DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning

ABSTRACT

Randomized controlled trials (RCTs) have been the gold-standard, in biomedical and social sciences among others, to assess causal effects of interventions on outcomes, e.g. medical treatments or lifestyle changes, and to reduce the risk of biased estimations. However, conducting RCTs is not always feasible due to operational or ethical constraints. Alternatively, causal effects can be drawn from observational, real-world data, but the data generation and collection process might not be randomized and contain underlying bias. Several techniques for addressing bias in treatment assignments, and for predicting individualized treatment effects (ITEs) –including counterfactuals, i.e., outcomes for alternative treatment scenarios– have been proposed, from propensity score matching, to ensemble tree-based learning, to recent breakthrough in deep learning, e.g., the Causal Effect Variational Autoencoder (CEVAE) or the Generative Adversarial Nets for inference of Individualized Treatment Effects (GANITE). In this work, we propose a novel deep learning approach, the Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning (DR-VIDAL) that incorporates the following key properties: (1) propensity weighting without loss of sample size; (2) latent-space design under causal independence assumptions; (3) feature attention on high-dimensional datasets and noise-reduction; (3) generation of intervention instances in addition to counterfactuals; and (4) doubly-robust causal effect estimation. Tests performed on real-world datasets showed that the DR-VIDAL outperforms several other state-of-the-art techniques. The utility of DR-VIDAL is not only with respect to prediction, for which the doubly robustness is assured, but also for more general inference tasks. The code is available under the MIT license on Github at: https://github.com/Shantanu48114860/Doubly_Robust_ITE

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

causal AI, biomedical informatics, generative adversarial networks, variational inference, information theory, doubly robust

ACM Reference Format:

. 2018. DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

Woodstock '18, June 03–05, 2018, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Evaluating causal effect of treatment assignment is one of the elementary problems in many research domains such as medicine, economics, policy making. However it is quite challenging to estimate the individualized treatment effect from the observational dataset as we do not have the access to the counterfactuals which deals with the question "what will be the outcome if the treatment will be intervened and assigned a different value?". This is why causal inference is different from traditional supervised learning problem as we can only observe the factual outcome for the treatment assignment, the counterfactual will not be observed at all. Also it is quite difficult to distinguish the conditional association from causation. However such distinction is extremely necessary in order to design the models to ascertain interventions which is necessary to estimate the causal effect of a treatment. A typical causal inference scenario in medicine is to determine whether a treatment (e.g., statins –a class of lipid-lowering medications) is effective with respect to an health outcome (e.g., reduce the risk of cardiovascular disease). Randomized control trial(RCT) is often considered to be the best practice in order to evaluate ITE [31]. In RCT, the treatment is assigned randomly to the samples so that the treatment assignment is free from any kind of selection bias making it independent of the individual's characteristics. Thus it ensures that the effect of the treatment on the outcome will entirely depends on the treatment and no other factors. However, RCT can be performed frequently due to ethical and legal constraints, e.g to evaluate whether college education is the cause of good salary, it is not ethical to randomly pick a teenager and force him/her not to get admitted to college. So, to estimate the effect of treatment on the outcome, observational data is the only source.

However, the observational dataset is often plagued with various biases - such as confounding (i.e., missing the true cause of an outcome but including spurious features correlated with the cause) and colliders (i.e., mistakenly including effects of an outcome as predictors), and other more complex configurations such as M-shaped or butterfly bias [11] - making it difficult to infer causal claims [26].

Traditional work in this domain to infer the causal effect of treatment from the observational data has been performed using propensity scores [10] and [23]. (which is the probability of receiving treatment given the covariates). The propensity score can be estimated using regression technique –often, regularized logistic regression is used [32]. Recently machine learning approaches to improve PSM have tackled problems such as calculating nonlinear propensity scores, and further exploited the *potential outcomes* statistical framework [27, 28], upon which algorithms are able to represent both factual and counterfactual outcomes and thus be

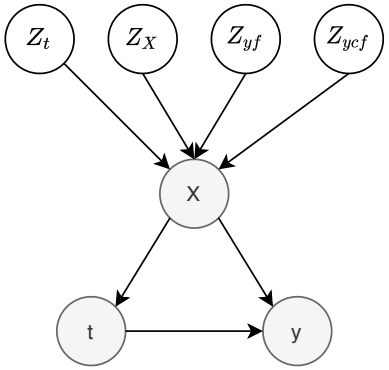

Figure 1: Directed Acyclic Graphs (DAGs) for estimating the causal effect of the treatment t on the outcome y with pre-treatment covariates X (hidden confounders).

used to calculate causal treatment effects [17], both at the individual level (i.e., individual treatment effect [ITE]) and at the population (average) level (i.e., average treatment effect [ATE]). Nonlinear counterfactual prediction models have used various learning algorithms such as Bayesian additive regression trees [15] and random forests [33]. The increasing availability of large amounts of electronic health record (EHR) data and advances in machine learning entailed the flourishing of a glut of deep learning models for causal inference, from which a number them are closely related to PSM/IPW such as the Treatment-Agnostic Representation Network (TARNet) [29], Dragonnet [30], GANITE [34] and CEVAE [21].

Contribution. In this work, we want to infer the causal effect of the treatment on the outcome by using latent variable models and doubly robust regression. Our idea is based on the figure 1 where the observed covariates x has a confounding effect on both the treatment assignment and the observed outcome. There are four independent latent variables $Z_t, Z_X, Z_{yf}, Z_{ycf}$, generating the covariates x. In this context we propose a novel deep learning approach DR-VIDAL - Doubly Robust Variational Information-theoretic Disentangled Adversarial Learning to disentangle the latent variables to estimate the ITE by generating the counterfactuals. Our major contributions are as follows:

- We estimate the counterfactual outcomes by inferring the latent variables that generates the covariates x first, by using a Variational Autoencoder (VAE) [19] and disentangling them. Next with the help of adversarial learning using a Generative Adversarial Network (GAN) [14] and variational information maximization [7], we generated the counterfactual outcomes so that the observational dataset complete will have the quadruples - covariates, treatment, factual/observed and counterfactual outcomes.
- After having the complete dataset, we next train a downstream doubly robust [12, 13] neural network, we estimated the ITE more accurately.
- We performed experiments on three real world datasets and our results have shown that we outperformed the existing state-of-art models.

In fact, for the first time in Causal inference literature, we amalgamated VAE, GAN, information theory and doubly robustness together to make our model more robust.

2 RELATED WORK

To estimate ITE, the classical works [3, 8–10, 22, 23, 25, 33] in focused on contemplating treatment as feature, with model learned for everything and the inconsistencies between the distribution of treated and controlled samples are adjusted to solve the problem of selection bias. For example tree-based methods have been used in [3, 8, 22, 33], doubly robust method has been deployed in [25] and propensity score based matching method has been used by [9, 10, 23]. With the recent popularity in deep learning, multitask networks have been utilized in TARNET and DCN-PD [1, 29] respectively to estimate the ITE. In TARNET [29], each sample from the treated and control group is associated with a weight indemnifying to the imbalance between the two groups. In DCN-PD [1], propensity score of each sample has been calculated to estimate the dropout probability for each sample to regularize the multitask network in alternating phases based on treated and control batches. Later, a modified architecture of TARNET - named as Dragonnet [30] has been developed to introduce the targeted regularization where the propensity score has been used. Along with this popular deep generative models like VAE [19] and GAN [14] have been employed by CEVAE [21, 34] to identify the latent variables from an observed proxy and generate counterfactuals respectively. We are motivated mostly from GANITE[34], CEVAE[21], Dragonnet[30] and TARNET [29], however the differences of the architectures and the training mechanisms of DR-VIDAL with these models are discussed in details in the following sections.

3 PROBLEM FORMULATION

We utilize the aforementioned potential outcomes framework [27, 28]. Let us consider a population sample of N individuals who can be prescribed a treatment T (binary, for simplicity), have a set of pre-treatment background covariates $\boldsymbol{X}$, and have a measured health outcome Y. We denote each subject i as the tuple $\{\boldsymbol{X}, T, Y\}_{i=1}^{N}$. For each individual i the potential outcomes are represented as Y_i^0 and Y_i^1 when applying treatments $T_i = 0$ and $T_i = 1$, respectively. The individualized treatment effect (ITE) $\tau(\boldsymbol{x})$ for an individual i with feature vector $\boldsymbol{X}_i = \boldsymbol{x}$, is defined as the difference in the mean potential outcomes under both treatment interventions (i.e., treated vs. not treated), conditional on the observed covariate vector $\boldsymbol{x}$

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_i^1 - Y_i^0 \mid \boldsymbol{X}_i = \boldsymbol{x}] \tag{1}$$

We can not calculate $\tau(\boldsymbol{x})$ directly, since we do not have the access to both potential outcomes, because an individual cannot be on and off treatment at the same time. Only one outcome (factual) can be observed, while the other (counterfactual) is missing.

If the potential outcomes are independent of the treatment assignment, conditionally on the background variables, i.e. $\{Y^1, Y^0\} \perp T \mid \boldsymbol{X}$, the assumption of strongly ignorable treatment assignment (SITA) is met [16, 24]. Under the assumption of SITA, the ITE can then be calculated as $\tau(\boldsymbol{x}) = \mathbb{E}[Y^1 \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y^0 \mid T = 0, \boldsymbol{X} = \boldsymbol{x}] = \mathbb{E}[Y \mid T = 1, \boldsymbol{X} = \boldsymbol{x}] - \mathbb{E}[Y \mid T = 0, \boldsymbol{X} = \boldsymbol{x}]$. Further, under SITA and by averaging over the distribution of $\boldsymbol{X}$,

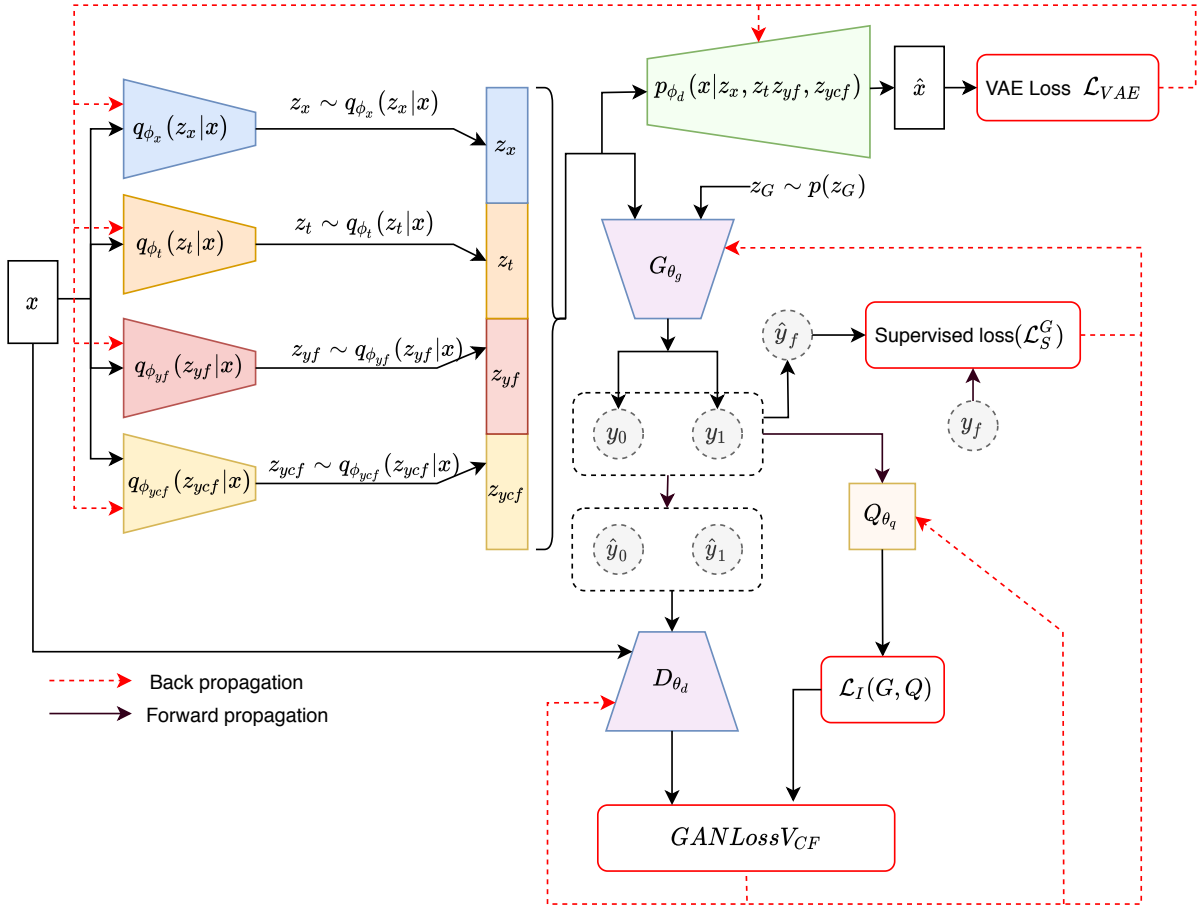

Figure 2: Architecture of the counterfactual network to estimate the counterfactual outcome.

the average treatment effect (ATE) τ_{01} can be calculated as

$$\tau_{01} = \mathbb{E}[\tau(\boldsymbol{X})] = \mathbb{E}[Y \mid T = 1] - \mathbb{E}[Y \mid T = 0] \tag{2}$$

ITE and ATE can be calculated with $\boldsymbol{x}$ being equally matched in treatment/control groups, but stratification becomes unfeasible as the dimension of $\boldsymbol{x}$ increases. With the help of PSM/IPW, through the conditional probability $\pi(\boldsymbol{x})$ one can balance the probability of receiving T given $\boldsymbol{X} = \boldsymbol{x}$ across the two comparison groups.

4 METHODOLOGY

In this section we describe the detailed description of the proposed model DR-VIDAL. It is mainly decomposed into 3 parts - 1) a VAE to disentanglement of the latent variables, 2) a GAN to generate the counterfactual outcomes 3) a doubly robust model to estimate the ITE. The architecture and the training algorithm of the VAE and GAN together is listed in figure 2 and algorithm 1 respectively. For the doubly robust estimator, the architecture and the training algorithm are mentioned in figure 3 and algorithm2 respectively. The detailed derivation of the ELBO loss of the VAE and the Information theoretic loss of GAN is derived in the Appendix A.1 and A.2 respectively.

4.1 Disentanglement of the latent variables using VAE

Our idea of decomposing the observed covariates is motivated from the figure 1. We assumed that the observed covariates x are generated from the independent latent instruments $z_x \sim p(z_x)$, $z_t \sim p(z_t)$, $z_{yf} \sim p(z_{yf})$, $z_{ycf} \sim p(z_{ycf})$ denoting the latent variables for the covariates x, treatment indicator t, factual outcome y_f and y_{cf} respectively. The goal is to infer the posterior distribution $p(z_x, z_t, z_{yf}, z_{ycf}|x)$, which is harder to optimize. So, we used the theory of variational inference [5] using a variational autoencoder [19] to learn the variational posteriors $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ using 4 different encoders which are neural networks with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively. Using the latent factors sampled from the learned variational posterior distributions, we reconstructed x by estimating the likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$ usinf a single decoder parametrized by ϕ_d. The latent factors assumed to be Gaussian distributions are defined as follows:

$$p(z_x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(z_{x_i}|0,1); \qquad p(z_t) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(z_{t_i}|0,1) \tag{3}$$

$$p(z_{yf}) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(z_{yf_i}|0,1); \quad p(z_{ycf}) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(z_{ycf_i}|0,1) \tag{4}$$

where $D_{z_x}, D_{z_t}, D_{z_{yf}}, D_{z_{ycf}}$ are the dimensions of the latent factors $z_x, z_t, z_{yf}, z_{ycf}$ respectively. The variational posteriors of the inference of models are defined as:

$$q_{\phi_x}(z_x|x) = \prod_{i=1}^{D_{z_x}} \mathcal{N}(\mu = \hat{\mu}_x, \sigma^2 = \hat{\sigma}_x^2) \tag{5}$$

$$q_{\phi_t}(z_t|x) = \prod_{i=1}^{D_{z_t}} \mathcal{N}(\mu = \hat{\mu}_t, \sigma^2 = \hat{\sigma}_t^2) \tag{6}$$

$$q_{\phi_{yf}}(z_{yf}|x) = \prod_{i=1}^{D_{z_{yf}}} \mathcal{N}(\mu = \hat{\mu}_{yf}, \sigma^2 = \hat{\sigma}_{yf}^2) \tag{7}$$

$$q_{\phi_{ycf}}(z_{ycf}|x) = \prod_{i=1}^{D_{z_{ycf}}} \mathcal{N}(\mu = \hat{\mu}_{ycf}, \sigma^2 = \hat{\sigma}_{ycf}^2) \tag{8}$$

where $\hat{\mu}_x, \hat{\mu}_t, \hat{\mu}_{yf}, \hat{\mu}_{ycf}$ and $\hat{\sigma}_x^2, \hat{\sigma}_t^2, \hat{\sigma}_{yf}^2, \hat{\sigma}_{ycf}^2$ are the means and variances of the Gaussian distributions parametrized by encoders $E_{\phi_x}, E_{\phi_t}, E_{\phi_{yf}}, E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively.

The overall ELBO loss of the VAE will be expressed as $\mathcal{L}_{ELBO}$ in the following equation,

$$\begin{aligned}
&\mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) \\
&= \mathbb{E}_{q_{\phi_x}, q_{\phi_t}, q_{\phi_{yf}}, q_{\phi_{ycf}}} [\log p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})] \\
&\quad - KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x))\big) \\
&\quad - KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t))\big) \\
&\quad - KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf}))\big) \\
&\quad - KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf}))\big)
\end{aligned}$$

We will minimise the optimization function of the VAE as $\mathcal{L}_{VAE}$ to obtain the optimal parameter of the encoders: $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ and decoder: ϕ_d respectively.

$$\begin{aligned}
&\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf}) = \\
&- \mathcal{L}_{ELBO}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})
\end{aligned} \tag{9}$$

4.2 Generation of counterfactual outcomes using GAN

After learning the hidden latent codes $z_x, z_t, z_{yf}, z_{ycf}$ from the VAE, we concatenated the latent codes to form z_c. z_c, along with a random noise $z_G \sim \mathcal{N}(0, Id)$ have been passed to the generator of the GAN block - denoted as G_{θ_g} which is parametrized by θ_g and outputs a vector of the potential outcomes (both factual and counterfactual) $\overline{y}$. We then replace the true factual outcome y_f in the generated outcome vector $\overline{y}$ to form $\hat{y}_0$ and $\hat{y}_1$. $\hat{y}_0$ and $\hat{y}_1$, along with the true covariate vector x will be passed to the counterfactual discriminator D_{θ_d}, parametrized by θ_d, which is then responsible

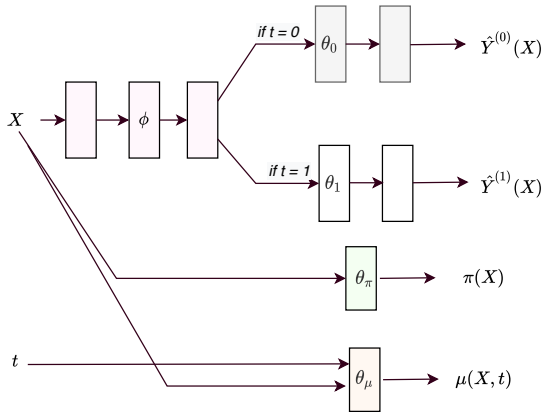

Figure 3: Architecture of the doubly robust network to calculate the average treatment effect.

to predict the treatment variable as proposed in GANITE [34]. The loss of the GAN block is given as:

$$\begin{aligned}
V_{GAN}(G, D) = \mathbb{E}_{x, z_G, z_c} \big[& t^T \log D(x, G(z_G, z_c)) \\
& + (1-t)^T \log(1 - D(x, G(z_G, z_c))) \big]
\end{aligned}$$

where $x \sim p(x), z_G \sim p(z_G)$ and z_c denotes the concatenated latent codes $z_x \sim q_{\phi_x}(z_x|x)$, $z_t \sim q_{\phi_t}(z_t|x)$, $z_{yf} \sim q_{\phi_{yf}}(z_{yf}|x)$ and $z_{ycf} \sim q_{\phi_{ycf}}(z_{ycf}|x)$.

From the $\overline{y}$, we also calculated the predicted factual outcome $\hat{y}_f$. As introduced in GANITE [34], we also include the supervised loss $\mathcal{L}_S^G(y_f, \hat{y}_f)$ which enforces the predicted factual outcome $\hat{y}_f$ to as close as to the true factual outcome y_f.

$$\mathcal{L}_S^G(y_f, \hat{y}_f) = \frac{1}{n} \sum_{i=1}^{n} \big(y_f(i) - \hat{y}_f(i)\big)^2 \tag{10}$$

The complete loss function of counterfactual GAN is given by

$$V_{CF}(G, D) = V_{GAN}(G, D) + \gamma \mathcal{L}_S^G(y_f, \hat{y}_f)$$

We also put additional regularization $\lambda I(z_c; G(z_G, z_c))$ to maximise the mutual information between the learnt concatenated latent code z_c and the generated output by the generator $G(z_G, z_c)$ as mentioned in [7].So, we propose to solve the following minimax game:

$$\min_G \max_D V_{CF\_I}(G, D) = V_{CF}(G, D) + \lambda I(z_c; G(z_G, z_c)) \tag{11}$$

As mentioned in [7], $I(z_c; G(z_G, z_c))$ is harder to solve because of the presence of the posterior $p(z_c|x)$, so we obtain the lower bound of it using an auxiliary distribution $Q(z_c|x)$ to approximate $p(z_c|x)$. Finally, the optimization function of the counterfactual information theoretic GAN incorporating the variational regularization of mutual information and a hyperparameter λ is given by:

$$\min_{G,Q} \max_D V_{CF\_infoGAN}(G, D, Q) = V_{CF}(G, D) - \lambda \mathcal{L}_I(G, Q) \tag{12}$$

The counterfactual InfoGAN is used to generate the missing counterfactual outcome y_{cf} to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^N$ and sent to the doubly robust block to estimate the ITE.

Algorithm 1 Training the Counterfactual Adversarial Generative Model to estimate counterfactual outcome

Input: Training set $\boldsymbol{X} = \{(x^{(1)}, t^{(1)}, y_f^{(1)}),..., (x^{(n)}, t^{(n)}, y_f^{(n)})\}$; hyper-parameters $\gamma > 0$; $\lambda > 0$; Encoders: $E_{\phi_x}, E_{\phi_t}, E_{\phi_{yf}}, E_{\phi_{ycf}}$ with parameters $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}$ respectively; Decoder D_{ϕ_d} with parameter D_{ϕ_d}; Generator G_{θ_g}, Discriminator D_{θ_d}, Q network D_{θ_q} with parameters $\theta_g, \theta_d, \theta_q$ respectively

Initialize parameters: $\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}, \phi_d, \theta_g, \theta_d, \theta_q$
while training **do**
  $x \leftarrow$ batch of samples from the dataset
  $z_{\mu_x}, z_{\sigma_x} \leftarrow E_{\phi_x}(x)$
  $z_{\mu_t}, z_{\sigma_t} \leftarrow E_{\phi_t}(x)$
  $z_{\mu_{yf}}, z_{\sigma_{yf}} \leftarrow E_{\phi_{yf}}(x)$
  $z_{\mu_{ycf}}, z_{\sigma_{ycf}} \leftarrow E_{\phi_{ycf}}(x)$
  $z_x \leftarrow z_{\mu_x} + \epsilon z_{\sigma_x}$, where $\epsilon \sim \mathcal{N}(0, Id)$
  $z_t \leftarrow z_{\mu_t} + \epsilon z_{\sigma_t}$, where $\epsilon \sim \mathcal{N}(0, Id)$
  $z_{yf} \leftarrow z_{\mu_{yf}} + \epsilon z_{\sigma_{yf}}$, where $\epsilon \sim \mathcal{N}(0, Id)$
  $z_{ycf} \leftarrow z_{\mu_{ycf}} + \epsilon z_{\sigma_{ycf}}$, where $\epsilon \sim \mathcal{N}(0, Id)$
  Concatenate $z_x, z_t, z_{yf}, z_{ycf}$ to form z_c
  $\hat{x} \leftarrow D_{\phi_d}(z_c)$
  Calculate $\mathcal{L}_{VAE}(\phi_x, \phi_t, \phi_{yf}, \phi_{ycf}; x, z_x, z_t, z_{yf}, z_{ycf})$
  $\phi_x \overleftarrow{-} \nabla_{\phi_x}\mathcal{L}_{VAE}$; $\phi_t \overleftarrow{-} \nabla_{\phi_t}\mathcal{L}_{VAE}$; $\phi_{yf} \overleftarrow{-} \nabla_{\phi_{yf}}\mathcal{L}_{VAE}$; $\phi_{ycf} \overleftarrow{-} \nabla_{\phi_{ycf}}\mathcal{L}_{VAE}$; $\phi_d \overleftarrow{-} \nabla_{\phi_d}\mathcal{L}_{VAE}$
  $z_G \sim \mathcal{N}(0, Id)$
  $y_0, y_1 \leftarrow G_{\theta_g}(z_G, z_c)$
  $\hat{y}_0 = ((1-T) * y_f + T * y_0)$
  $\hat{y}_1 = (T * y_f + (1-T) * y_1)$
  $d_{logit} \leftarrow D_{\theta_d}(x, \hat{y}_0, \hat{y}_1)$
  Calculate $\mathcal{L}^D(\theta_d)$
  $\theta_d \overleftarrow{-} \nabla_{\theta_d}\mathcal{L}^D(\theta_d)$
  $\hat{y}_f \leftarrow T * y_1 + (1-T) * y_0$
  Compute $\mathcal{L}_S^G(y_f, \hat{y}_f)$
  Concatenate y_0, y_1 to form q_{input}
  $q_\mu, q_\sigma \leftarrow Q_{\theta_q}(q_{input})$
  Compute $\mathcal{L}_I(G, Q)$ by treating $Q(c|x)$ as factored Gaussian using q_μ, q_σ and z_c
  Compute $\mathcal{L}^G(\theta_g)$
  $\theta_g \overleftarrow{-} \nabla_{\theta_g}\mathcal{L}^G(\theta_g)$
end while

GAN optimization

The objective of the discriminator D_{θ_d} and the generator G_{θ_g} of the GAN are to correctly and incorrectly identify the treatment indicator respectively - this is the adversarial learning employed in the training objective. To get the optimal Discriminator $D^*_{\theta_d}$, we need to maximise $V_{CF\_infoGAN}$ as

$$\max_D \mathcal{L}^D(\theta_d) = V_{CF\_infoGAN}(G, D, Q) \tag{13}$$

To get the optimal Generator $G^*_{\theta_g}$, we need to maximise $V_{CF\_infoGAN}$ as

$$\min_{G,Q} \mathcal{L}^G(\theta_g) = V_{CF\_infoGAN}(G, D, Q) \tag{14}$$

4.3 Estimation of ATE using Doubly Robust regression model

Algorithm 2 Training the Doubly Robust Multitask Model to estimate ITE

Input: Complete dataset $\tilde{X} = \{(x^{(1)}, t^{(1)}, y_f^{(1)}, y_{cf}^{(1)}),..., (x^{(n)}, t^{(n)}, y_f^{(n)}, y_{cf}^{(n)})\}$ after training the Counterfactual generative model; hyper-parameters $\alpha > 0$; $\beta > 0$; outcome heads with shared parameters ϕ and outcome specific parameters θ_0, θ_1; propensity head with parameters θ_π; regressor head with parameters θ_μ

Initialize parameters: $\theta_0, \theta_1, \theta_\pi, \theta_\mu$
while training **do**
  $x \leftarrow$ batch of samples from the dataset
  Calculate $\hat{y}_i^{(0)}, \hat{y}_i^{(1)}, \hat{y}_f^{(i)}, \hat{y}_{cf}^{(i)}$
  Calculate the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$
  Calculate $\hat{y}_{fDR}^{(i)}, \hat{y}_{cfDR}^{(i)}$
  Calculate the Doubly Robust loss $\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
  Calculate the final loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
  Calculate gradients of the loss $\mathcal{L}_{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi)$
  Update the parameters $\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi$
end while

The propensity score $\pi(x)$ represents the probability of receiving a treatment $T = 1$ [27] (assuming that the alternative is no treatment $T = 0$) conditioned on the pre-treatment covariates X, denoted as

$$\pi(\boldsymbol{x}) = P(T = 1|\boldsymbol{X} = \boldsymbol{x}). \tag{15}$$

Combining the strategy of Inverse Probability of Treatment weighting (IPW) [4] and regressing the outcome using treatment variable and the covariates, Jonsson introduced the Doubly Robust estimation of causal effect [13] as the following:

$$\hat{\delta}_{DR} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i t_i - (t_i - \pi(x_i))\mu(x_i, t_i)}{\pi(x_i)} - \frac{y_i(1-t_i) - (t_i - \pi(x_i))\mu(x_i, t_i)}{1-\pi(x_i)}\right] \quad \text{where,} \tag{16}$$

$$\mu(x, t) = \hat{\alpha_0} + \hat{\alpha_1}x_1 + \hat{\alpha_2}x_2 + \cdots + \hat{\alpha_n}x_n + \hat{\delta}t \tag{17}$$

The term in red is used to augment the IPW estimator. After getting the counterfactual outcome y_{cf} from the Counterfactual Generative algorithm to form the quadruple $\{X, t, y_f, y_{cf}\}_{i=1}^N$ we passed this as the input to the Doubly Robust Multitask Network to estimate the Individual Treatment Effect (ITE) shown in figure 3. To predict the outcomes $y^{(0)}$ and $y^{(1)}$, we used the model, motivated by TARNET [29] which contained some shared layers denoted by f_ϕ parametrized by ϕ and two outcome specific heads denoted by f_{θ_0} and f_{θ_1} parametrized by θ_0 and θ_1 to predict $y^{(0)}$ and $y^{(1)}$ respectively. To make the model Doubly Robust, we also introduced

Methods	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS/LR1	5.8 ± 0.3	0.73 ± 0.04	5.8 ± 0.3	0.94 ± 0.06
OLS/LR2	2.4 ± 0.1	0.14 ± 0.01	2.5 ± 0.1	0.31 ± 0.02
BLR	5.8 ± 0.3	0.72 ± 0.04	5.8 ± 0.3	0.93 ± 0.05
k-NN	2.1 ± 0.1	0.14 ± 0.01	4.1 ± 0.2	0.79 ± 0.05
BART	2.1 ± 0.2	0.23 ± 0.01	2.3 ± 0.1	0.34 ± 0.02
R Forest	4.2 ± 0.2	0.73 ± 0.05	6.6 ± 0.3	0.96 ± 0.06
C Forest	3.8 ± 0.2	0.18 ± 0.01	3.8 ± 0.2	0.40 ± 0.03
BNN	2.2 ± 0.1	0.37 ± 0.03	2.1 ± 0.1	0.42 ± 0.03
TARNET	0.88 ± 0.02	0.26 ± 0.01	0.95 ± 0.02	0.28 ± 0.01
CFR_{WASS}	0.71 ± 0.0	0.25 ± 0.01	0.76 ± 0.0	0.27 ± 0.01
GANITE	1.9 ± 0.4	0.43 ± 0.05	2.4 ± 0.4	0.49 ± 0.05
CEVAE	2.7 ± 0.1	0.34 ± 0.01	2.6 ± 0.1	0.46 ± 0.02
DR-VIDAL	**1.44 ± 0.03**	**0.44 ± 0.11**	**1.46 ± 0.12**	**0.44 ± 0.19**

Table 1: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the IHDP dataset.

Methods	$R_{Pol}^{within-s}$	$\epsilon_{ATT}^{within-s}$	$R_{Pol}^{out-of-s}$	$\epsilon_{ATT}^{out-of-s}$
OLS/LR1	0.22 ± 0.0	0.01 ± 0.00	0.23 ± 0.0	0.08 ± 0.04
OLS/LR2	0.21 ± 0.0	0.01 ± 0.01	0.24 ± 0.0	0.08 ± 0.03
BLR	0.22 ± 0.0	0.01 ± 0.01	0.25 ± 0.0	0.08 ± 0.03
k-NN	0.02 ± 0.0	0.21 ± 0.01	0.26 ± 0.0	0.13 ± 0.05
BART	0.23 ± 0.0	0.02 ± 0.00	0.25 ± 0.0	0.08 ± 0.03
R Forest	0.23 ± 0.0	0.03 ± 0.01	0.28 ± 0.0	0.09 ± 0.04
C Forest	0.19 ± 0.0	0.03 ± 0.01	0.20 ± 0.0	0.07 ± 0.03
BNN	0.20 ± 0.0	0.03 ± 0.01	0.24 ± 0.0	0.09 ± 0.04
TARNET	0.17 ± 0.0	0.05 ± 0.02	0.21 ± 0.0	0.11 ± 0.04
CFR_{WASS}	0.17 ± 0.0	0.04 ± 0.01	0.21 ± 0.0	0.09 ± 0.03
GANITE	0.13 ± 0.01	0.01 ± 0.01	0.14 ± 0.01	0.06 ± 0.03
CEVAE	0.15 ± 0.0	0.02 ± 0.01	0.26 ± 0.0	0.03 ± 0.01
DR-VIDAL	**0.09 ± 0.005**	**0.04 ± 0.03**	**0.10 ± 0.01**	**0.05 ± 0.02**

Table 2: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the Jobs dataset.

two more heads to predict the propensity score $\pi(X) = \mathbb{P}(t = 1|X)$ and the regressor $\mu(X, t)$. These two were calculated using two neural networks parametrized by θ_π and θ_μ respectively. The factual and counterfactual outcome were calculated from the predicted potential outcomes (from the doubly robust network) $y_i^{(0)}$ and $y_i^{(1)}$ of the i^{th} sample as:

$$\hat{y}_i^{(0)} = f_{\theta_0}(f_\phi(x_i)) \qquad \text{if } t_i = 0 \tag{18}$$

$$\hat{y}_i^{(1)} = f_{\theta_1}(f_\phi(x_i)) \qquad \text{if } t_i = 1 \tag{19}$$

$$\hat{y}_f^{(i)} = t_i\hat{y}_i^{(1)} + (1 - t_i)\hat{y}_i^{(0)} \tag{20}$$

$$\hat{y}_{cf}^{(i)} = (1 - t_i)\hat{y}_i^{(1)} + t_i\hat{y}_i^{(0)} \tag{21}$$

Next, the predicted loss $\mathcal{L}_i^p(\theta_1, \theta_0, \phi)$ will be calculated as:

$$\mathcal{L}_i^p(\theta_1, \theta_0, \phi) = (\hat{y}_f^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf}^{(i)} - y_{cf}^{(i)})^2 + \alpha\text{BinaryCrossEntropy}(\pi(x_i), t_i) \tag{22}$$

where α is a hyperparameter. With the help of the propensity score $\pi(X)$ and the regressor $\mu(X, t)$, the Doubly Robust outcomes were calculated as

$$\hat{y}_{f_{DR}}^{(i)} = t_i\left[\frac{t_i\hat{y}_i^{(1)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{\pi(x_i)}\right] + (1 - t_i)\left[\frac{(1 - t_i)\hat{y}_i^{(0)} - (t_i - \pi(X_i)\mu(x_i, t_i))}{1 - \pi(x_i)}\right] \tag{23}$$

$$\hat{y}_{cf_{DR}}^{(i)} = (1 - t_i)\left[\frac{(1 - t_i)\hat{y}_i^{(1)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{\pi(x_i)}\right] + t_i\left[\frac{t_i\hat{y}_i^{(0)} - (t_i - \pi(x_i)\mu(x_i, t_i))}{1 - \pi(x_i)}\right] \tag{24}$$

Finally the Doubly Robust loss $\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\phi, \theta_\mu, \phi)$ was calculated as:

$$\mathcal{L}_i^{DR}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = (\hat{y}_{f_{DR}}^{(i)} - y_f^{(i)})^2 + (\hat{y}_{cf_{DR}}^{(i)} - y_{cf}^{(i)})^2 \tag{25}$$

The final loss function of the Doubly robust will be denoted as:

$$\mathcal{L}^{ITE}(\theta_1, \theta_0, \theta_\pi, \theta_\mu, \phi) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathcal{L}_i^p + \beta\mathcal{L}_i^{DR}\right) \tag{26}$$

where β is a hyperparameter and the whole network is trained using end-to-end strategy .

Comparison with CEVAE and GANITE

Like CEVAE[21] and GANITE [34], the counterfactual outcome predictor block of DR-VIDAL also uses deep generative models with some significant differences. CEVAE uses a VAE to estimate the ITE, employing a causal graph where the latent variable Z and the model aims to infer the observed proxy X from Z. On the other hand, DR-VIDAL is highly motivated by the GANITE architecture with the same adversarial training objective to estimate the counterfactual outcome. However, unlike GANITE and CEVAE, we consider a different causal graph with multiple latent instruments responsible for the generating the common observed confounder - so, first we disentangle the multiple latent factors using a VAE, employed the adversarial learning of GANITE along with the mutual information maximization objective to generate the counterfactual outcomes to estimate the entire potential outcome vector and then evaluate the ITE using doubly robust network.

Comparison with TARNET and Dragonnet

The design doubly robust model of DR-VIDAL is closely related to that of TARNET [29] and Dragonnet [30]. The doubly robust network without the propensity score and the regressor heads is the TARNET and only the he propensity score head is essentially the Dragonnet. However the in TARNET, the weights corresponding to each of samples is calculated as the probability of the treatment

Methods	$\sqrt{\epsilon_{PEHE}^{within-s}}$	$\epsilon_{ATE}^{within-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
OLS/LR1	0.319 ± 0.005	0.0038 ± 0.0025	0.297 ± 0.016	0.0069 ± 0.0056
OLS/LR2	0.320 ± 0.001	0.0039 ± 0.0025	0.318 ± 0.007	0.0070 ± 0.0059
BLR	0.312 ± 0.002	0.0057 ± 0.0036	0.320 ± 0.003	0.0334 ± 0.0092
k-NN	0.333 ± 0.003	0.0028 ± 0.0021	0.323 ± 0.018	0.0051 ± 0.0039
BART	0.347 ± 0.009	0.1206 ± 0.0236	0.338 ± 0.016	0.1265 ± 0.0234
R Forest	0.306 ± 0.002	0.0049 ± 0.0034	0.321 ± 0.005	0.0080 ± 0.0051
C Forest	0.366 ± 0.003	0.0286 ± 0.0035	0.316 ± 0.011	0.0335 ± 0.0083
BNN	0.325 ± 0.003	0.0056 ± 0.0032	0.321 ± 0.018	0.0203 ± 0.0071
TARNET	0.317 ± 0.005	0.0108 ± 0.0017	0.315 ± 0.003	0.0151 ± 0.0018
CFR_{WASS}	0.315 ± 0.007	0.0112 ± 0.0016	0.313 ± 0.008	0.0284 ± 0.0032
GANITE	0.289 ± 0.12	0.0058 ± 0.0017	0.297 ± 0.05	0.0089 ± 0.0075
CEVAE	n.r	n.r	n.r	n.r
DR-VIDAL	**0.317 ± 0.002**	**0.0102 ± 0.0128**	**0.318 ± 0.008**	**0.0111 ± 0.0137**

Table 3: Performance on the within-sample and out-of-sample test sets (mean ± st.dev) of various models on the Twins dataset.

assignment whereas in DR-VIDAL, the weights are calculated based on the propensity score which is the probability of treatment assignment given the covariates. For the Dragonnet, the targeted regularization was implemented without taking the regressed out into account which is estimated using the treatment assignment and the covariates together as in DR-VIDAL. Another major difference between TARNET and Dragonnet with DR-VIDAL is the training strategy employed. For both TARNET and Dragonnet, the counterfactual outcome does not exist, so for each sample the overall loss function has been estimated with the factual outcome only and the parameters of the outcome head of the factual outcome have been updated during training. In DR-VIDAL, we have the entire potential outcome vector comprising of both the factual and the counterfactual outcomes, so for each training sample, the loss function is calculated for both the outcomes and the corresponding parameters of both the outcome heads are updated as shown in algorithm 2.

5 EXPERIMENTAL SETUP

5.1 Datasets

Estimating ITE from a observational dataset is dataset is a difficult task due to the unavailability of the counterfactuals. However, the importance of the dataset holds immense importance in estimating the ITE. In this work, we used two semi-synthetic datasets Infant Health and Development Program (IHDP)[15], Twins [2] and a real world dataset - Jobs [20]. These datasets are well described in [21, 29, 34]. In all the experiments for all the datasets, we used the same settings described in GANITE [34], where all the datasets are divided into 56/24/20 % train-validation-test splits. We ran 1000, 10 and 100 realizations of IHDP, Jobs and Twins datasets respectively and report the performance metrics on both the within and out-of-samples.

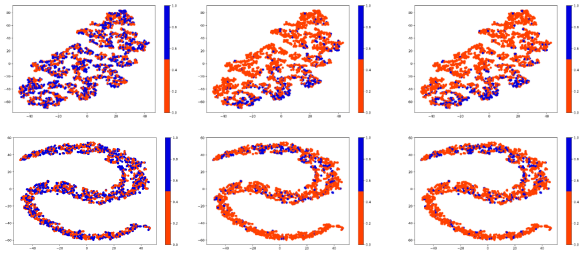

Figure 4: t-SNE visualization of the Latent representation of the Twins dataset, learned by the encoder of the counterfactual network. From the left to right, the plots show the representations of treatment, factual and counterfactual outcomes respectively. The top and bottom panels show the representations before and after training the network respectively.

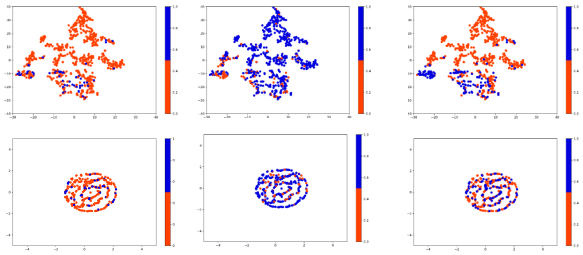

Figure 5: t-SNE visualization of the Latent representation of the Jobs dataset, learned by the encoder of the counterfactual network. From the left to right, the plots show the representations of treatment, factual and counterfactual outcomes respectively. The top and bottom panels show the representations before and after training the network respectively.

5.2 Deep learning module's settings

Adversarial Module To reduce the model complexity and parameters for the encoder of the VAE, we have a shared neural network connected to 4 other neural networks for estimating the 4 posterior distributions $q_{\phi_x}(z_x|x), q_{\phi_t}(z_t|x), q_{\phi_{yf}}(z_{yf}|x), q_{\phi_{ycf}}(z_{ycf}|x)$. The shared neural network has 3 layers, each having 15 nodes. The networks with outputs as $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ have a single layer with 5, 1, 1, 1 nodes respectively. The decoder is 4-layered neural network, each with 15 nodes to calculate the data likelihood $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{ycf})$. For the GAN, the generator network has 2 shared layers and 2 outcome specific layers for each outcomes, each with 100 nodes. The discriminator and Qnetwork is a 3-layered neural network, each with 30 nodes and 8 nodes respectively. All the layers of the VAE and GAN used RELU activation functions and the parameters were updated using Adam optimizer [18]. The random noise z_G is sampled from a 92-dimensional $\mathcal{N}(0, 1)$.

The hyperparameters γ is set as 1 for the 3 datasets and λ is set as 0.2, 0.01 and 10 for IHDP, Jobs and Twins datasets respectively.

	IHDP		Jobs		Twins	
	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$	$R_{Pol}^{out-of-s}$	$\epsilon_{ATT}^{out-of-s}$	$\sqrt{\epsilon_{PEHE}^{out-of-s}}$	$\epsilon_{ATE}^{out-of-s}$
DR-VIDAL	**1.453 ± 0.11**	**0.45 ± 0.12**	**0.102 ± 0.01**	**0.056 ± 0.02**	**0.318 ± 0.008**	**0.0111 ± 0.0137**
DR-VIDAL (w/o DR loss)	1.476 ± 0.17	0.46 ± 0.18	0.110 ± 0.01	0.091 ± 0.04	0.324 ± 0.007	0.0131 ± 0.0152
DR-VIDAL (w/o Info loss)	1.461 ± 0.11	0.45 ± 0.12	0.109 ± 0.01	0.056 ± 0.02	0.318 ± 0.012	0.0115 ± 0.0171
DR-VIDAL (w/o DR loss & Info loss)	1.476 ± 0.13	0.46 ± 0.15	0.113 ± 0.01	0.094 ± 0.04	0.326 ± 0.008	0.0124 ± 0.0172

Table 4: Performance on the out-of-sample test sets (mean ± st.dev) of the all the variants of DR-VIDAL algorithm on 100, 10, 100 realizations of the IHDP, Jobs and Twins datasets respectively.

The batch sizes of IHDP, Jobs and Twins are set as 64, 64, 256 for the for IHDP, Jobs and Twins datasets respectively. The learning rate of the VAE, generator and discriminator is set as 1e-3, 1e-4, 5e-4 respectively.

Doubly robust Module For the Doubly robust module, the shared network f_ϕ and outcome specific networks f_{θ_0} and f_{θ_1} are both 3-layered neural network each with 200 and 100 nodes respectively. The propensity network π has 2 layers each with 200 nodes. The regressor network μ has 6 layers with 200 nodes and 100 nodes in the first and last 3 layers respectively. All the layers of the VAE and GAN used ELU activation functions and the parameters were updated using Adam optimizer [18]. The batch sizes of IHDP, Jobs and Twins are set as 64, 64, 256 for the for IHDP, Jobs and Twins datasets respectively. We set the learning rate of all the networks as 1e-4 and the hyperparameters α and β are both set as 1 for all 3 datasets.

5.3 Validation and performance metrics

We report the expected Precision in Estimation of Heterogeneous Effect (PEHE) (ϵ_{PEHE}), average treatment effect (ATE) (ϵ_{ATE}) as discussed in [15, 29, 34] for datasets IHDP and Twins due to the availability of the factual and the counterfactual outcomes both. For Jobs, the counterfactual outcome does not exist, so we report the policy risk ($R_{pol}(\pi)$) and the error in average treatment effect on the treated (ATT) (ϵ_{ATT}) as mentioned by [29, 34].

$$\epsilon_{PEHE} = \frac{1}{N}\sum_{n=0}^{N}\Big(\mathbb{E}_{y_j(n)\sim\mu_j(n)}\big[y_1(n) - y_0(n)\big] - \big[\hat{y_1}(n) - \hat{y_0}(n)\big]\Big)^2 \tag{27}$$

$$\epsilon_{ATE} = ||\frac{1}{N}\sum_{n=0}^{N}\mathbb{E}_{y(n)\sim\mu(n)}[y(n)] - \frac{1}{N}\sum_{n=0}^{N}\hat{y}(n)||_2^2 \tag{28}$$

$$R_{pol}(\pi) = \frac{1}{N}\sum_{n=0}^{N}\Big[1 - \Big(\sum_{i=1}^{k}[\frac{1}{|\Pi_i \cap T_i \cap E|} \sum_{x(n)\in\Pi_i\cap\ T_i\cap E} y_i(n) \times \frac{|\Pi_n \cap E|}{|E|}]\Big)\Big] \tag{29}$$

where $\pi_i = \{\boldsymbol{x(n)} : i = \arg\max \hat{\boldsymbol{y}}$,
$T_i = \boldsymbol{x(n)} : t_i(n) = 1\}$, and E is the randomized sample. For datasets where only factual outcomes are available with treatment being binary, such as Jobs and a randomized controlled trail(RCT) generates the testing set, the true average treatment effect on the treated (ATT) and the error ϵ_{ATT} are define by [29] as follows:

$$ATT = \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E} Y_1(x_i) - \frac{1}{|T_0 \cap E|}\sum_{x_i\in C\cap E} Y_0(x_i) \tag{30}$$

$$\epsilon_{ATT} = \Big|ATT - \frac{1}{|T_1 \cap E|}\sum_{x_i\in T_1\cap E}\hat{Y}_1(x_i) - \hat{Y}_0(x_i)\Big| \tag{31}$$

where T_1, T_0 and E are the subsets corresponding to treated, controlled samples and randomized controlled trials respectively.

6 RESULTS

We evaluated DR-VIDAL for 3 datasets and compare the performance with least squares regression using treatment as a feature (OLS/LR1), separate least squares regressions for each treatment (OLS/LR2), balancing linear regression (BLR) [17], k-nearest neighbor (k-NN) [9], Bayesian additive regression trees (BART) [8], random forests (RForest) [6], causal forests (C Forest) [33], balancing neural network (BNN) [17], treatment-agnostic representation network (TARNET) [29], counterfactual regression with Wasserstein distance (CFRW ASS) [29], CEVAE [21], and GANITE [34]. We report the performance for both the in-sample and out-of-sample in Table 1, 2 and 3 for IHDP, Jobs and Twins respectively. We can see that DR-VIDAL outperformed all the other models for Jobs dataset. For IHDP, it also surpassed all the models by a significant margin except TARNET and CFR_{WASS}. Due to the disentanglement of the hidden factors and the adversarial learning, the performance gain for the IHDP dataset was consequential compared to the other 2 generative models CEVAE and GANITE even with the large number of parameters of DR-VIDAL for a small dataset like IHDP. For Twins, the performance of DR-VIDAL is fairly competitive with most of the state of the art methods. The representations learned by the VAE of the adversarial model of DR-VIDAL for Twins and Jobs dataset is shown in Figures 4 and 5 respectively for both before and after training. From these figures, we can observe how the learned representations are clustered within a small space after we train the model. Also Table 4 shows the different variations of DR-VIDAL and how DR-VIDAL with the doubly robust loss and information loss performs the best for all the three datasets

7 DISCUSSION

In this work, DR-VIDAL - a novel deep learning based model showed the power of adversarial representation learning with doubly robust

regression. However, there are some limitation of this work. One limitation is the causal graph that we assumed to work on is fairly simple. Also it will be quite, interesting to see how TARNET and Dragonnet will perform as a downstream model after the counterfactuals, generated from the adversarial network of Dr-VIDAL. Another possible extension will be the usage of attention in the encoded representations in VAE and while calculating the propensity score which will give us the more important covariates in the covariate space.

In conclusion, DR-VIDAL framework is a promising approach to estimate the counterfactuals and the ITE. Our experiments proved that this method outperforms the state-of-the-art models empirically making the estimation more robust.

REFERENCES

[1] Ahmed M Alaa, Michael Weisz, and Mihaela Van Der Schaar. 2017. Deep counterfactual networks with propensity-dropout. *arXiv preprint arXiv:1706.05966* (2017).

[2] Douglas Almond, Kenneth Y Chay, and David S Lee. 2005. The costs of low birth weight. *The Quarterly Journal of Economics* 120, 3 (2005), 1031–1083.

[3] Susan Athey and Guido Imbens. 2016. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* 113, 27 (2016), 7353–7360.

[4] Peter C Austin and Elizabeth A Stuart. 2017. The performance of inverse probability of treatment weighting and full matching on the propensity score in the presence of model misspecification when estimating the effect of treatment on survival outcomes. *Statistical methods in medical research* 26, 4 (2017), 1654–1670.

[5] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association* 112, 518 (2017), 859–877.

[6] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. 2016. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems* 29 (2016), 2172–2180.

[8] Hugh A Chipman, Edward I George, Robert E McCulloch, et al. 2010. BART: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298.

[9] Richard K Crump, V Joseph Hotz, Guido W Imbens, and Oscar A Mitnik. 2008. Nonparametric tests for treatment effect heterogeneity. *The Review of Economics and Statistics* 90, 3 (2008), 389–405.

[10] Rajeev H Dehejia and Sadek Wahba. 2002. Propensity score-matching methods for nonexperimental causal studies. *Review of Economics and statistics* 84, 1 (2002), 151–161.

[11] Peng Ding and Luke W. Miratrix. 2015. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference* 3, 1 (2015), 41 – 57. https://doi.org/10.1515/jci-2013-0021

[12] Miroslav Dudík, Dumitru Erhan, John Langford, Lihong Li, et al. 2014. Doubly robust policy evaluation and optimization. *Statist. Sci.* 29, 4 (2014), 485–511.

[13] Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M Alan Brookhart, and Marie Davidian. 2011. Doubly robust estimation of causal effects. *American journal of epidemiology* 173, 7 (2011), 761–767.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014), 2672–2680.

[15] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

[16] Guido W Imbens. 2000. The role of the propensity score in estimating dose-response functions. *Biometrika* 87, 3 (2000), 706–710.

[17] Fredrik Johansson, Uri Shalit, and David Sontag. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*. 3020–3029.

[18] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[19] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[20] Robert J LaLonde. 1986. Evaluating the econometric evaluations of training programs with experimental data. *The American economic review* (1986), 604–620.

[21] Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. Causal effect inference with deep latent-variable models. In *Advances in neural information processing systems*. 6446–6456.

[22] Min Lu, Saad Sadiq, Daniel J Feaster, and Hemant Ishwaran. 2018. Estimating individual treatment effect in observational data using random forest methods. *Journal of Computational and Graphical Statistics* 27, 1 (2018), 209–219.

[23] Jared K Lunceford and Marie Davidian. 2004. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in medicine* 23, 19 (2004), 2937–2960.

[24] J. Pearl, M. Glymour, and N.P. Jewell. 2016. *Causal Inference in Statistics: A Primer*. Wiley. https://books.google.com/books?id=L3G-CgAAQBAJ

[25] Kristin E Porter, Susan Gruber, Mark J Van Der Laan, and Jasjeet S Sekhon. 2011. The relative performance of targeted maximum likelihood estimators. *The International Journal of Biostatistics* 7, 1 (2011).

[26] Mattia Prosperi, Yi Guo, Matt Sperrin, James S. Koopman, Jae S. Min, Xing He, Shannan Rich, Mo Wang, Iain E. Buchan, and Jiang Bian. 2020. Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nature Machine Intelligence* 2 (2020), 369–375. https://doi.org/10.1038/s42256-020-0197-y

[27] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (04 1983), 41–55. https://doi.org/10.1093/biomet/70.1.41

[28] Donald B Rubin. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* 66, 5 (1974), 688–701.

[29] Uri Shalit, Fredrik D. Johansson, and David Sontag. 2017. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 3076–3085. http://proceedings.mlr.press/v70/shalit17a.html

[30] Claudia Shi, David Blei, and Victor Veitch. 2019. Adapting neural networks for the estimation of treatment effects. In *Advances in neural information processing systems*. 2507–2517.

[31] Bonnie Sibbald and Martin Roland. 1998. Understanding controlled trials: Why are randomised controlled trials important? *BMJ* 316, 7126 (1998), 201. https://doi.org/10.1136/bmj.316.7126.201

[32] Yuxi Tian, Martijn J Schuemie, and Marc A Suchard. 2018. Evaluating large-scale propensity score performance through real-world and synthetic data experiments. *International journal of epidemiology* 47, 6 (2018), 2005–2014.

[33] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (2018), 1228–1242.

[34] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. 2018. GANITE: Estimation of Individualized Treatment Effects using Generative Adversarial Nets. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=ByKWUeWA-

A APPENDIX

A.1 Derivation of the Loss function ELBO loss of VAE

From figure 1, $p_{\phi_d}(x|z_x, z_t, z_{yf}, z_{y_cf})$ and $p_{\phi_d}(z_x, z_t, z_{yf}, z_{y_cf}|x)$ are the true likelihood and true posterior respectively. The posterior is hard to evaluate, so we have to approximate the true posterior to the product of the factorized known distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$ and $q_{\phi_{ycf}}(z_{ycf}|x)$ by minimising the KL divergence as follows,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)|| \\
&\quad p_{\phi_d}(z_x,z_t,z_{yf},z_{ycf}|x)\big) \\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x) \\
&\quad \Big[\log\frac{q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_x,z_t,z_{yf},z_{ycf}|x)}\Big]dz_x dz_t dz_{yf} dz_{ycf} \\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x) \\
&\quad \Big[\log q_{\phi_x}(z_x|x)+\log q_{\phi_t}(z_t|x) \\
&\quad +\log q_{\phi_{yf}}(z_{yf}|x)+\log q_{\phi_{ycf}}(z_{ycf}|x) \\
&\quad -\log p_{\phi_d}(z_x,z_t,z_{yf},z_{ycf}|x)\Big]dz_x dz_t dz_{yf} dz_{ycf} \\
&= \int\int\int\int q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x) \\
&\quad \Big[\log q_{\phi_x}(z_x|x)+\log q_{\phi_t}(z_t|x) \\
&\quad +\log q_{\phi_{yf}}(z_{yf}|x)+\log q_{\phi_{ycf}}(z_{ycf}|x) \\
&\quad -\log p_{\phi_d}(x|z_x,z_t,z_{yf},z_{ycf})-\log p_{\phi_d}(z_x,z_t,z_{yf},z_{ycf}) \\
&\quad +\log p_{\phi_d}(x)\Big]dz_x dz_t dz_{yf} dz_{ycf} \\
&= \int q_{\phi_x}(z_x|x)\log\frac{q_{\phi_x}(z_x|x)}{p_{\phi_d}(z_x)}dz_x \\
&\quad +\int q_{\phi_t}(z_t|x)\log\frac{q_{\phi_t}(z_t|x)}{p_{\phi_d}(z_t)}dz_t \\
&\quad +\int q_{\phi_{yf}}(z_{yf}|x)\log\frac{q_{\phi_{yf}}(z_{yf}|x)}{p_{\phi_d}(z_{yf})}dz_{yf} \\
&\quad +\int q_{\phi_{ycf}}(z_{ycf}|x)\log\frac{q_{\phi_x}(z_{ycf}|x)}{p_{\phi_d}(z_x)}dz_{ycf} \\
&\quad -\int\int\int\int \big[q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x) \\
&\quad q_{\phi_{ycf}}(z_{ycf}|x)\log p_{\phi_d}(x|z_x,z_t,z_{yf},z_{ycf})\big]dz_x dz_t dz_{yf} dz_{ycf} \\
&\quad +\log p_{\phi_d}(x) \\
&= KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x)\big)+KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t)\big) \\
&\quad +KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf})\big) \\
&\quad +KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf})\big) \\
&\quad -\mathbb{E}_{q_{\phi_x},q_{\phi_t},q_{\phi_{yf}},q_{\phi_{ycf}}}[\log p(x|z_x,z_t,z_{yf},z_{ycf})] \\
&\quad +\log p_{\phi_d}(x)
\end{aligned}
$$

where, the distributions $q_{\phi_x}(z_x|x)$, $q_{\phi_t}(z_t|x)$, $q_{\phi_{yf}}(z_{yf}|x)$, $q_{\phi_{ycf}}(z_{ycf}|x)$ and $p_{\phi_d}(x|z_x,z_t,z_{yf},z_{ycf}|x)$ are parametrized by the parameters $\phi_x,\phi_t,\phi_{yf},\phi_{ycf},\phi_d$. The KL divergence of two distributions is always greater than or equal to zero. So,

$$
\begin{aligned}
&KL\big(q_{\phi_x}(z_x|x)q_{\phi_t}(z_t|x)q_{\phi_{yf}}(z_{yf}|x)q_{\phi_{ycf}}(z_{ycf}|x)|| \\
&\quad p_{\phi_d}(z_x,z_t,z_{yf},z_{ycf}|x)\big)\geq 0, \\
&\log p_{\phi_d}(x)\geq \mathcal{L}_{ELBO} \qquad \text{where,} \\
&\mathcal{L}_{ELBO}(\phi_x,\phi_t,\phi_{yf},\phi_{ycf};x,z_x,z_t,z_{yf},z_{ycf}) \\
&= \mathbb{E}_{q_{\phi_x},q_{\phi_t},q_{\phi_{yf}},q_{\phi_{ycf}}}[\log p(x|z_x,z_t,z_{yf},z_{ycf})] \\
&\quad -KL\big(q_{\phi_x}(z_x|x)||p_{\phi_d}(z_x)\big)-KL\big(q_{\phi_t}(z_t|x)||p_{\phi_d}(z_t)\big) \\
&\quad -KL\big(q_{\phi_{yf}}(z_{yf}|x)||p_{\phi_d}(z_{yf})\big) \\
&\quad -KL\big(q_{\phi_{ycf}}(z_{ycf}|x)||p_{\phi_d}(z_{ycf})\big)
\end{aligned}
$$

A.2 Variational information maximization

$$
\begin{aligned}
&I(z_c;G(z_G,z_c)) = H(z_c)-H(z_c|G(z_G,z_c)) \\
&= H(z_c)+\int\int p(Z_c=z_c',X=G(z_G,z_c)) \\
&\quad \log p(Z_c=z_c'|X=G(z_G,z_c))\,dz_c dx \\
&= H(z_c)+\mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log(p(z_c'|x)) \\
&= H(z_c)+\mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\left[\frac{p(z_c'|x)}{Q(z_c'|x)}Q(z_c'|x)\right] \\
&= H(z_c)+\mathbb{E}_{x\sim G(z_G,z_c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\left[\frac{p(z_c'|x)}{Q(z_c'|x)}\right] \\
&\quad +\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\Big[Q(z_c'|x)\Big] \\
&= H(z_c)+\mathbb{E}_{x\sim G(z,c)}\int p(z_c'|x)\log\frac{p(z_c'|x)}{Q(z_c'|x)}dc' \\
&\quad +\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{c'\sim p(z_c|x)}\log\Big[Q(z_c'|x)\Big] \\
&= H(z_c)+\mathbb{E}_{x\sim G(z,c)}\big[KL\big(p(z_c'|x||Q(z_c'|x))\big)\big] \\
&\quad +\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\Big[Q(z_c'|x)\Big] \\
&\geq H(z_c)+\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\Big[Q(z_c'|x)\Big] \\
&\geq H(z_c)+\mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)}\mathbb{E}_{z_c'\sim p(z_c|x)}\log\Big[Q(z_c'|x)\Big] \\
&\geq H(z_c)+\mathbb{E}_{z_c\sim p(z_c)}\mathbb{E}_{x\sim G(z,c)}\log\Big[Q(z_c|x)\Big] \\
&\big(\text{by Lemma 5.1 of [7]}\big) \\
&= L_I(G,Q)
\end{aligned}
$$