On Mitigating Bias Using (Caus-)Deep Generative Models

Ajay Gopi

Abstract

Deep learning models achieve impressive results across computer vision tasks. The higher accuracy comes at a cost of poor generality towards minority groups and underrepresented samples in a dataset. Inspired by causality, this research evaluates existing deep generative models designed to generate bias-conflicting counterfactuals, with the goal of improving fairness and robustness of the models towards under-represented groups in a dataset curated across bias ratios of increasing bias.

1 Introduction

Correlation does not imply causation. In deep learning terminology, this means that the correlations made by a deep learning model do not impute an outcome y. Although traditional deep learning models achieve impressive accuracies on benchmarks, often struggle with poor generality towards Underrepresented samples or Out of Distribution samples [2] in a dataset. The higher benchmarks often result from the models making correlations within the dataset that do not account for meaningful representations of the object being modeled. These are often attributed as spurious correlations. Usage of such biased models in real-world applications involving human subjects could lead to amplifying pre-existing societal bias by the model. We conduct a thorough study on two computer vision datasets, Colored MNIST, a synthetically generated dataset with digit and color attributes, where digit is the core feature and color is the spurious feature and MIMIC, a real-world medical image dataset containing patient lung scans, where presence of a disease in a scan is a core feature and the age of the patient is the spurious feature. We evaluate existing Deep Generative Models capabilities and their impact in addressing such dataset bias using principles inspired from Causality literature in Graphical Modeling.

2 Literature Review

2.1 Existing Debiasing Strategies

Although numerous studies have explored debiasing methods, GroupDRO by Sagawa et al. [7] remains a prominent approach for mitigating bias in classifiers. However, they have varying definition on number of epochs to train the subgroups to de-bias a biased model between datasets using the

same technique. Making it challenging for extending their applicability on a new dataset. We mainly investigate approaches that rely on generative-augmentation of bias conflicting minority groups. Jaeju et al [1] and Goel et al [3] in their work demonstrated that minority groups augmentation resulted in significant improvement in test time accuracies for bias-conflicting groups. BiasSwap by Kim et al [4] propose an unsupervied bias feature augmentation for model robustness. However, none of the previous strategies use xAI lens to model Deep Generative Models, with Counterfactual Generative Networks by Sauer et al [8] an exception. They report their metrics on Correlated MNIST dataset but a detailed study is required on their performance on real-world datasets.

2.2 Pearl's Ladder Of Causation and Causal Models

Pearl's ladder of causation involves 3 rungs. The first level Association, dealing with identifying associations in passively observed data. The second level, Intervention, commonly referred to by the do-operator, involves actively inferring what happens if I do action A? The third level, Counterfactual Generation, involves generating plausible outcomes by modeling what would have happened if I did action B instead of action A? We apply the same Counterfactual Generation technique in our work on Generative Modeling to infer bias-conflicting subgroups in the dataset.

2.3 Causality In Deep Learning

In the last few years, there have been several works [5] integrating principles of Causal Inference in Deep Learning to have model interpretability and steerability. Consequently, there have been two notable lines of work in this emerging field, Causal Representation Learning (CRL) and Causal CounterFactual Generations (CCG) where former involves Counterfactual in the Representation Space, the later involves Counterfactual Generation of images. The current work studies the impact of these using the lens of Causal Generative Models, specifically 3 main model families, Causal VAE's, Causal GAN's and Causal Diffusion Models for counterfactual augmentation of bias conflicting minority groups. Conditional Diffusion Generative model [6] are referenced interchangeably as DGMs and the generative capabilities across other families and their limitations are reported in the Conclusion.

3 Methodology

The strategies explored in this study of DGMs are organized into three main sections, each focusing on the methodologies related to Model Architecture, Training and Generation. Each of these topics are covered in depth in subsequent sections.

3.1 Model (Decoder) Architecture

For the Diffusion Model family, we start with a base architecture from Improved Diffusion [6]. Several key modifications were made to make the model architecture compatible for our current setting. These

are covered in subsequent sections.

3.1.1 Dual Attribute Conditioning

The original architecture proposed for Improved Diffusion were designed to condition for single class labels. To modify the architecture, for conditioning on two different attributes, two strategies were initially tested. The first involved concatenating the images with one hot style embedding across subgroups in a dataset. The latter involved slicing the embedding dimension by half, each half used for conditioning the core attribute and the spurious attribute and concatenating the same to the image. The size of the embedding dimension chosen was 100 for cmnist and 4 for MIMIC, compared to the original dimension of 1000 corresponding to number of ImageNet classes. On initial testing for Balanced CMNIST scenario, we noticed an approximately 10% improvement in success rates from employing the embedding slicing strategy. And the rest of the experiments followed the same strategy for integrating conditional signals. And on additional testing, we noticed size of the embedding dimension did not impact the success rates but to maintain consistency with the original architecture, the size of the dimensions were chosen accordingly.

3.1.2 Attention Resolutions

The attention blocks applied to the Score Model in the DGMs were chosen based on the input resolutions of the dataset. For CMNIST, the attention resolutions were applied at scales 7,4 and for MIMIC, the attention resolutions were applied at scales 32, 16, 8 accounting for change in input image resolution of 28*28 for CMNIST and 64*64 for MIMIC. And the number of attention heads for both MIMIC and CMNIST were set to 2. The choice of these hyperparameter were decided after qualitatively inspecting the generations.

3.2 Training

We explore two common strategies usually employed for training Deep Learning Models. Regular Sampling and Over Sampling, based on the weights of the samples seen during the training phase of a DGMs. In Regular Sampling, a DGM is trained with all the training samples having equal weights while Over Sampling upweights the training samples inversely proportional to their class frequencies. We employ both the sampling strategy during the training phase of a DGM, and the models are then subsequently trained on both CMNIST and MIMIC.

3.3 Generation

The trained DGMs are evaluated on their capabilities for minority generations and majority generations. We refer majority generations as the samples produced by the DGM that are bias aligned and minority generations as the samples produced by the DGM that are bias conflicting. An in-depth view

on the nature of the bias is covered in experimental setup. DGMs capabilities of majority and minority generations are further studied across two different settings, Random and Inferred Generations.

3.3.1 Random Generation

This is a standard type of generation normally followed in any generative setting. We first sample noise from a Standard Gaussian Distribution and these are then fed to the decoder to generate subgroups by passing condition signals accordingly. And the generations obtained from the same are referred to as Random Generations.

3.3.2 Counterfactual / Inferred Generation

Generations that are decoded where the latents of a particular attribute serve as start noise are referred to as Inferred / Counterfactual generations. Generating such counterfactuals are further divided into two settings, generation of counterfactuals from core features and generation of counterfactuals from spurious feature depending on the random noise initialized during the sampling time for generating an image.

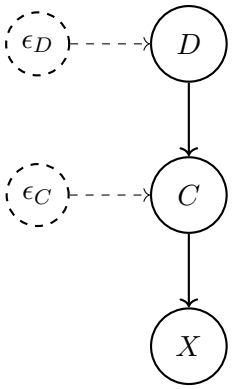


Figure 2: Hierarchical Causal Graph of Digit, Color, and Sample X

Given that VAE's can be viewed as a simplified form of diffusion process, most of the following notations will follow a typical notation used by VAE's, i.e the timesteps would not be explicitly mentioned.

A typical conditional latent variable model formulation involves the following. Let $x \in \mathbb{R}^d$ and $u \in \mathbb{R}^m$ be two observed random variables where x represents the observed samples and u represents the the conditional signal used for latent factorization, and let $z \in \mathbb{R}^n$ be a latent variable, where $n \leq d$. Define the parameters as $\theta = (f, T, \lambda)$ for the following conditional latent generative model:

$$p_\theta(x, z \mid u) = p_f(x \mid z) p_{T,\lambda}(z \mid u)$$

Here T defines the sufficient statics of the latent variable and λ would simply be a lookup table or a dictionary mapping. This means that x can be decomposed as,

$$x := f(z) + \epsilon$$
$$\text{Alternatively,} \quad x^{(k)} := f^{(k)}(\text{pa}(k), u^{(k)})$$

Where, f is a set of structural assignment given the latent prior conditional decomposition and u is the exogenous noise to not have degenerate solutions.

Given this factorization, generation of counterfactual would be done in the following 3 steps. First, abduction of the exogenous noise. This process becomes clearer when we look at the very first step in generative process, we z is usually independently sampled. But, in the case of inferred generation, a particular image with a desired attribute is fed into the encoder to abduct the exogenous noise for that specific image. The second step would be to perform an intervention mechanism, usually denoted by the do-operator in the known causal graph given the factorized latent representations. And the last step would be to generate counterfactuals based on this performed intervention in the generative process. Based on the type of intervention, we can further divide inferred generation as Counterfactual generation from a core feature and Counterfactual generation from a spurious feature. The sampling strategies stated above are applied across all the trained DGMs to sample 25000 generations across each of the unique 54 different sampling combinations.

4 Experimental Design

4.1 Benchmark Datasets for Debiasing

Several datasets, both synthetic and real world datasets are currently used in debiasing research, and are covered in subsequent sections.

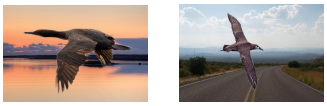
Coloured MNIST (CMNIST), a dataset with task relevant feature i.e shape of the digit, vs task irrelavant features i.e color of the digits is used due to it's simplicity in synthetically generating any combination of bias aligned and conflicting samples as shown in Fig 1a. Similarly, WaterBirds is another synthetically generated dataset by combining bird segmentations from CUB dataset with Places365 as backgrounds to generate bias aligned and conflicting groups as showin in Fig 1b.

MIMIC is a real-world dataset with patient lung scans, in which task relevant feature is presence of a disease vs task irrelevant feature age of the patient or vice versa as shown in Fig 1c.

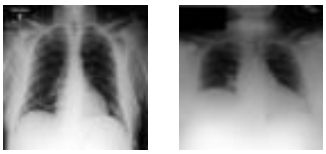
4.2 Data Preprocessing

To study the effect of bias against minority groups, five different versions of the dataset are segregated depending on the amount of bias exhibited on the minority groups within them. To illustrate, consider CMNIST, with digit and color attribute, with each attribute ranging between 0 to 9 values. Subsequently, this would result in 100 unique groups within the dataset. Among the 100 groups, the samples that share the same value for digit and color attribute are addressed as majority subgroups

(a) Dataset: CMNIST
Task-relevant / Core: Digit Shape
Task-irrelevant / Spurious: Digit Color



(b) Dataset: WaterBirds
Task-relevant / Core: Bird Features
Task-irrelevant / Spurious: Background



(c) Dataset: MIMIC
Task-relevant / Core: Presence of Diseased Lung
Task-irrelevant / Spurious: Age of the Patient

Figure 1: Examples of biased datasets with task-relevant and task-irrelevant features.

and conversely, the samples that do not share the same value are addressed as minority subgroups. Depending on the sample distribution in each of these subgroups, the dataset is segregated into Balanced, 95 %, 98 %, 99 %, 99.5 % Versions, each informing the difference in sample count between majority and minority subgroups within the dataset. Similarly, MIMIC shares two attribute, disease and age attributes for values ranging between 0 to 1, each indicating the presence of a disease and old vs young respectively.

4.3 Evaluation Metrics

The evaluation metrics for comparing the DGMs capabilities in generating minority and majority samples are covered in this section. The generated samples are subjected to the following 4 metrics: Success Ratio, FID, Coverage and Density.

4.3.1 Success Rate

The Success ratios are further divided into DGM's capability for successfully generating samples with core feature, spurious feature and both core and spurious feature in the image as three independent metrics. And a high success rate corresponds to better generation quality corresponding to a particular attribute.

4.3.2 FID

FID is a standard metric used to evaluate DGMs capability in modeling the underlying data distribution. We feed both real and fake generated images to an inception model and find the euclidian distance in the feature space. And a higher FID for a generation corresponds to poor generation quality.

4.3.3 Coverage and Density

We additionally use Coverage and Density Metrics to accurately compare the underlying data distribution with the distribution of the generated samples from the DGM. To illustrate, consider P(X) as the data distribution and P(Y) as the data distribution of the generated fake samples, the Density metric evaluates whether a portion of P(Y) can be reconstructed from P(X) indicating DGMs capability in generating samples with high fidelity with respect to the original distribution P(X), and Coverage indicates if a sample from P(X) can be reconstructed from P(Y). The Coverage metric can be interpreted as the DGMs capability in modelling the underlying factors of variation from the True Data Distribution P(X).

4.4 Results

The results presented in Fig 2 pertain to a representative subset of minority generations, averaged over three independent runs. For brevity, only one experiment is detailed, despite the broader scale of the conducted experiments. The results from the fig 2 indicate that as the bias increases in a generative model, the generative model struggle to generate accurate minority samples reflecting the data distribution. The values for coverage, density and success rates follow a downward trend indicating the impact of the increased bias. And the FID shows an upward trend, again indicating the sample generations worsened as the bias increased. To assess the stability of model performance across three independent runs, we conducted a one-way ANOVA, which yielded no statistically significant difference between the run means $p > 0.05$, indicating consistent and reliable results. Overall results across both the datasets CMNIST and MIMIC and their metrics across both random and counterfactual generation are accumulated but will be presented for a conference paper being written. Additionally we notice a similar trend for generative models across other common architectures involving VAE and GAN, but the presented results are pertaining to Diffusion Generative Models.

5 Discussion & Conclusion

This paper introduces a novel approach to addressing bias using causal perspective to generations, which can enhance the robustness and interpretability of decision-making systems in healthcare. Additionally this study shows that, although generative models have improved over the years, their ability to conditionally generate samples is heavily reliant on the amount of supervision signal that

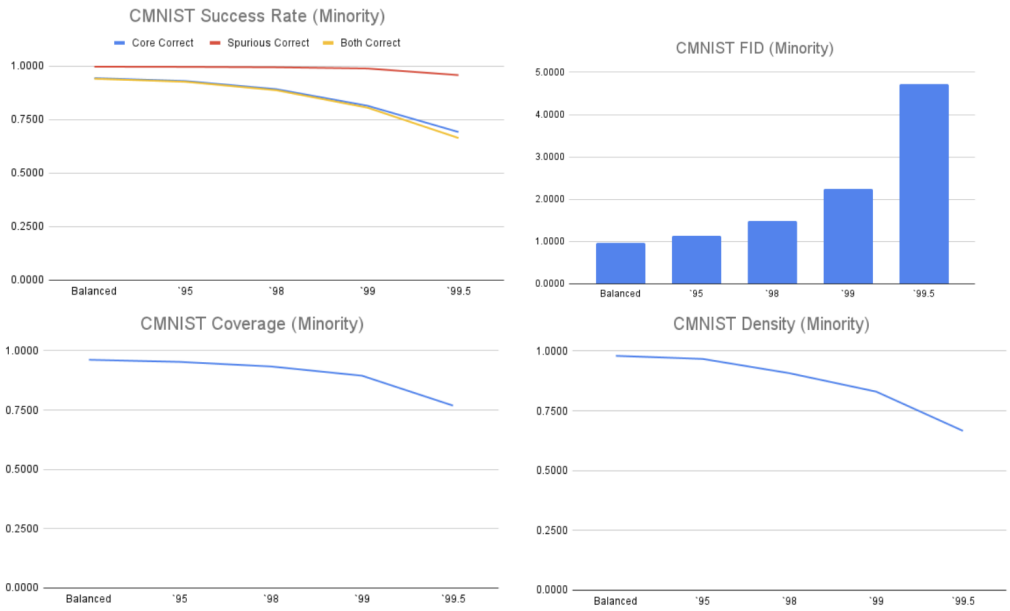


Figure 2: Select Metrics Displayed For Random Minority Generations CMNIST Averaged Across 3 Different Seeds

we add during the training phase of the diffusion models. Additionally, this study also shows that the choice of condition integration mechanism does have impact on generative capabilities of diffusion models. Both these impact the quality of minority generations that can be added to augment the base dataset to improve the robustness against minority samples in a dataset. Additionally, from the results 2, given a clear downward trend across multiple metrics indicate that minority generations is not a viable solution for reducing the bias induced by majority subgroups in a dataset. This shows that better strategies need to be studied to improve robustness of a model. And additional experiments suggested generative models innately struggled generating samples across the WaterBirds dataset due to the diverse representations of bird in a dataset compared to CMNIST and MIMIC, again indicating that addressing bias induced by majority subgroups need to be dealt on a case by case basis. One limitation of this study is the over-reliance on domain specific knowledge to model causal graphs, which may not be feasible when the data has unknown set of confounders and the complexity of the data-distribution as the attributes pertaining to a dataset increase. Given that, the model performance is impacted by the confounding association across bias ratios, this can have a ripple effect as the number of confounding associations increase. A key threat to validity is the lack of evaluation against Causal Representation Learning benchmarks to study how they improve robustness on minority subgroups. Future work could involve applying this framework to real-world observational datasets in genomics to assess its robustness. And study the impact of VLM embeddings used as condition signals for training the generative model. Additionally, incorporating domain-specific priors into the causal graph structure may enhance interpretability, which is critical in health care systems.

6 Acknowledgements

I would like to acknowledge Ruby Shrestha and Nilesh Kumar for their contributions in implementation and interpretations of evaluation metrics used to report results pertaining to the Diffusion Generative Models.

References

[1] Jaeju An, Taejune Kim, Donggeun Ko, Sangyup Lee, and Simon S Woo. Aˆ 2: Adaptive augmentation for effectively mitigating dataset bias. In *Proceedings of the Asian Conference on Computer Vision*, pages 4077–4092, 2022.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2020.

[3] Karan Goel, Albert Gu, Yixuan Li, and Christopher Ré. Model patching: Closing the subgroup performance gap with data augmentation, 2020.

[4] Eungyeup Kim, Jihyeon Lee, and Jaegul Choo. Biaswap: Removing dataset bias with bias-tailored swapping augmentation, 2021.

[5] Aneesh Komanduri, Xintao Wu, Yongkai Wu, and Feng Chen. From identifiable causal representations to controllable counterfactual generation: A survey on causal generative modeling, 2024.

[6] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021.

[7] 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, 2020.

[8] Axel Sauer and Andreas Geiger. Counterfactual generative networks, 2021.