Counterfactual Fairness with Disentangled Causal Effect Variational Autoencoder

Hyemi Kim\(^1\), Seungjae Shin\(^1\), JoonHo Jang\(^1\), Kyungwoo Song\(^1\), Weonyoung Joo\(^1\), Wanmo Kang\(^2\), Il-Chul Moon\(^1\)

\(^1\) Department of Industrial and Systems Engineering
\(^2\) Department of Mathematical Sciences

Korea Advanced Institute of Science and Technology (KAIST), Daejeon, Korea
{khm0308, tmdwo0910, adkto8093, gtshs2, es345, wanmo.kang, icmoon}@kaist.ac.kr

Abstract

The problem of fair classification can be mollified if we develop a method to remove the embedded sensitive information from the classification features. This line of separating the sensitive information is developed through the causal inference, and the causal inference enables the counterfactual generations to contrast the what-if case of the opposite sensitive attribute. Along with this separation with the causality, a frequent assumption in the deep latent causal model defines a single latent variable to absorb the entire exogenous uncertainty of the causal graph. However, we claim that such structure cannot distinguish the 1) information caused by the intervention (i.e., sensitive variable) and 2) information correlated with the intervention from the data. Therefore, this paper proposes Disentangled Causal Effect Variational Autoencoder (DCEVAE) to resolve this limitation by disentangling the exogenous uncertainty into two latent variables: either 1) independent to interventions or 2) correlated to interventions without causality. Particularly, our disentangling approach preserves the latent variable correlated to interventions in generating counterfactual examples. We show that our method estimates the total effect and the counterfactual effect without a complete causal graph. By adding a fairness regularization, DCEVAE generates a counterfactual fair dataset while losing less original information. Also, DCEVAE generates natural counterfactual images by only flipping sensitive information. Additionally, we theoretically show the differences in the covariance structures of DCEVAE and prior works from the perspective of the latent disentanglement.

Introduction

Machine learning has penetrated our lives so deep, and its fairness and societal utilization have become a growing concern in our society (Aleo and Svirsky 2008; Kim, Ghorbani, and Zou 2019). The incident of COMPAS (Brennan, Dieterich, and Ehret 2009) shows that the learning model can be a source of unfairness in our judicial system by discriminating people by race. Given that the learners become unfair only because of training data (Hardt, Price, and Srebro 2016), we ask the question of whether it is feasible to correct its unfairness from the data or not. Particularly, our concept of unfairness comes from our societal principle on equal treatments across races, gender, religion, etc., a.k.a. sensitive variables, without prejudice. Then, the key question on the machine learning research becomes whether we can separate such prejudices embedded in the data algorithmically or not.

Considering the prejudice by the sensitive variable, the causal inference is an interesting tool to separate the factors contributing to decision-making. The objective of causal inference is learning the causal effect of an intervention variable, \(a\), on individual features, \(x\), and an outcome, \(y\). Here, if we regard the intervention variable in causality as the sensitive variable in decision-making, the learning fairness can be formulated as the causal inference task (Zhang, Wu, and Wu 2018; Chiappa 2019; Wu, Zhang, and Wu 2019; Kilbertus et al. 2017). For example, a causal model estimates the effect of sensitive variables, such as race and gender, on an admission result (Kusner et al. 2017). Another study shows a causal model predicting a medication’s effect on a patient’s prognosis (Pfohl et al. 2019). If we focus on modeling the exogenous uncertainty with Variational Autoencoder (VAE)
is the set of endogenous variables, and \( \hat{Y} \) becomes two different variations by either \( a \) or \( \neg a \). The counterfactual fairness asserts that the estimated distributions on \( \hat{Y} \)'s should be identical regardless of the sensitive value, \( a \).

**Causality with Variational Autoencoder**

Amidst the limitations of model-based approaches with a single latent variable to model all features complex counterfactual example generations because the negation only affects \( a \), not the embedded components in \( u \).

\[
\mathcal{M}_{CEVAE} = \mathbb{E}_{q(u|a,x,y)}[\log p(u) + \log p(a,x|u) + \log p(y|a,u) - \log q(u|a,x,y)] + \log q(a^*|x^*) + \log q(y^*|x^*, a^*) ,
\]

where \( x, a^*, y^* \) being the observed values in the training set.

To compensate this potential problem in CEVAE, modified version of CEVAE (Pfohl et al. 2019), or mCEVAE, assumed that \( x \) and \( y \) are caused by \( a \) and \( u \). \( \mathcal{M}_{mCEVAE} \) uses the maximum mean discrepancy (MMD) to regularize the generations to remove the information of \( a \) from \( u \), but this MMD regularization removes \( u \) components that is simply correlated to \( a \), not caused by \( a \).

\[
\mathcal{M}_{mCEVAE} = \mathbb{E}_{q(u|a,x)}[\lambda_x \log p(x|a,u) + \lambda_y \log p(y|a,u)] + \lambda_1 D_{MMD}(q(u||p(u)) + \lambda_2 \sum_{a_k \in A} D_{MMD}(q(u|a = a_k)||p(u)) ,
\]

We hypothesize that the accurate counterfactual generation lies in the middle of these two models. We define \( x_{d} \subset x \) is a subset of features caused by \( a \) whereas \( x_{r} \subset x \) is the other subset of irrelevant features to the intervention. Similarly, we define the exogenous variables of \( x_{r} \) and \( x_{d} \) to be \( u_{r} \) and \( u_{d} \), respectively. When the counterfactual generation is required, the intervention on \( a \) should be imposed on \( x_{d} \), and \( x_{r} \) should be maintained. This disentanglement is the fundamental motivation of our model, DCEVAE.

Also, CEVAE separates the decoder network into two alternative functions: \( f_{a=0} \) and \( f_{a=1} \). Therefore, either decoder tends to be updated according to the observed sensitive variables. For example, for data \( (x_i, a_i = 1, y_i) \) is utilized to learn the parameters of \( f_{a=1} \). When CEVAE makes the counterfactual samples, the latent values from a data instance with \( a \) are propagated to the decoder \( f_{a=0} \). However, the lack of examples of \( a \) in the training process of \( f_{a=0} \) can cause inaccuracy. To resolve this issue, mCEVAE regularizes latent variables \( u_{a=0} \) and \( u_{a=1} \) to be similar by using...
Disentanglement with Total Correlation

A latent variable, $u$, is considered to be disentangled if $u_j$ are independent when $j$ indicates a dimension of the variable. The total correlation is $TC = KL(q(u)||\Pi_{j=1}^d q(u_j))$, where $d$ is a dimension of $u$. When we minimize the TC, the latent is disentangled (Kim and Mnih 2018).

Methodology

Causal Structure of Disentangled Causal Effect Variational Autoencoder

We design the structure of DCEVAE from the causal graph in Figure 2a. The causal graph specifies the separated causalities from the sensitive variable, $x$, to the feature variables, $x$. Hence, $x$ is divided into the feature variables caused by sensitive information, $x_d$, and the other feature variables, $x_r$. This paper assumes that the causal graph gives the attribute association to either $x_r$ or $x_d$. For instance, let us assume that there is a woman (a: Gender) who went to a women’s only school (x_d: School). When we compute counterfactual value by intervening $a = Male$, then the value of school will change. However, this person’s birth year (x_r), which is not a descendant of gender, will not be changed. This consideration enables a more general causal graph ordering assumption to be incorporated for counterfactual generation based on VAEs.

Bayesian Network of Disentangled Causal Effect Variational Autoencoder

The causal graph in Figure 2a translates to a Bayesian network in Figure 2c. Two exogenous variables, $u_r$ and $x_d$, in the causal graph are translated into two latent variables in the Bayesian network. This Bayesian network corresponds to the neural network in Figure 2h. The neural network consists of an inference structure on two latent random variables ($q_{\phi_r}$ and $q_{\phi_d}$), and the neural network also includes a structure for disentanglement, as $D_{\psi}$. The objective function of DCEVAE, Eq 4, is devised to satisfy the above model structure. The latent variables are inferred by the optimization of $M_{ELBO}$. The disentanglement of the latent variables is solved by reducing the total correlation, $L_{TC}$. In the computation of the total correlation, we use the discriminator $D_{\psi}$ from $M_D$, so we add an optimization on $M_D$ to our objective function, as well. Eventually, our objective becomes the min-max structure to correspond to the counterfactual generation and the latent disentanglement.

$$\min_{\theta,\phi} L_{DCEVAE} := -M_{ELBO} + \beta_{tc} L_{TC}$$

$$\max_{\psi} M_D$$

(4)

Counterfactual Inference and Counterfactual Example Generation

Besides the model structure, a causal graph requires the inference on the exogenous variable to estimate the causal effect, including the total effect and the counterfactual effect.
We match two types of exogenous variables, $u_d$ and $u_r$, in the causal graph to two corresponding latent variables of DCEVAE. According to the counterfactual inference [Pearl 2009], a counterfactual $y$ can be inferred in following steps:

1. **Abduction** Infer the distribution of $u_d$ and $u_r$ with the encoder network of DCEVAE: $q(u_d|a,x_d,y)$ and $q(u_r|a,x_r,y)$.

2. **Action** Substitute $A$ with $\neg a$.

3. **Prediction** Compute the probability of counterfactual $y$, with the decoder network of DCEVAE, $p(y|\neg a,u_d,u_r)$.

In the third step, prediction, $x_d$ could be potentially change for intervention $\neg a$, but $x_r$ maintain its value because we consider causal ordering as $u_r \rightarrow x_r$ and $(a,u_d) \rightarrow x_d$.

The identification is shown as follows.

**Proposition 1** If we recover $p(u_d,u_r,a,x_d,x_r,y)$, then we recover counterfactual effect under the causal model Fig[27]

**Proof 1** $p(y | x_d, x_r, do(a = 1))$

$$= \int_{u_d} \int_{u_r} p(y | x_d, x_r, do(a = 1), u_d, u_r) \times p(u_d, u_r | x_d, x_r, do(a = 1)) du_d du_r$$

$$= \int_{u_d} \int_{u_r} p(y | x_d, x_r, a = 1, u_d, u_r) \times p(u_d, u_r | x_d, x_r, a = 1) du_d du_r$$

$p(y|x_d,x_r,a=1,u_d,u_r)$ and $p(u_d,u_r|x_d,x_r,a=1)$ can be identified from the distribution of $p(u_d,u_r,a,x_d,x_r,y)$.

**Evidence Lower Bound of DCEVAE**

We propose the ELBO to disentangle $u_d$ and $u_r$ by following the Bayesian network structure. Figure [27] We assume that $x_d$ and $u_r$ is independent given $u_d$; and $x_r$ and $(a,u_d)$ are independent given $u_r$, as well. Then, the decoder distribution, $p(y|x_d,x_r,u_d,u_r|a)$, can be factorized as the below:

$$p(y|x_d,x_r,u_d,u_r|a) = p(u_d)p(u_r)p(y|x_d,a|u_d) \times p(y|x_r,u_r)p(y|a,u_d,u_r)$$

(5)

Also, we assume that the posterior, $q_\phi(u_d,u_r|a,x_d,x_r,y)$, can be factorized as the Eq.[6]

$$q_\phi(u_d,u_r|a,x_d,x_r,y) = q_\phi(u_d|a,x_d,y)q_\phi(u_r|a,x_r,y)$$

(6)

Given an approximate posterior $q_\phi(u_d,u_r|a,x_d,x_r,y)$, we obtain the variational lower bound as Eq.[7]

$$\log p_\theta(y|x_d,x_r,y|a)$$

$$\geq E_{q_\theta(u_d|x,a,y)q_\phi(u_r|x,a,r,y)} [\log p_\theta(y|a,u_d,u_r)]$$

$$+ E_{q_\phi(u_d|a,x,y)} [\log p_\phi(y|x_d,a|u_d)]$$

$$+ E_{q_\phi(u_r|a,x,y)} [\log p_\phi(y|x_r,u_r)]$$

$$+ KL(q_\phi(u_d|a,x,d,y)||p(u_d)) + KL(q_\phi(u_r|a,x,r,y)||p(u_r))$$

$$=: \mathcal{ELBO}$$

(7)

In practice, we use the neural network layers to infer the parameters of a Gaussian distribution over the joint space of $u_d$ and $u_r$. To obtain the posterior distribution of $q_\phi, p(u_d)$, and $p(u_r)$ are the prior distributions following the Gaussian distribution, and we utilize the reparameterization, accordingly. The below is the encoder structure for $u_d$ and $u_r$.

$$p(u_d) = \mathcal{N}(u_d|0,I); \quad p(u_r) = \mathcal{N}(u_r|0,I);$$

$$q_\phi(u_d|a,x,d,y) = \mathcal{N}(\mu_d, \sigma_d^2 I); \quad q_\phi(u_r|a,x,r,y) = \mathcal{N}(\mu_r, \sigma_r^2 I);$$

$$\mu_d = g_\theta^d(a,x,d,y); \quad \sigma_d = g_\phi^d(a,x,d,y); \quad \mu_r = g_\theta^r(a,x,r,y); \quad \sigma_r = g_\phi^r(a,x,r,y).$$

(8)

With decoder $p_\theta$, DCEVAE provides the generative process for $a, x_d, x_r, y$. Since $a, x_r, y$ are different in nature, we differentiate their distributions. We assume that if $x$ are continuous variables, they follow the Gaussian distribution. We let $y$ be the binary variable of the Bernoulli trial. $x_d$ and $y$ are influenced by $a$. On the other hand, $x_r$ is only determined by $u_r$ and not determined by $a$. The decoder structure for a factual data instance is specified as below:

$$p_\theta(y|x_d,a,u_d) = \mathcal{N}(\mu - \mu_r, \sigma_r^2)$$

$$p_\theta(x_r|u_r) = \mathcal{N}(\mu_r, \sigma_r^2); \quad \mu_r = g_\phi^r(a,u_d, u_r)$$

$$\mu_r, \sigma_r^2 = f_\phi^r(a,u_d) \quad f_\psi^r(u_r)$$

$$p(y|a,u_d,u_r) = \text{Bern}(\pi = \hat{\pi}_y); \quad \hat{\pi}_y = f_\phi(a,u_d,u_r)$$

For counterfactual data generation, we use the same encoder and decoder structure. The only difference is using $\neg a$ for a decoder output. As mentioned before, a counterfactual $x_r$ is same with the factual $x_r$, while a counterfactual $x_d$ and $y$ are influenced by the counterfactual $\neg a$.

**Disentanglement Loss of DCEVAE**

Up to this point, we treat each pair $(a, u_d)$ and $(u_d, u_r)$ to be disentangled. This assumption minimize $\mathcal{L}_{TC} = KL(q(a,u_d,u_r)||q(a,u_r)q(u_d))$. This KL divergence is intractable because both $q(a,u_d,u_r)$ and $q(a,u_r)q(u_d)$ are conditioned on $a$, $x_d$, and $x_r$. Therefore, we take an alternative approach adopted in FactorVAE [Kim and Mnih 2018]. Algorithm 1 in Appendix 3.3 specifies the sampling from $q(a,u_d,u_r)$ and $q(a,u_r)q(u_d)$ under the conditions, and we apply the permutation to minimize $\mathcal{L}_{TC}$. $\mathcal{L}_{TC}$ is related to the discriminator $D_\psi$ by the density-ratio trick, which is approximated by a neural network. The output of $D_\psi([a,u_d,u_r])$ estimates the probability when the density function takes a sampled input from $q(a,u_d,u_r)$, rather than from $q(a,u_r)q(u_d)$. $\mathcal{L}_{TC}$ is expressed by $D_\psi$ as in Eq.[10]

$$\mathcal{L}_{TC} = KL(q(a,u_d,u_r)||q(a,u_r)q(u_d))$$

$$\approx E_{q(a,u_d,u_r)} [\log \frac{D_\psi(a,u_d,u_r)}{1 - D_\psi(a,u_d,u_r)}].$$

(10)

For training the network $D_\psi$, we should maximize $\mathcal{M}_{D_\psi}$.

$$\mathcal{M}_{D_\psi} = E_{q(a,u_d,u_r)} [\log (D_\psi([a,u_d,u_r]))]$$

$$+ E_{q(a,u_r)q(u_d)} [\log 1 - D_\psi([a,u_d,u_r])].$$

(11)

Appendix 3.3 provides a whole algorithm of DCEVAE including the minimization of $\mathcal{L}_{DCEVAE}$ and the maximization phase of $\mathcal{M}_{D}$. 


Theoretic Analysis on Covariance Structure

Eq. [12] defines the posterior distribution of the latent variable, and we derive the stationary point of $\Sigma$. (Lucas et al. 2019).

$$q(u|a, x, y) = N(V_u ([a, x, y] - \mu), \Sigma)$$  \hspace{1cm} (12)

Here, $u$ is $[u_r, u_d]$, and $\Sigma$ is a covariance matrix of the joint distribution of latent variables. $\Sigma$ is a covariance matrix of the permuted $u$ by discriminator $D_q$.

If $[u_r, u_d]$ is well disentangled, $\Sigma$ should show two block diagonal matrices corresponding to $u_r$ and $u_d$. To theoretically analyze the disentanglement, we provide Proposition 2 and Eq. [13].

**Proposition 2** Let $\Sigma^*$ be the stationary point of the $\mathcal{L}_{DCEVAE}$. For a linear DCEVAE, $\Sigma^*$ has the form:

$$\Sigma^* = \left\{ \frac{1}{1 + \beta} \left( \frac{1}{\sigma^2} W^T W + \frac{1}{\sigma^2} W^T M'M + \mathbf{I} \right) \right\}^{-1}$$

with $n = |u_r|$ and $m = |u_d|$.

**Proof 2** see Appendix 1.1 for the proof.

We observed that $u_r$ and $u_d$ has their distinct covariance blocks from the real dataset due to the masking effect of $M_r$ and $M_d$. This theoretically shows the disentangling effects of DCEVAE. Despite the clear disentangling effect from the masks, the off-diagonal covariance can be feasible by $\Sigma$, so we provide Corollary [11]. Also, the part of covariance, $M_d^T \Sigma$, is constructed to be independent to $a$ by the Total Correlation loss, $\mathcal{L}_{TC}$, which enforces $u_d$ to be independent to $a$.

**Corollary 1** As $\beta \to \infty$, $\Sigma^*$ becomes a covariance matrix with two blocks on diagonal.

**Proof 3** $\beta \to \infty \Rightarrow \frac{1}{1 + \beta} \to 0$, $\frac{1}{1 + \beta} \to 1 \Rightarrow \Sigma^* \to \Sigma$. Also, $\Sigma$ is designed to permute $u_d$ and $u_r$ by following Line 8-11, Algorithm 1, Appendix 3.3; cov$(u_d, u_r) \to 0$. Therefore, $\Sigma^*$ has two blocks of $u_d$ and $u_r$, dimensions.

Figure 3c and 3d contrast the covariance structure of DCEVAE to the CEVAE. There is no disentanglement effect on the covariance of CEVAE, so it cannot distinguish the correlated latent variable from the caused ones. Theoretic analyses on the covariance of CEVAE are in Appendix 1.2.

![Figure 3](image-url)

Figure 3: (a,b) The covariance matrix of sampled latent values from DCEVAE with $\beta = 0$ and 1000. The dimension of $u_d$ is five, and so is $u_r$. (c,d) Each covariance matrix $\Sigma_0$ and $\Sigma_1$ of sampled latent values from each decoder of CEVAE.

**Application**

**Causal Fair Classification** The task of causal fairness requires estimating $y$ with the minimized influence of $a$. When we assume $p(y|a, u_d, i, u_r, i) = 0$ for $u_d, i$ and $u_r, j$ from $i$-th data instance; we say that the counterfactual fairness is satisfied for the data instance. Therefore, we alter the objective function of DCEVAE by adding a regularization $\mathcal{L}_f$ as the below:

$$\min \mathcal{L}_{fair} = \mathcal{L}_{DCEVAE} + \beta \mathcal{L}_f, \max M_D$$

After optimizing Eq.[14] we train a classifier, such as a logistic regression (LR), with the pairs of $\hat{y}$ with $x$ through the decoder, $p_\theta(x_d, x_r, y|a, u_d, u_r)$ and $p_\theta(x_d, x_r, y|\neg a, u_d, u_r)$. The test procedure utilizes the raw input of the testing feature $a$ and $x$ without any modifications.

**Counterfactual Image Generation** The counterfactual image generation task allocates $y$ to be the image and $x$ to be the labels which describe the image, so is the label which we want to intervene on. Unlike the fairness dataset, we modified the encoder, $q_{\theta_d}(u_d, u_r|a, y)$, because the information of $x$ is already embedded on an image, $y$. The counterfactual images are sampled from the decoder, $p_\theta(y|\neg a, u_d, u_r)$, while $u_d$ and $u_r$ are obtained from the encoder.

**Experiments**

**Datasets and Baselines**

Appendix 4.1 provides the details of datasets; and Appendix 4.2 enumerates the causal graphs and their paired attributes, $x_d$ and $x_r$.

**Causal Estimation and Fair Classification** We use the UCI Adult income (Asuncion and Newman 2007) dataset for causal estimation and fair classification tasks. We treat gender as a sensitive variable (or intervention) $a$; income as the outcome $y$; race, age, and native country as $x_d$; and other variables as $x_r$. We benchmark the causal effect estimation on five baselines: CausalGAN (Kocaoglu et al. 2017), CausalGAN-Incomplete (CausalGAN-IC), conditional VAE (CVAE) (Sohn, Lee, and Yan 2015), CEVAE, and mCEVAE. CausalGAN-IC has the same generator and discriminator structures as CausalGAN, but CausalGAN-IC used the same causal graph DCEVAE used. For fairness experiments, we additionally use Unawareness (CF-UA), Additive Noise (CF-AN) (Kusner et al. 2017), and CFGAN (Xu et al. 2019) as baselines.

**Counterfactual Image Generation** We use the CelebA dataset (Liu et al. 2018) for the counterfactual image generation. We treat Mustache as an intervention attribute $a$; an image as $y$; and the other attributes as $x$. Appendix 5.1 provides a similar experiment with Smiling as an intervention attribute. We separate $x_d$ and $x_r$ as the assumed causal graph in Figure 4. We consider the following baselines for counterfactual image generations: CVAE, CEVAE, mCEVAE, Conditional GAN (CGAN) (Mirza and Osindero 2014) with Wasserstein distance (Arjovsky, Chintala, and Bottou 2017) (denoted as CWGAN), CausalGAN, and CausalGAN-IC.
Table 1: The total effect and counterfactual effect of real and generated datasets (\(O = \{\text{race}, \text{native country}\}\)). CE error is \(\sum_{i,j \in \{0,1\}} |\hat{o}_{ij} - o^*_{ij}|\) with true CE, \(o^*\). The numbers in bold indicates the best performance, and the underlined numbers indicate the second best performance.

| Evaluation Metrics |
|---------------------|
| Causal Estimation and Fairness Task |
| We evaluate the performance on the causal estimation and fairness task with following metrics used in CFGAN: **Total effect**, \(TE(a_1, a_0) = \tilde{P}(y_{a_1}) - \tilde{P}(y_{a_0})\), measures the change likelihood of \(A\) from \(a_0\) to \(a_1\) on \(Y\). **Counterfactual effect**, \(CE(a_1, a_0|o) = \tilde{P}(y_{a_1}|o) - \tilde{P}(y_{a_0}|o)\), is the total effect conditioned on the observation, \(o\). A **Logistic Regression (LR)** and a **Support Vector Machine (SVM)** are trained with generated datasets from the model, and their test accuracy is also used from the original dataset. **Chi square distance** (\(\chi^2\)) indicates the similarity between the generated and the real datasets (Daliri 2013). We follow MaskGAN (Lee et al. 2019) to evaluate quality of counterfactually generated images: **Semantic-level Evaluation** evaluates generated images by measuring the preservation of the original values of \(x_r\). We trained a classifier with ResNet-18 (He et al. 2016) to examine the accuracy of \(x_r\) in generated images. **Distribution-level Evaluation** measures the quality and the diversity of generated images, and we used the Frechet Inception Distance (FID) (Cao et al. 2013). **Identity Preserving Evaluation** evaluates the identity preservation ability, so we conducted a face matching experiment for whole pairs of original and counterfactual images with ArcFace (Deng et al. 2019). We use 1,544 pairs for **Mustache** and 10,000 pairs for **Smiling**.

Results

Causal Estimation and Fairness Task

Without the regularization of the fairness \(L_{fair}\), we calculate the total effect (TE) and the counterfactual effect (CE) of our model and baselines in Table 1 for UCI Adult. DCEVAE estimates the total effect and the counterfactual effect close to the original dataset through DCEVAE does not know the exact causal graph structure, unlike CausalGAN. CausalGAN estimates the true TE and CE well, but CausalGAN-IC has lower CE and TE estimation accuracies when the incomplete causal graph is given. CEVAE has lower performance on TE and CE estimations compared to DCEVAE, which is caused by the latent variable in CEVAE with the correlated information from \(o\) to \(X_d\). Therefore, CEVAE maintains correlated information even in the counterfactual prediction.

Generated Counterfactual Images

Figure 5 visualizes the counterfactual images from the VAE based models for the real image. VAE based models can infer the exogenous variables from a specific real image. Counterfactual images from DCEVAE preserve the identity of real images except...
Table 2: (1) Real: comparison between real images and generated counterfactual examples; Pair: comparison between reconstructed images and counterfactual images with the same latent values (2) Target Accuracy, (3) Identity Preserving (IP) scores (threshold = 0.2 for Real, and 0.6 for Pair), (4) The attribute classification accuracy for unaltered labels, (5) FID Score (FID). The numbers in bold indicate the best performance, and the underlined numbers indicate that the second best performance.

for the intervention variables. For example, female images on the first row in Figure 5 are kept to be female while the intervened Mustache (M) is added. On the contrary, CVAE makes images with Mustache, but its gender is altered. On the other hand, CVAE creates blurry images because CVAE separates the images of real images in a latent space. However, we did not include the real-world instances. Our model, DCEVAE, estimates counterfactual images because the GAN-based counterfactual generation because the GAN variations cannot produce the counterfactual image matched to a real-world image.

Quantitative Analysis on Image Generation Task This section shows the quantitative evaluations on generated counterfactual images with label classifier accuracies, Frechet Inception Distance (FID) score, and identity preserving (IP) metrics. Table 2 shows the result of counterfactual generation on Mustache. We compare the real and the counterfactual generated images of VAE-based approaches. DCEVAE has the highest attribute classification accuracies, so DCEVAE maintains the attribute of $x_r$ intact. Also, DCEVAE has the highest IP and the lowest FID scores, so the generated images are evaluated to be more natural than the other models. The generation of Mustache is also measured by a classifier accuracy, and CVAE, mCEVAE, and DCEVAE have similar accuracies. We compare reconstructed images and counterfactual images from VAEs and GANs. Except FID score, DCEVAE preserves attributes causing Mustache, i.e. preserved Male attribute.

![Figure 5](image.png)

Figure 5: (rows 1-3) Real image and its counterfactual images for Mustache from DCEVAE, CVAE, CEVAE, and mCEVAE. (row 4) tSNE (Maaten and Hinton 2008) of latent variables $u_r$, $u_d$ in each model. Here, $M = 0$ (yellow) and $M = 1$ (blue).

Conclusion

This paper disentangles the exogenous uncertainty into two latent variables of 1) independent to interventions ($u_d$), and 2) correlated to interventions without causality ($u_r$). The disentanglement of latent variables resolves the limitation in previous works, including maintaining causality from the intervention ($a$) and altering all correlated information for counterfactual instances. Our model, DCEVAE, estimates the total effect and the counterfactual effect without a complete causal graph. In experiments, we showed that DCEVAE is comparable with other models with and without the complete causal graph. Both applications on the fair classification and the counterfactual generation showed the best quantitative performance by utilizing a counterfactual instance matched to the real-world instances.
Ethical Impact

Besides the COMPAS incident (Brennan, Dieterich, and Ehret, 2009), governments and corporates utilize the AI-based screening and recommendation systems on a massive scale, and these applications are prone to the fairness question, particularly when the subject individual has minority backgrounds. This paper discusses the triad of 1) the fair classification, 2) the causality-based counterfactual generation, and 3) the latent disentanglement. If we were to maximize the classification accuracy, the proposed method would be irrelevant from such efficiency-oriented perspectives. However, our society always asks what-if questions, i.e., the veil of ignorance by Rawls (Rawls, 2009). The limitation of the accuracy can be acceptable in two conditions: 1) the damage to the accuracy performance should be controlled and minimal, and 2) the limitation satisfies the argument that ”I would accept the classification result under my altered background.” This ”altered background” is, in fact, an identical argument to the justice concept suggested by Rawls, which argues designing a taxation concept before determine whether you will be born in either high-income or low-income families. An emerging question is whether or not an AI can come up with a justifiable altered concept on what-if scenarios, so we work on the counterfactual generation to satisfy the fairness concept defined in the above. This counterfactual generation is further elaborated if an AI carefully dissect the context of an individual subject, and this is the disentanglement process when causality and a correlation should be distinguished. Given this series of arguments and necessity, this work is an important contribution in promoting the fairness of deployed AI systems, which are already running without a user’s perception of its background operation.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(NRF-2018R1C1B600865213)

References

Aleo, M.; and Svirsy, P. 2008. Foreclosure fallout: The banking industry’s attack on disparate impact race discrimination claims under the fair housing act and the equal credit opportunity act. BU Pub. Int. LJ 18: 1.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein gan. arXiv preprint arXiv:1701.07875.

Asuncion, A.; and Newman, D. 2007. UCI machine learning repository.

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. Criminal Justice and Behavior 36(1): 21–40.

Cao, C.; Weng, Y.; Zhou, S.; Tong, Y.; and Zhou, K. 2013. Facewarehouse: A 3d facial expression database for visual computing. IEEE Transactions on Visualization and Computer Graphics 20(3): 413–425.

Chiappa, S. 2019. Path-specific counterfactual fairness. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, 7801–7808.

Daliri, M. R. 2013. Chi-square distance kernel of the gaits for the diagnosis of Parkinson’s disease. Biomedical Signal Processing and Control 8(1): 66–70.

Deng, J.; Guo, J.; Xue, N.; and Zafeiriou, S. 2019. Arcface: Additive angular margin loss for deep face recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4690–4699.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. In Advances in neural information processing systems, 3315–3323.

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.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In Advances in Neural Information Processing Systems, 656–666.

Kim, H.; and Mnih, A. 2018. Disentangling by factorising. arXiv preprint arXiv:1802.05983.

Kim, M. P.; Ghorbani, A.; and Zou, J. 2019. Multiaccuracy: Black-box post-processing for fairness in classification. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 247–254.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114.

Kocaoglu, M.; Snyder, C.; Dimakis, A. G.; and Vishwanath, S. 2017. Causalgan: Learning causal implicit generative models with adversarial training. arXiv preprint arXiv:1709.02023.

Kusner, M. J.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. In Advances in Neural Information Processing Systems, 4066–4076.

Lee, C.-H.; Liu, Z.; Wu, L.; and Luo, P. 2019. MaskGAN: towards diverse and interactive facial image manipulation. arXiv preprint arXiv:1907.11922.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2018. Large-scale celebfaces attributes (celeba) dataset. Retrieved August 15: 2018.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In Advances in Neural Information Processing Systems, 6446–6456.

Lucas, J.; Tucker, G.; Grosse, R. B.; and Norouzi, M. 2019. Don’t Blame the ELBO! A Linear VAE Perspective on Posterior Collapse. In Advances in Neural Information Processing Systems, 9408–9418.

Maaten, L. v. d.; and Hinton, G. 2008. Visualizing data using t-SNE. Journal of machine learning research 9(Nov): 2579–2605.
Mirza, M.; and Osindero, S. 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Pearl, J. 2009. Causality. Cambridge university press.

Pföhl, S.; Duan, T.; Ding, D. Y.; and Shah, N. H. 2019. Counterfactual Reasoning for Fair Clinical Risk Prediction. arXiv preprint arXiv:1907.06260.

Rawls, J. 2009. A theory of justice. Harvard university press.

Shalit, U.; Johansson, F. D.; and Sontag, D. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In Proceedings of the 34th International Conference on Machine Learning-Volume 70, 3076–3085. JMLR.org.

Sohn, K.; Lee, H.; and Yan, X. 2015. Learning structured output representation using deep conditional generative models. In Advances in neural information processing systems, 3483–3491.

Wu, Y.; Zhang, L.; and Wu, X. 2019. Counterfactual fairness: Unidentification, bound and algorithm. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI, 10–16.

Xu, D.; Wu, Y.; Yuan, S.; Zhang, L.; and Wu, X. 2019. Achieving causal fairness through generative adversarial networks. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.

Zhang, L.; Wu, Y.; and Wu, X. 2018. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. IEEE Transactions on Knowledge and Data Engineering 31(11): 2035–2050.