoder consists 4 $\times$ Conv2DTranspose layers with a 3×3 kernel. A batch normalization layer and Leaky ReLu activation function are added after the Conv2D and Conv2DTranspose layers. The latent feature dimension is set to be 128. We add a fourth channel to each image to encode the sensitive attributes.

Training details. We use Adam optimizer throughout all the experiments in the paper. All models are trained with a learning rate of 0.001 and a batch size of 64. We train the classification models for 5 epochs using the cross-entropy loss. We train CVAEs for 50 and 20 epochs for BiasedMNIST and CelebA, respectively, with binary cross-entropy loss as the reconstruction objective. We generate counterfactual interpolations for the whole training set using the pre-trained CVAE following our CIA approach for BiasedMNIST. Since CelebeA dataset is much larger, with more than 160,000 images in the training set, we randomly select 10,000 samples from the training set and generate their counterfactual interpolations.

5.3 Baseline Methods

LAFTR. [Madras *et al.*, 2018] explore adversarial representation learning ensuring group fairness (e.g., demographic parity, equalized odds, and equal opportunity) to different adversarial objectives.

PriorTraining. [Wang *et al.*, 2021] propose a general framework for learning interpretable fair representations by introducing an interpretable "prior knowledge" during the representation learning process. They add an adversarial loss similar to LAFTR as fairness constraints. Another prior loss is used to ensure the interpretable feature learning.

Group DRO. [Sagawa *et al.*, 2019] aim to minimize "worst-case" training loss over a set of pre-defined groups. The authors expect that models that learn the spurious correlation between sensitive attributes and target variables would perform poorly on groups for which the correlation does not hold. By adding a strong regularization on the worst-case groups, Group DRO can prevent the models from learning pre-specified spurious correlations.

LfL. [Sagawa *et al.*, 2019] propose a failure-based debiasing scheme by training a pair of neural networks simultaneously. The first network is trained to be biased by repeatedly amplifying its "prejudice". They debias the training of the second network by focusing on samples that go against the prejudice of the first network.

5.4 BiasedMNIST Results

We compare our method with the vanilla models (Vanilla, plain training without any debiasing procedure), LAFTR [Madras *et al.*, 2018], and PriorTraining [Wang *et al.*, 2021]. We quantitatively assess the effectiveness of different methods via comparing classification performance on training and test sets. The results are shown in Table **??**. The vanilla model heavily relies on the spurious correlation between color (sensitive attribute) and digit (target), so it fails to learn the digit shape during training, resulting in a large accuracy drop on the test set (79.48 $\rightarrow$ 18.08). LAFTR and PriorTraining apply adversarial training to remove sensitive information from the learned features, which may compromise the model performance on the main classification task. Our CIA debiases the training set using counterfactual interpolations and consequently achieves the highest training and test accuracies. The number of generated counterfactual interpolations benefits the performance of CIA, i.e., CIA-30 achieves the best performance.

We qualitatively compare the explanation performance of our CGI with two baselines, IG and BlurIG [Xu *et al.*, 2020], in Figure 4. IG applies a black image as the baseline for gradients integration whereas BlurIG defines the path integral by successively blurring the original input.

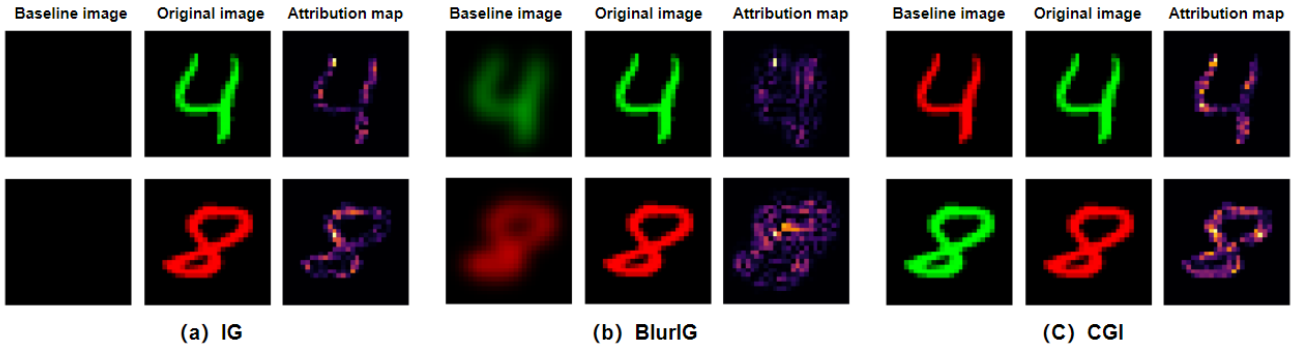


Figure 4: Examples of attribution heatmaps obtained by IG, BlurIG, and CGI. CGI demonstrates to generate higher quality attribution heatmaps with clearer digits shape, less noise, and focuses more densely on the digits (i.e., more bright masks).

Method	Training Acc	Test Acc
Vanilla	79.48	18.08
LAFTR	74.14	75.22
PriorTraining	74.62	75.46
CIA-10	79.64	78.16
CIA-20	79.95	78.23
CIA-30	**79.97**	**78.69**

Table 1: Fairness evaluation on BiasedMNIST. CIA-10, CIA-20 and CIA-30 denotes our CIA method with 10, 20 and 30 generated counterfactual interpolations for each sample, respectively. Best performing results are marked in bold.

5.5 CelebA Results

We compare our method with LfF [Nam *et al.*, 2020] and Group DRO [Sagawa *et al.*, 2019] with results shown in Table **??**. The vanilla model spuriously uses the sensitive attributes for target variable prediction, leading to low accuracies, especially on the bias-conflict sets. Notably, there are large accuracy gaps (i.e., unbiased dataset: 69.14 $\rightarrow$ 85.60 and 61.45 $\rightarrow$ 68.39; bias-conflict dataset: 50.26 $\rightarrow$ 84.17 and 31.56 $\rightarrow$ 50.16) between the vanilla model and our model demonstrating the effectiveness of CIA for bias mitigation. Our model outperforms Group DRO and LfF on most evaluation data sets. We note that CIA is a pre-processing approach that is both algorithm- and model-agnostic. As such, it is compatible with many other in-processing and post-processing fairness algorithms. We mainly demonstrate the advantage of only using CIA coupled with plain training in this work and leave the combination of CIA with other algorithms as our future works.

Figure 5 shows a qualitative example demonstrating CGI is capable of generating higher quality attribution map. Note that there is substantial noise in IG's attribution map due to the arbitrary choice of the baseline. Both CGI and BlurIG have captured the meaningful facial features (e.g., eyes and lips) related to the target attribute $HeavyMakeup$. While CGI's attribution map has higher density masks demonstrating a focus more densely on these facial features.

5.6 Quantitative Performance

We use insertion score and deletion score [Petsiuk *et al.*, 2018] to quantitatively evaluate the interpretation quality of different attribution methods. An attribution method should yield a high insertion score while keeping a low deletion score. We select 1000 samples from BiasedMNIST and 128 samples from CelebA (target variable: $Heavymakeup$) and report the quantitative results in Table **??**. Our CGI outperforms other attribution methods evident by higher insertion and lower deletion scores.

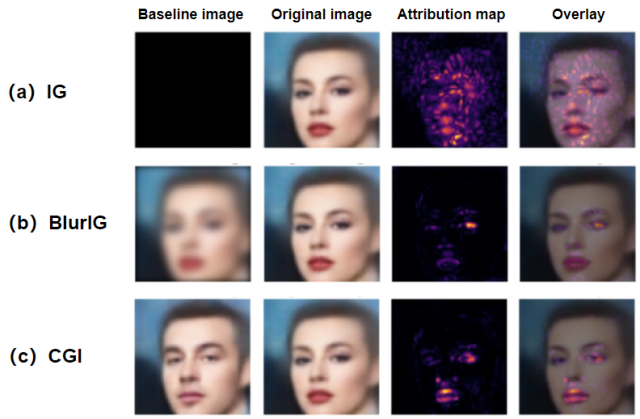


Figure 5: Examples of attribution maps obtained by IG, BlurIG, and CGI. The target attribute is $HeavyMakeup$.

6 Discussion

6.1 Fair Explanation

Although these explanation methods can generate attributions to interpret the model predictions, it is still unclear whether the attributions are generated from the discriminative features or the sensitive attribute since we do not have the ground truth attributions available for evaluation [Zhou *et al.*, 2021]. We illustrate an example from BiasedMNIST in Figure 6 to examine whether these methods are making fair explanations. Both IG and CGI can generate high-quality attribution maps with clear digit shape. While CGI's attribution map clearly shows that the attributions are captured from the digit shape rather than the color. This is because our CGI applies the counterfactual interpolations for gradients integration, which counteracts the effect of the sensitive attribute.

Target	Acc.Type	Vanilla	Group DRO	LfF	**CIA-10**	**CIA-20**	**CIA-30**
HairColor	Unbiased	69.14	85.43	84.24	84.95	85.12	**85.60**
	Bias-conflict	50.26	83.40	81.24	83.16	83.69	**84.17**
HeavyMakeup	Unbiased	61.45	64.88	66.20	67.86	68.04	**68.39**
	Bias-conflict	31.56	**50.24**	45.48	48.07	49.26	50.16

Table 2: Evaluation results on CelebA. *Gender* is the sensitive attribute. The results of Group DRO and LfF are cited from [Nam *et al.*, 2020]. We report the average accuracy over all (y, s) pairs.

Method	BiasedMNIST		CelebA	
	Deletion↓	Insertion↑	Deletion↓	Insertion↑
IG	0.2080	0.5591	0.1038	0.2514
BlurIG	0.2693	0.5014	**0.0638**	0.3016
CGI(ours)	**0.1649**	**0.6253**	0.0746	**0.3264**

Table 3: Quantitative results using deletion score and insertion score.

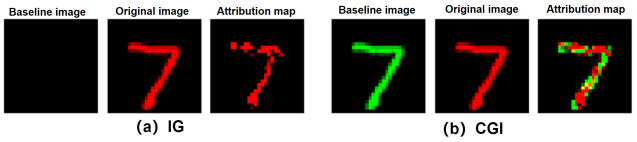


Figure 6: An showcase example demonstrates CGI is capable of generating fair explanation.

6.2 Investigating Saturation Effects

[Miglani *et al.*, 2020] split the area along the integral path as the saturated region where the model outputs changes minimally, and unsaturated region where the model outputs changes substantially. The gradients from the saturated region dominate the calculation of IG. Nevertheless, the integrated gradients of the saturated region seem to be noisier and substantially less faithful than the unsaturated region. Therefore, it is desirable to have a larger unsaturated region to convey feature importance via gradients integration.

Here, we conduct experiment to investigate the saturation regions of IG and CGI. Figure 7 clearly shows that our CGI integrates gradients in a larger unsaturated region than IG does, which contributes proportionately to the computed attribution leading to better explanations as shown in Figure 4. This further demonstrates the effectiveness of our CGI approach in the aspect of gradient saturation effect.

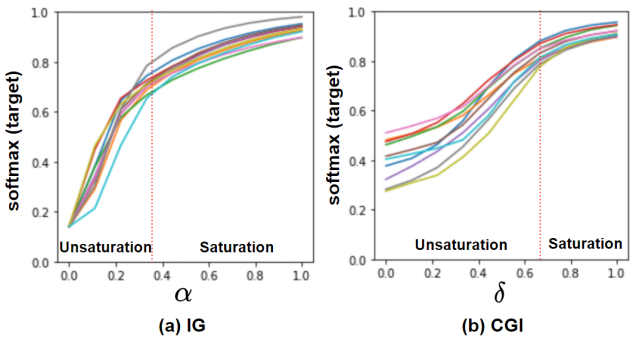


Figure 7: Comparing saturation regions of IG and CGI. We randomly select 10 samples from BiasedMNIST and report the model predictive probability (y-axis) along α and δ (x-axis) integral path.

7 Conclusion

We propose CIA as a pre-processing method to improve DNN's fairness via de-correlating the target variable with the sensitive attribute in training set. CIA generates counterfactual interpolations from a generative model. We further develop a gradient-based feature attribution method leveraging the counterfactual interpolations from CIA to generate high quality and fair explanations. Our experimental results demonstrate the outstanding performance of our approach via quantitative and qualitative evaluations using benchmark datasets. In the future, we will investigate the problem of fair explanation generation with implicit bias mitigation.

Acknowledgments

This work is supported by the National Science Foundation under grant CNS-2043611.

References

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

[Buolamwini and Gebru, 2018] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[Chuang and Mroueh, 2021] Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

[Erion *et al.*, 2021] Gabriel Erion, Joseph D Janizek, Pascal Sturmfels, Scott M Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, pages 1–12, 2021.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, et al. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

[Haug *et al.*, 2021] Johannes Haug, Stefan Zürn, Peter El-Jiz, and Gjergji Kasneci. On baselines for local feature attributions. *arXiv preprint arXiv:2101.00905*, 2021.

[Hong and Yang, 2021] Youngkyu Hong and Eunho Yang. Unbiased classification through bias-contrastive and bias-balanced learning. *Advances in Neural Information Processing Systems*, 34, 2021.

[Izzo *et al.*, 2020] Cosimo Izzo, Aldo Lipani, Ramin Okhrati, and Francesca Medda. A baseline for shapley values in mlps: from missingness to neutrality. *arXiv preprint arXiv:2006.04896*, 2020.

[Kapishnikov *et al.*, 2021] Andrei Kapishnikov, Subhashini Venugopalan, Besim Avci, Ben Wedin, Michael Terry, and Tolga Bolukbasi. Guided integrated gradients: An adaptive path method for removing noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5050–5058, 2021.

[Kim *et al.*, 2019] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 9012–9020, 2019.

[Kim *et al.*, 2021] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14992–15001, 2021.

[Kusner *et al.*, 2017] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[Li *et al.*, 2020] Xin Li, Xiangrui Li, Deng Pan, and Dongxiao Zhu. Improving adversarial robustness via probabilistically compact loss with logit constraints. *arXiv preprint arXiv:2012.07688*, 2020.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

[Madras *et al.*, 2019] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Fairness through causal awareness: Learning causal latent-variable models for biased data. In *FaCCT*, pages 349–358, 2019.

[Miglani *et al.*, 2020] Vivek Miglani, Narine Kokhlikyan, Bilal Alsallakh, Miguel Martin, and Orion Reblitz-Richardson. Investigating saturation effects in integrated gradients. *arXiv preprint arXiv:2010.12697*, 2020.

[Nam *et al.*, 2020] 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.

[Pan *et al.*, 2020] Deng Pan, Xiangrui Li, Xin Li, and Dongxiao Zhu. Explainable recommendation via interpretable feature mapping and evaluation of explainability. *arXiv preprint arXiv:2007.06133*, 2020.

[Pan *et al.*, 2021] Deng Pan, Xin Li, and Dongxiao Zhu. Explaining deep neural network models with adversarial gradient integration. In *Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[Petsiuk *et al.*, 2018] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[Pfohl *et al.*, 2019] Stephen R Pfohl, Tony Duan, Daisy Yi Ding, and Nigam H Shah. Counterfactual reasoning for fair clinical risk prediction. In *Machine Learning for Healthcare Conference*, 2019.

[Qiang *et al.*, 2020] Yao Qiang, Xin Li, and Dongxiao Zhu. Toward tag-free aspect based sentiment analysis: A multiple attention network approach. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.

[Sagawa *et al.*, 2019] 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.

[Sohn *et al.*, 2015] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

[Sturmfels *et al.*, 2020] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

[Tong and Kagal, 2020] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations. *arXiv preprint arXiv:2012.05463*, 2020.

[Wachter *et al.*, 2017] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[Wang *et al.*, 2021] Tianhao Wang, Zana Buçinca, and Zilin Ma. Learning interpretable fair representations. Technical report, Technical report, Harvard University, 2021.

[Xu *et al.*, 2020] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

[Zhang and Sang, 2020] Yi Zhang and Jitao Sang. Towards accuracy-fairness paradox: Adversarial example-based data augmentation for visual debiasing. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 4346–4354, 2020.

[Zhou *et al.*, 2021] Yilun Zhou, Serena Booth, Marco Tulio Ribeiro, and Julie Shah. Do feature attribution methods correctly attribute features? *arXiv preprint arXiv:2104.14403*, 2021.