On Mitigating Bias Using Causal Generative Models

Ajay Gopi Linwei Wang

Abstract

While deep learning models have demonstrated superior performance across computer vision tasks, the higher performance comes at the cost of poor generality towards out-of-distribution and / or underrepresented minority samples in a dataset. Causality-inspired research in deep learning presents a lucrative avenue to address these problems by modeling cause-effect relationships in data and providing tools for interventional strategies to answer counterfactual queries. This research aims to evaluate the performance of existing causality-inspired models in datasets that are known to exhibit spurious correlations, with the goal of understanding the ability and limitations of these models to address these biases and to improve interpretability and robust modeling across diverse samples within a given dataset.

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. Traditional deep learning models are known [2] to make such correlations often addressed as spurious with respect to a task to get better train time accuracies, often resulting in poor generalization in real world scenarios.

Consider the scenario, where we need to classify between waterbirds and landbirds, due to the nature of the data collected, most waterbirds are often found in a lake or water background, while the landbirds are often found in forest or land backgrounds. This results in unintended bias during the data collection process. Given a particular type of bird, we have certain task relevant features for eg, shape of the beak and task irrelevant features eg, background of the subject. Causality-Inspired deep generative models aims to model these task relevant features as causal features and promise generating counterfactuals that are bias conflicting to debias models for downstream computer vision tasks such as classification or generations. This research aims to evaluate the generative capabilities of causal deep-generative models, an emerging but understudied area, particularly in biased settings to debias a model.

2 Literature Review

2.1 Existing Debiasing Strategies

Although there have been numerous studies in debiasing groupDRO by Sagawa et al [6] promises to debias a classifier however they have varying definition on how long to train the subgroups to debias a model between datasets using the same technique, 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 [7] an exception, where they report accuracies on correlated mnist dataset but not on real-world datasets.

2.2 Pearl's Ladder Of Causation

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 modelling what would have happened if I did action B instead of action A?

2.3 Causality In Deep Learning

In the last few years, there have been several works [5] integrating causal inference principles 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 counterfactuals in the representation space, the later involves counterfactual generation of images. The current work studies the impact of these causal generative models, specifically 3 main model familes, Causal VAE's, Causal GAN's and Causal Diffusion Models for counterfactual augmentation of bias conflicting minority groups.

3 Methodology

3.1 Causal VAE & Diffusion Models

The current work is on an ongoing work on integrating different types of conditioning signals in Deep Generative models [8] with a known causal graph of the data generative process. The proposed methodology is completely hypothetical, still working on it at the time of writing, building upon existing work on identifiable guarantees using conditional factorization of priors. 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 or conditionally sampled based on conditional factorization. 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 causal process.

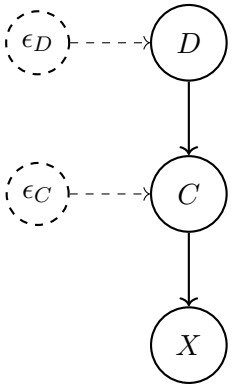


Figure 2: Hierarchical Causal Graph of Digit, Color, and Sample X

There are several current limitations to this approach, given that in the above causal graph, the reverse process can also be true given the data, where color is the parent of digit. Additionally, the presence of confounding association makes isolation of the interventional mechanisms / treatment

assignment mechanism challenging. Exisiting work, for eg, DiffSCM considers a very trivial causal graph where label is the parent in the causal graph of X. And counterfactual inference is performed by obtaining latents from inversion with a desired sample which is then intervened for new unobserved samples. Current work planned is to extend the current formulation of the causal graph to integrate in the deep generative model setting.

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.

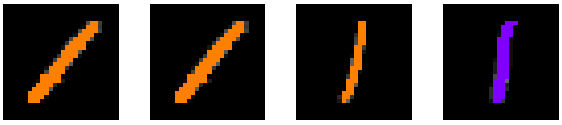
CelebA is a real-world dataset with celebrity faces, in which task relevant feature is hair color vs task irrelevant feature gender or vice versa as shown in Fig 1c.

4.2 Data Preprocessing

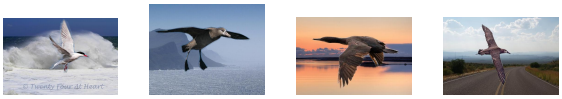
To study the impact of bias on deep learning models, the dataset is prepared in with increasing bias strengths i.e, balanced setting, 95%, 97%, 98% and 99.5%. We consider both synthetic and real-world datasets. For CMnist dataset, there exists datasets that are already synthetically created across these bias ratios. For WaterBirds, the places dataset is combined with CUB segments to synthetically generate bird samples that are present on land backgrounds and water backgrounds. This dataset was further simplified by considering land and water backgrounds that had lesser image features. For eg, for land backgrounds desert and for water background ocean was considered compared to original forest and lake background. The choice for this was motivated after failed experiments in the past to generate samples from the original datasets with complex backgrounds w.r.t image features. And for CelebA we construct biased versions of this dataset by filtering the attributes accordingly between bias aligned and bias conflicting attributes. The attribute that was considered was hair and gender to represent bias ratios.

4.3 Evaluation

A causal generative model is first trained across these bias ratios. Then the generative capabilities of these models are evaluated across bias ratios across two different types of generations random



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



(b) Dataset: WaterBirds
Task-relevant: Bird Features
Task-irrelevant: BackGround

(c) Dataset: CelebA
Task-relevant: Gender
Task-irrelevant: Hair Color

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

and inferred. The former involves generating minority samples from pure noise and the latter involves intervening on latents obtained by a majority sample from the encoder head of a generative model to generate a minority counterfactual. A classifier model trained on a balanced version of the dataset is used as a baseline to evaluate the generative capabilities of these models across bias ratios. The following metrics FID, Density, Coverage, Success Rate of Majority Samples, Success Rate of Minority Samples then computed for each bias setting across three different random runs. To ensure reproducibility all the experiments are seeded both machine level and code level. FID measures the diversity of a generative dataset with a base dataset. Success rate metrics gives a score for the correctly classified generative samples evaluated across majority and minority groups separately. Coverage and Diversity metrics provide an comparative estimate of precision and recall at a manifold level across real and generated data distributions. This experimental is conducted across different families of generative models and variants within the generative model. For eg, for a diffusion family of generative model, the above mentioned experiments are thoroughly repeated across 3 different seeds for conditional generative and the same experiments are repeated for causal generative models. The randomized trails ensures that experiments are not by chance and ensures reproducibility across different runs.

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] 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.

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

[8] Zheyuan Zhan, Defang Chen, Jian-Ping Mei, Zhenghe Zhao, Jiawei Chen, Chun Chen, Siwei Lyu, and Can Wang. Conditional image synthesis with diffusion models: A survey, 2024.