DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation

Kailiang Zhong*
Fengtong Xiao*
brice.zkl@alibaba-inc.com
fengtong.xiao@alibaba-inc.com
Alibaba Group, China, Singapore

Yan Ren
william.ry@lazada.com
Alibaba Group
China

Yaorong Liang
yaorong.lyr@lazada.com
Alibaba Group
China

Wenqing Yao
wenqing.ywq@alibaba-inc.com
Alibaba Group
China

Xiaofeng Yang
xiaofeng.yang@alibaba-inc.com
Alibaba Group
Singapore

Ling Cen
ling.cen@lazada.com
Alibaba Group
Singapore

ABSTRACT

Causal Inference has wide applications in various areas such as E-commerce and precision medicine, and its performance heavily relies on the accurate estimation of the Individual Treatment Effect (ITE). Conventionally, ITE is predicted by modeling the treated and control response functions separately in their individual sample spaces. However, such an approach usually encounters two issues in practice, i.e. divergent distribution between treated and control groups due to *treatment bias*, and significant *sample imbalance* of their population sizes. This paper proposes Deep Entire Space Cross Networks (DESCN) to model treatment effects from an end-to-end perspective. DESCN captures the integrated information of the treatment propensity, the response, and the hidden treatment effect through a cross network in a multi-task learning manner. Our method jointly learns the treatment and response functions in the entire sample space to avoid treatment bias and employs an intermediate pseudo treatment effect prediction network to relieve sample imbalance. Extensive experiments are conducted on a synthetic dataset and a large-scaled production dataset from the E-commerce voucher distribution business. The results indicate that DESCN can successfully enhance the accuracy of ITE estimation and improve the uplift ranking performance. A sample of the production dataset and the source code are released to facilitate future research in the community, which is, to the best of our knowledge, the first large-scale public biased treatment dataset for causal inference[1].

CCS CONCEPTS

• **Mathematics of computing** → **Causal networks**; • **Computing methodologies** → **Causal reasoning and diagnostics**.

*Both authors contributed equally to this work.
[1]https://github.com/kailiang-zhong/DESCN

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539198

KEYWORDS

Causal Inference; Uplift Modeling; Individual Treatment Effect Estimation; Deep Learning; Entire Space Modeling

ACM Reference Format:
Kailiang Zhong, Fengtong Xiao, Yan Ren, Yaorong Liang, Wenqing Yao, Xiaofeng Yang, and Ling Cen. 2022. DESCN: Deep Entire Space Cross Networks for Individual Treatment Effect Estimation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3534678.3539198

1 INTRODUCTION

Individual-level causal inference is a predictive analytics technique to estimate the Individual Treatment Effect (ITE)[2] of a single or multiple treatments. This technique has wide applications such as identifying the most effective medication to patients [8], and optimizing cross-selling for personalized insurance products [5]. It is also popular in the E-commerce domain as a profit-driven business like voucher distribution and targeted advertising. In this paper, we focus on the ITE estimation task where only a single treatment (i.e., treated or non-treated[3]) exists.

One of the fundamental challenges in the treatment effect modeling (a.k.a uplift modeling) is the existence of *counterfactual outcomes*, that is, we could not observe both treated and control responses in the exact same context for an individual. As a result, the direct measurement of treatment effect from an individual is not observable. A common solution to this challenge is to build models on both response functions separately to derive the counterfactual prediction.

Besides, there are two additional significant issues encountered in practice. First, the underlying distributions for the treated and control groups could be divergent due to the existence of confounding factors. We refer to this as *treatment bias* issue. Confounding factors exist when the treated and control groups are selected not by random, but by some specific factors. Take the E-commerce voucher distribution scenario as an example, a certain discount offer (a.k.a voucher) could only be provided to inactive users with the aim to improve the customer retention rate. The active users,

[2]ITE is sometimes also referred as Conditional Average Treatment Effect (CATE) at the individual level.
[3]Non-treated group is also known as control group.

nevertheless, are not given vouchers for cost-saving purposes. As a result, the users in the control group tend to be more active than those in the treated group. It makes a model difficult to learn an unbiased representation between those two groups through loss functions for ITE estimation, like Counter-factual Regression (CFR)[18].

Second, the sample size between the treated and control group could vary significantly in practice due to particular treatment strategies, which leads to the *sample imbalance* problem. In E-commerce domain, free membership (i.e. treatment) could be given to the majority group of users to promote new products. This makes the sample size of the treated group much larger than the control group. By contrast, in a voucher distribution scenario, it is common that vouchers are only distributed to the users who are promotion-sensitive (i.e. users who will only purchase when a promotion is given), resulting in comparatively small treated group size. Such an imbalanced data could make the model very difficult to learn an accurate estimation of the treatment effects and need extra calibration efforts.

To tackle the above issues, we propose a multi-task cross network to capture the factors of the treatment propensity, the true responses, and the pseudo treatment effect in an integrated framework named Deep Entire Space Cross Networks (DESCN). We introduce an Entire Space Network (ESN) to jointly learn the treatment and response functions in the entire sample space. Instead of separately modeling the Treated Response (TR) and Control Response (CR) in their individual sample space, ESN applies a propensity network to learn the treatment propensity and then connect with TR and CR to derive Entire Space Treated Response (ESTR) and Entire Space Control Response (ESCR). In this manner, the model could be trained on ESTR and ESCR directly where both response functions could leverage the entire samples to address the *treatment bias* issue. Furthermore, it introduces an additional novel cross network (X-network) on top of the shared structure to learn the hidden pseudo treatment effect, which acts as an intermediate variable to connect with the two true response functions (TR and CR) to simulate the counterfactual responses. In this way, we could learn an integrated representation that contains all the responses and the treatment effect information to alleviate the *sample imbalance* issue.

The main contributions of this paper are as follows:

- An end-to-end multi-task cross network, DESCN, is proposed which captures the relationships between the treatment propensity, the true responses, and the pseudo treatment effect in an integrated manner to alleviate *treatment bias* and *sample imbalance* issues simultaneously. Extensive experiments conducted on a synthetic dataset and a large-scale production dataset indicate that DESCN outperforms the baseline models in both ITE estimation accuracy and uplift ranking performance by over +4% to +7%.
- An Entire Space Network (ESN) for modeling the joint distribution between treatment and response functions in the entire sample space. In this way, we could derive the counterfactual information for both treated and control groups in an integrated manner to leverage from entire samples and address the *treatment bias* issue. Note that the ESN is not limited to the DESCN model but could also be applied to other existing individual response functions estimation based uplift models.
- A large-scale production dataset on voucher distribution was collected from the E-commerce platform Lazada. We specifically design the experiment to generate strong treatment bias in the training set but use randomized treatment in the testing set in order to better evaluate the model performance. To the best of our knowledge, this is the first industrial production dataset with both biased and randomized treatments in training and testing set simultaneously, we hope this could help facilitate future research in causal inference.

2 RELATED WORK

Meta learning [7, 26, 27] is a popular framework to estimate ITE, which uses any machine learning estimators as base learners, and identifies the treatment effect by comparing the estimated responses returned by the base models. S-learner and T-learner are two commonly adopted meta learning paradigms. S-learner is trained over the entire space combining both treated and control samples, with the treatment indicator as an additive input feature. T-learner uses two models built individually on the treatment and control sample spaces. In both S-learner and T-learner, when the population size of the two groups is unbalanced, the performance between the corresponding base models could be inconsistent and hurt ITE estimation performance. To overcome this problem, X-learner [10] is proposed, which adopts information learned from the control group to give a better estimation of the treated group and vice versa. The first step of X-learner is to learn two T-learner like models separately. Then, it calculates the difference between the observed outcome and the estimated outcome from both treated and control groups as the imputed treatment effects, which are further used to train a pair of ITE estimators. Finally, the overall ITE is calculated as the weighted average of the two estimators. As with other meta-learning methods, the performance of the base model still has a strong cascading effect in X-learner.

In meta learning, linear methods (e.g., Least Absolute Shrinkage and Selection Operator (LASSO) [21]) may not perform well if the treatment indicator feature is less important or even ignored in modeling. Moreover, the difference of sample size between the treated and control group will significantly influence the training loss contribution in the linear model, which means it can't handle *sample imbalance* issue well. Tree-base methods like Bayesian Additive Regression Trees (BART) [2] and Causal Forest(CF) [23] can alleviate the sample imbalanced problem to a certain degree. BART is employed as a base model to estimate heterogeneous treatment effects in the S-learner, which is able to deal with high-dimensional predictors, yields coherent uncertainty intervals as well as handles continuous treatment variables and missing data for the outcome variable. In [14], a classification tree algorithm is designed with the splitting criteria minimizing the heterogeneous treatment effects in the sub-trees. A non-parametric Causal Forest algorithm proposed first builds a large number of causal trees with different sub-sampling rates and then estimates heterogeneous treatment effects by taking an average of the outcomes from these individual uplift trees. It has the advantage of exploiting huge-sized data

in a large feature space and abstracting inherent heterogeneity in treatment effects.

Recently, representation learning using neural networks becomes the mainstream treatment distribution debiasing methods. Those methods apply networks consisting of both group-dependent layers and shared layers over the entire treated and control groups, and utilize several extended regularisation strategies to address treatment bias inherent in the observation data [16]. Balancing Neural Network (BNN) [9] uses Integral Probability Metric (IPM) to minimize the discrepancy between the distributions of treatment and control samples. Treatment-Agnostic Representation Network (TARNet) and Counterfactual Regression (CFR) [18] extend BNN into a two-headed structure in the multi-task learning manner, learning two functions based on a shared and balanced feature representation across the treated and control spaces. The similarity-preserved individual treatment effect (SITE) estimation method [24] improves the CFR by adding a position-dependent deep metric (PPDM) and middle-point distance minimization (MPDM) constraints, which not only balances the distributions of the treated and control population but also preserves the local similarity information. Furthermore, the Perfect Match (PM) method [17] extends the TARNet to multi-treatments scenarios, augmenting the samples within a min-batch and matching their propensity with nearest neighbours. An adapting neural network, DragonNet [19] is designed for predicting propensity score and conditional outcome in an end-to-end procedure. However, those methods do not address the previously mentioned two issues together.

In recommendation systems, the Entire Space Multi-Task Model (ESMM) [11] has been introduced to solve the sample selection bias problem by exploiting the sequential pattern of user behaviors to model both user clicking and user purchasing in the entire space. In this paper, we apply a similar idea of multi-task learning and entire sample space modeling to causal inference domain. We exploit the structural properties among the treatment propensity, the true response, and the pseudo treatment effect to estimate ITE in the entire sample space to alleviate the *treatment bias* and *sample imbalance* issues.

3 METHODOLOGY

This section introduces the DESCN model for ITE estimation. It starts with the problem definition, followed by the illustration of DESCN and its two components: Entire Space Network (ESN) and X-network.

3.1 Problem Definition

We follow the Neyman-Rubin potential outcome framework [13] to define the ITE estimation problem. To be specific, assuming we observe samples $\mathcal{D} = \{(y_i, x_i, w_i)\}_{i=1}^{n}$, with $(Y, X, W) \overset{i.i.d.}{\sim} \mathbb{P}$. Here, for each sample i, $y_i \in \{0, 1\}$ is a binary treatment outcome of interest; $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional covariate or feature vector; $w_i \in \{0, 1\}$ denotes the binary treatment. Let $y_i(1)$ and $y_i(0)$ denote the potential outcome of individual i when i is being treated ($w_i = 1$) or non-treated ($w_i = 0$). The treatment is assigned according to the propensity score $\pi(x) = P(W = 1|X = x)$.

Further, let $T = \{i : w_i = 1\}$ and $C = \{i : w_i = 0\}$ denotes the set of treated samples and control samples respectively. In this way, T and C represent the treated and control sample space, and we refer to them as sub-sample spaces. The union of T and C represents the whole sample set and referred as the entire sample space.

The major challenge in ITE estimation is that only one potential outcome, either $y_i(0)$ or $y_i(1)$, is observable for each individual i but not both, that is, $y_i = w_i y_i(1) + (1 - w_i) y_i(0)$, where $w_i \in \{0, 1\}$. Hence, there is no true uplift value of each sample $y_i(1) - y_i(0)$. Our interest is to estimate the expected individual treatment effect (ITE) with covariate values $X = x$, denoted as $\tau(x)$:

$$\tau(x) = \mathbb{E}_{\mathbb{P}}(Y(1) - Y(0)|x), \tag{1}$$

We consider the problem under three standard assumptions:

- Consistency: if individual i is assigned treatment w_i, we observe a consistent associated potential outcome $y_i = y_i(w_i)$.
- Ignorability: there are no unobserved confounders, such that $Y(1), Y(0) \perp\!\!\!\perp W|X$.
- Overlap: treatment assignment is non-deterministic, i.e. $0 < \pi(x) < 1, \forall x \in \mathcal{X}$.

Treated Response (TR) and Control Response (CR) are the outcome of interest under the treated and non-treated (control) groups, which are denoted as $\mu_1(x)$ and $\mu_0(x)$:

$$\begin{aligned} \mu_1(x) &= \mathbb{E}_{\mathbb{P}}(Y|W = 1, X = x), \\ \mu_0(x) &= \mathbb{E}_{\mathbb{P}}(Y|W = 0, X = x) \end{aligned} \tag{2}$$

In this way, ITE could be written as $\tau(x) = \mu_1(x) - \mu_0(x)$. In order to estimate τ from observational data, a natural idea is to estimate μ_1 and μ_0 on the treated and control sub-sample space respectively, and then get $\hat{\tau} = \hat{\mu}_1 - \hat{\mu}_0$[4]. However, such a method has the following two problems:

- **Treatment Bias:** As the treatment is assigned according to the propensity score π, there is systematic difference between the treated and control group data distributions.
- **Sample Imbalance:** The population size of treated and control subspace could vary significantly.

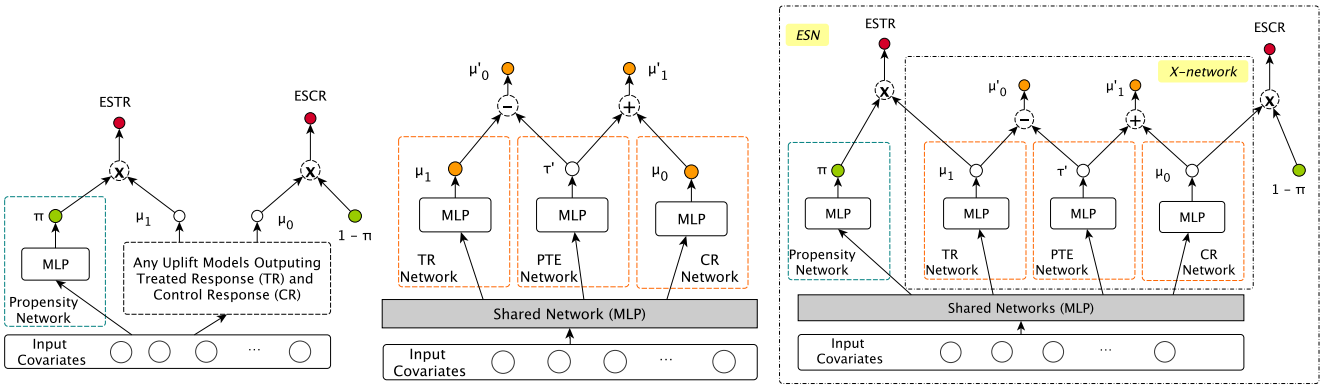
In order to address those two issues, we propose a novel Entire Space Network (ESN) and a multi-task cross network (X-network). The combination of ESN and X-network gives us Deep Entire Space Cross Networks (DESCN). Details of each network will be introduced in the following sections.

3.2 Entire Space Network (ESN)

The estimation of ITE requires the estimation of μ_1, μ_0. Inspired by ESMM[11], we define two associated probabilities: Entire Space Treated Response (ESTR) as $P(Y, W = 1|X)$, and the Entire Space Control Response (ESCR) as $P(Y, W = 0|X)$. Their probabilities follows Eq.(3):

$$\begin{aligned} \underbrace{P(Y, W = 1|X)}_{ESTR} &= \underbrace{P(Y|W = 1, X)}_{TR} \cdot \underbrace{P(W = 1|X)}_{\pi} \\ &= \mu_1 \cdot \pi \\ \underbrace{P(Y, W = 0|X)}_{ESCR} &= \underbrace{P(Y|W = 0, X)}_{CR} \cdot \underbrace{P(W = 0|X)}_{1-\pi} \\ &= \mu_0 \cdot (1 - \pi) \end{aligned} \tag{3}$$

[4] $\hat{f}$ denotes the estimation of the true function f.



(a) Entire Space Network (ESN) (b) X-network (c) Deep Entire Space Cross Networks (DESCN)

Figure 1: Model Architectures of Entire Space Network (ESN), X-network, and DESCN. Note that DESCN is the combination of ESN and X-network. All nodes associated with training loss are highlighted in color.

Note that ESTR, ESCR, and propensity score could be modeled on all samples to get the corresponding unbiased estimated values. Hence, $\hat{\mu_1}$ and $\hat{\mu_0}$ could be derived from them by dividing the estimated ESTR and ESCR by π and $1-\pi$ respectively. We could get the estimation of ITE thereafter.

With such structural properties in conditional probabilities, we propose an Entire Space Network (ESN) as shown in Figure 1a. The leftmost part of the network is to model the propensity score $\pi(X)$, with the output node π. On the right are two additional nodes μ_1 and μ_0, representing the modeling of TR and CR respectively. The output of μ_1 node and π node are multiplied to get ESTR, while the output of μ_0 is multiplied with 1 - π to get ESCR as shown in Eq.3. In this way, instead of learning μ_0 and μ_1 in their individual sub-sample spaces, we derive them in integration by modeling the ESTR, ESCR, and π in the entire sample space. As a result, a sample from the treated group will not only contribute to the learning of treated response function but also contribute to the learning of control response function, and vice versa. The counterfactual information for both treated and control groups could be derived in the integrated manner from entire samples, which could alleviate the *treatment bias* issue. Note that ESN has to be used with other uplift models that output the estimation of TR and CR, such as TARNet, CFR [18], X-network, etc. A more specific performance comparison on adding ESN to those existing methods is included in Section 5.2.

The training loss function of ESN is the weighted sum of the following[5]

$$
\begin{aligned}
L_{\pi} &= \frac{1}{n}\sum_{i} l(t_i, \hat{\pi}(x_i)) \\
L_{ESTR} &= \frac{1}{n}\sum_{i} l\big(y_i \& w_i, \hat{\mu}_1(x_i)\cdot\hat{\pi}(x_i)\big) \\
L_{ESCR} &= \frac{1}{n}\sum_{i} l\big(y_i \& (1-w_i), \hat{\mu}_0(x_i)\cdot(1-\hat{\pi}(x_i))\big)
\end{aligned}
\tag{4}
$$

[5]& denotes the logical AND operation between two binary label y_i and w_i.

that is,

$$
L_{ESN} = \alpha \cdot L_{\pi} + \beta_1 \cdot L_{ESTR} + \beta_0 \cdot L_{ESCR} \tag{5}
$$

where $l(\cdot)$ denotes the loss function (i.e. cross entropy loss) and α, β_1, β_0 are the hyper-parameters denotes the weight of corresponding loss functions.

3.3 X-network

Inspired by X-learner [10], to further address the *sample imbalance* issue, we propose a cross network named X-network as shown in Figure 1b. Besides the regular networks modeling Treated Response (TR Network) and Control Response (CR Network), we introduce the Pseudo Treatment Effect network (PTE network) to learn the Pseudo Treatment Effect (PTE) τ', that is, the hidden observable effect brought by the treatment. PTE variable τ' acts as an intermediate variable to connect with both TR and CR, which can balance the learning between those two response functions during the training process.

We add the outputs from the PTE Network and the CR Network to derive the Cross Treated Response $\mu_1' := \mu_0 + \tau'$, which represents the counterfactual results if the individuals of the control group were given treatment. Similarly, we deduct the pseudo treatment effect from TR to derive the Cross Control Response $\mu_0' := \mu_1 - \tau'$, which represents the counterfactual results when the individuals of the treated group did not receive treatment. With these additional nodes, we build the connection between TR and CR through the PTE τ', and learn a shared representation in the bottom shared network. This helps to learn both TR and CR in a more balanced and constrained manner, especially when one of the treated or control sample size is significantly smaller or larger than the other. Note that we do not force the model to learn a similar representation across treated and control groups through regularization as introduced in CFR[18], instead, we utilize the PTE to connect those two groups. In this manner, we are not smoothing over different units and thus improve the ITE estimation performance. The X-network has 4 losses in the treated or control sub-sample

Dataset	Covariates	Training Data				Testing Data			
		Treated	Control	Total	Positive Outcome (percentage)	Treated	Control	Total	Positive Outcome (percentage)
Epilesy	178	20.2K	19.8K	40k	18.4k (45.9%)	19.8K	20.1K	40k	18.1k (45.3%)
Production	83	0.92M	3.25M	4.17M	83.0k (2.0%)	0.47M	0.43M	0.91M	31.9k (3.5%)

Table 1: Statistics of the synthetic dataset *Epilesy* and real-world Production Dataset.

spaces defined as:

$$
\begin{aligned}
L_{TR} &= \frac{1}{|T|}\sum_{i\in T} l(y_i, \hat{\mu}_1(x_i)),\\
L_{CR} &= \frac{1}{|C|}\sum_{i\in C} l(y_i, \hat{\mu}_0(x_i)),\\
L_{CrossTR} &= \frac{1}{|T|}\sum_{i\in T} l(y_i, \hat{\mu}_1'(x_i))\\
&= \frac{1}{|T|}\sum_{i\in T} l\Big(y_i, \sigma\big(\sigma^{-1}(\hat{\mu}_0(x_i)) + \sigma^{-1}(\hat{\tau}'(x_i))\big)\Big),\\
L_{CrossCR} &= \frac{1}{|C|}\sum_{i\in C} l(y_i, \hat{\mu}_0'(x_i))\\
&= \frac{1}{|C|}\sum_{i\in C} l\Big(y_i, \sigma\big(\sigma^{-1}(\hat{\mu}_1(x_i)) - \sigma^{-1}(\hat{\tau}'(x_i))\big)\Big)
\end{aligned}
\tag{6}
$$

where σ is the sigmoid function, L_{TR} and L_{CR} represents the loss for Treated Response and Control Response, $L_{CrossTR}$ and $L_{CrossCR}$ represents the loss for Cross Treated Response and Cross Control Response respectively.

Note that instead of learning the direct values of μ_1 and μ_0, we learn their logit value (i.e. the input of sigmoid function) in TR network and CR network respectively for numerical stability. Specifically, $\hat{\tau}'$ is added(subtracted) on logits of $\hat{\mu}_0(\hat{\mu}_1)$ to drive $\hat{\mu_1}'(\hat{\mu_0}')$, i.e. $\hat{\mu_1}' = \sigma\big(\sigma^{-1}(\hat{\mu}_0) + \sigma^{-1}(\hat{\tau}')\big)$, and likewise for $\hat{\mu_0}'$. Another benefit of such transformation is that we do not need worry about the addition or subtraction getting out of the range $[0, 1]$. Besides, when $\hat{\mu}_0$ and $\hat{\mu}_1$ is very close to 0 or 1, the σ^{-1} function could magnify the uplift signal and make the MLP learning process easier. Empirically, we find that this transformation performs better than directly outputting the true response values.

We can see that our X-network is similar to X-learner in that both try to directly learn the counterfactual treatment effect. In X-learner, ITE is learned based on the results from base learners, and its performance is heavily subject to that of the base models. By contrast, in X-network, ITE is learned together with the base learners in an integrated way.

3.4 Model Architecture of DESCN

The overall model architecture of DESCN is presented in Figure 1c, which is the direct application of Entire Space Network (ESN) on X-network. In this way, both *treatment bias* and *sample imbalance* mentioned in 3.1 could be alleviated. Furthermore, the shared network in DESCN could learn both the propensity score and the control response with pseudo treatment effect simultaneously, which could also help learn a comprehensive representation capturing all this information.

For the overall training, we have the final loss for DESCN which is the weighted sum of the propensity score loss L_π, the Cross Treated Response $L_{CrossTR}$, and Cross Control Response losses $L_{CrossCR}$:

$$
\begin{aligned}
L_{DESCN} &= L_{ESN} + \gamma_1 \cdot L_{CrossTR} + \gamma_0 \cdot L_{CrossCR}\\
&= \alpha \cdot L_\pi + \beta_1 \cdot L_{ESTR} + \beta_0 \cdot L_{ESCR}\\
&+ \gamma_1 \cdot L_{CrossTR} + \gamma_0 \cdot L_{CrossCR}
\end{aligned}
\tag{7}
$$

Note that we remove L_{TR} and L_{CR} losses in Eq.(6) in section 3.3 since TR and CR are now connected with propensity π as ESTR and ESCR, which are trained in the entire sample space instead.

4 EXPERIMENTS

4.1 Dataset

To compare the model performance from both ITE estimation accuracy and uplift ranking perspectives, we use a synthetic dataset where the ground truth treatment effect is available and another real-world dataset where the *treatment bias* and *sample imbalance* issues exist. The statistics of those two datasets are summarized in Table 1.

The first dataset is a synthetic dataset generated based on the public code from the 2019 American Causal Inference Conference (ACIC) data challenge[6]. We use the provided Data Generator Processor (DPG) on the covariates collected from the high dimensional Epileptic Seizure Recognition Dataset (*Epilepsy*) [1] to generate 40k observational samples[7]. The advantage of this synthetic dataset is that the ground truth treatment effect is fully known from DPG, thus the *Epilepsy* dataset is ideal to evaluate the model performance on ITE estimation accuracy.

To further evaluate the model performance on uplifting ranking performance, we use another large-scale production dataset from the real voucher distribution business scenario in Lazada, a leading South-East Asia (SEA) E-commerce platform of Alibaba Group. In the real production environment, the treatment assignment is selective due to the operation targeting strategy, and we collect those data as our training set which includes strong treatment bias. We also have a slightly smaller size of users who are not affected by the targeting strategy and the treatment assignment follows the randomized controlled trials (RCT). We use them as our testing dataset.

By the above setup, the training set contains the biased treated data while the testing set only contains unbiased treated data. In the training set, the treated and control distribution is naturally divergent due to *treatment bias* issue and the sample size between the two groups differ naturally in the real-world scenario as shown in Table 1.

[6]https://sites.google.com/view/acic2019datachallenge/home
[7]We adopt method that involves complex models and treatment effect heterogeneity, indicated as *Mod 4* in the original source code.

Model	Epilepsy Dataset			Production Dataset		
	$\sqrt{\epsilon_{PEHE}}$ mean ± s.e.	Impr (CFR_{mmd})	ϵ_{ATE} mean ± s.e.	AUUC mean ± s.e.	Impr (CFR_{mmd})	ϵ_{ATT} mean ± s.e.
X-leaner (NN)	0.1556 ± 0.0018	-15.8%	0.0378 ± 0.0059	0.0234 ± 0.0035	-27.9%	0.0076 ± 0.0009
Causal Forest	0.1519 ± 0.0042	-13.0%	0.0663 ± 0.0086	0.0132 ± 0.0008	-59.2%	0.0123 ± 0.0003
BART	0.1387 ± 0.0004	-3.2%	0.0389 ± 0.0004	0.0222 ± 0.0003	-31.5%	0.0312 ± 0.0001
TARNet	0.1373 ± 0.0028	-2.2%	0.0405 ± 0.0091	0.0309 ± 0.0021	-4.6%	0.0106 ± 0.0016
CFR_{wass}	0.1363 ± 0.0031	-1.4%	0.0263 ± 0.0097	0.0261 ± 0.0002	-19.4%	0.0266 ± 0.0013
CFR_{mmd}	0.1344 ± 0.0027	0.0%	0.0305 ± 0.0069	0.0324 ± 0.0029	0.0%	0.0258 ± 0.0015
X-network	0.1289 ± 0.0026	4.1%	0.0245 ± 0.0044	0.0324 ± 0.0016	0.0%	0.0048 ± 0.0010
DESCN	**0.1241 ± 0.0009**	**7.6%**	**0.0058 ± 0.0014**	**0.0340 ± 0.0006**	**4.9%**	**0.0039 ± 0.0007**

Table 2: Model performance evaluated by ϵ_{PEHE}, ϵ_{ATE} on the Epilepsy Dataset and by AUUC, ϵ_{ATT} on the Production Dataset with corresponding mean and standard error. Note that for ϵ_{PEHE}, ϵ_{ATE} and ϵ_{ATT}, smaller value is better, and for AUUC larger value is better. Best results of all methods are highlighted in boldface.

4.2 Compared Methods

We compare DESCN with the following models which are commonly adopted in ITE estimation to evaluate its performance. All models use all input dense covariates in the two datasets.

- **X-learner (NN)**[10]: We take X-learner as the representative meta-learners. Neural Network (NN) is chosen as the base learner, which has a better non-linear learning advantage and provides a fair comparison with our proposed NN-based model. Two response fitting models, two imputed treatment effects fitting sub-models, and one propensity model are included. All models are trained separately without the shared network parameters. Estimated propensity scores are used for the final ITE output for better performance.
- **BART**[2]: Bayesian Additive Regression Trees (BART) is a sum-of-trees model, used to assess the performance of non deep-learning models.
- **Causal Forest**[23]: Causal Forest is a non-parametric Random Forest based tree model that directly estimates the treatment effect, which is another representative of tree-based uplift models.
- **TARNet**[18]: TARNet is a commonly used deep learning uplift model. Compared with X-learner (NN), it omits the additional imputed treatment effects fitting sub-models but introduces the shared layers for treated and control response networks. The shared network parameters could help alleviate the *sample imbalance*.
- **CFR**[18]: CFR applies an additional loss to TARNet, which forces the learned treated and control covariate distributions to be closer. This could help to learn a balanced representation between those two groups and address the imbalance issue. We report the CFR performance using two distribution distance measurement loss functions, Wasserstein [3, 22] (denoted as $\mathbf{CFR}_{wass}$) and Maximum Mean Discrepancy (MMD) [4] (denoted as $\mathbf{CFR}_{mmd}$).
- **X-network**: X-network is introduced in section 3.3. It can be regarded as a variant of our final model DESCN where the ESN part is removed.

4.3 Evaluation Metrics

We adopt different commonly used uplift modeling evaluation metrics for different datasets based on whether the dataset contains the true uplift treatment effect value. More specifically, for *Epilesy* synthetic dataset where the response generating process is known, we use the expected Precision in Estimation of Heterogeneous Effect ϵ_{PEHE} [17, 18] and the absolute error in average treatment effect ϵ_{ATE}:

$$\begin{aligned} \epsilon_{PEHE} &= \frac{1}{n}\sum_i \left[\left(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right) - \tau(x_i)\right]^2 \\ \epsilon_{ATE} &= \left|\frac{1}{n}\sum_i \left(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right) - \frac{1}{n}\sum_i \tau(x_i)\right| \end{aligned} \tag{8}$$

Generally, ϵ_{PEHE} could better measure the CATE prediction accuracy at an individual level, while ϵ_{ATE} is a better indicator of the average effect difference across the given sample groups.

Since the ground truth for uplift value is impossible to retrieve on the production dataset, we report the normalized Area Under the Uplift Curve[8] (AUUC) [6, 15, 20, 25] value on the testing dataset, which is an indicator evaluating the uplift score ranking performance. We also compute the error of the Average Treatment Effect on the Treated group ϵ_{ATT} on this randomized test set:

$$\begin{aligned} ATT &= \frac{1}{|T|}\sum_{i\in T} y_i - \frac{1}{|C|}\sum_{i\in C} y_i \\ \epsilon_{ATT} &= \left|\frac{1}{|T|}\sum_{i\in T}\left(\hat{\mu}_1(x_i) - \hat{\mu}_0(x_i)\right) - ATT\right| \end{aligned} \tag{9}$$

Besides the evaluation metrics, each experiment is repeated 5 times. Mean and standard error of all metrics are reported. Finally, we also calculate the relative improvement of a model over the baseline model for any evaluation metric E as:

$$Impr(BaselineModel) = \frac{E(Model) - E(BaselineModel)}{E(BaselineModel)} \times 100\% \tag{10}$$

[8] Also known as *Qini coefficient* in some literature.

4.4 Hyper-parameter Settings

We keep the same hyper-parameters in the common components across different models. Specifically, in the *Epilepsy* dataset, all Neural Networks across different deep models consist of 128 hidden units with 3 fully connected layers. L2 regularization is applied to alleviate over-fitting with coefficient of 0.01, and the learning rate is set as 0.001 without decay. All the models are trained with a batch size of 500 and epoch as 15, the only different hyper-parameters are those associated with each specific network setting, such as the weights of different loss functions. Similarly, for the production dataset, all Neural Networks also consist of 3 fully connected layers. The shared network contains 128 hidden units and other sub-models contain 64 hidden units. The L2 regularization value is set as 0.001, and the learning rate is set as 0.001 with the decay rate as 0.95. All the models are trained with a batch size of 5000 and 5 epoches[9].

5 RESULTS AND DISCUSSIONS

In this section, we present the overall performance of DESCN and the other methods to be compared. The following questions are also proposed and investigated:

- Q1: Is the proposed Entire Space Network (ESN) useful in improving the overall performance and alleviating the *treatment bias* issue?
- Q2: Are the proposed cross network (X-network) and the pseudo treatment effect (PTE) network helpful in learning balanced treated and control response functions?

The experimental results of DESCN and other baseline models are summarized in Table 2.

5.1 Overall Performance

Based on the results from both datasets in Table 2, we have the following observations:

- DESCN consistently outperforms the other baselines in both datasets in terms of different metrics, with reliability indicated by the small standard errors. Compared to CFRmmd, one of the most commonly used deep uplift models, DESCN achieves over +7% relative improvement in ϵ_{PEHE} on the Epilepsy dataset, and +4% relative improvement in AUUC on the highly biased production dataset. This indicates that DESCN is capable of handling both *treatment bias* and *sample imbalance* issues well and can achieve a higher performance in both individual-level treatment effect estimation accuracy and uplift ranking.
- DESCN achieves a more significant improvement in terms of ϵ_{ATE} and ϵ_{ATT} on the Epilepsy and the production dataset by +80% and +84%, respectively. With the help of X-network modeling treated and control distribution in a more balanced manner, DESCN could also improve the average treatment effect estimation accuracy for treated and control groups.

5.2 Entire Space Network (ESN) Analysis (Q1)

To further demonstrate the effectiveness by introducing the Entire Space Network (ESN) for jointly learning the treatment and response function in the entire sample space, we remove the ESN network from DESCN, which is the same as X-network as shown in Figure 1b. The comparison of the performance with the full structure is as listed in the last two rows of Table 2:

- From the table, we could find that after introducing ESN structure, DESCN improves X-network on ϵ_{PEHE} +3.7% and ϵ_{ATE} +76.3% on Epilepsy dataset. It also improves AUUC +4.9% and ϵ_{ATT} +18.7% on the production dataset. This indicates that by combining the learned treatment propensity score with learned responses in their individual sample space and directly learning ESTR and ESCR, the model could better capture the joint distribution in the entire sample space.

Moreover, the ESN is not limited to the DESCN model but also has the ability to extend to other existing individual response estimation methods to enhance the performance. We apply ESN to both TARNet and CFR in the same way as described in Section 3.2, and show their performance in Table 3 and Table 4 on both datasets:

Model	$\sqrt{\epsilon_{PEHE}}$ mean ± s.e.		ϵ_{ATE} mean ± s.e.
TARNet	0.1373 ± 0.0028		0.0405 ± 0.0091
ESN+TARNet	0.1320 ± 0.0008	+3.9%	0.0233 ± 0.0053
CFR_{wass}	0.1344 ± 0.0027		0.0305 ± 0.0069
ESN+CFR_{wass}	0.1361 ± 0.0034	-1.3%	0.0271 ± 0.0053
CFR_{mmd}	0.1363 ± 0.0031		0.0263 ± 0.0097
ESN+CFR_{mmd}	0.1577 ± 0.0013	-15.7%	0.0183 ± 0.0063

Table 3: Experimental results of applying ESN to TARNet and CFR on Epilepsy Dataset. Evaluated by ϵ_{PEHE} and ϵ_{ATE}, the ϵ_{PEHE} improvement after applying ESN is included.

Model	AUUC mean ± s.e.		ϵ_{ATT} mean ± s.e.
TARNet	0.0309 ± 0.0021		0.0106 ± 0.0016
ESN+TARNet	0.0340 ± 0.0008	+10.0%	0.0165 ± 0.0017
CFR_{wass}	0.0261 ± 0.0002		0.0266 ± 0.0013
ESN+CFR_{wass}	0.0264 ± 0.0018	+1.1%	0.0212 ± 0.0015
CFR_{mmd}	0.0324 ± 0.0029		0.0258 ± 0.0015
ESN+CFR_{mmd}	0.0331 ± 0.0005	+2.1%	0.0207 ± 0.0010

Table 4: Experimental results of applying ESN to TARNet and CFR on Production Dataset. Evaluated by AUUC and ϵ_{ATT}, the AUUC improvement after applying ESN is included.

- With the help of alleviating treatment bias issue by learning treatment and response functions in the entire sample space, ESN enhances the performance of TARNet by +3.9% in ϵ_{PEHE} on Epilepsy Dataset and 10.0% in AUUC on the production dataset. This indicates that ESN is capable of enhancing the individual ITE estimation accuracy and uplifting ranking performance. However, it does not show much improvement over CFR on the Epilepsy dataset and with a slight improvement on the production dataset over those

[9] More details regards experiment hyper-parameter settings could be found online.

two metrics. A possible reason is that CFR already uses the additional loss function (either WASS or MMD) to enforce a similar representation of treated and control response in their individual sample space. It might be in conflict with the additional propensity information learned in the ESN.

- Furthermore, we observe that ESN improves ϵ_{ATE} +42.5% on TARNet, +11.1% on CFRwass, and +30.4% on CFRmmd on the Epilepsy dataset. For the production dataset, although it does not improve on ϵ_{ATT} on the TARNet, it largely improves ϵ_{ATT} on CFRwass and CFRmmd by +20.3% and +19.8% respectively. A similar trade-off between estimating average effect and estimating individual effect is also observed in [18].
- We believe that debiasing treated and control distribution through ESN or loss functions in CFR will improve the individual estimation accuracy but hurt the average estimation performance to a certain degree. In the production dataset, the treatment bias is more significant than it in the Epilepsy dataset, hence the ESN+TARNet performs slightly worse than TARNet in ϵ_{ATT}. Nonetheless, ESN could still slightly improve CFR performance on both datasets, which shows that training ESTR and ESCR in the entire sample space is still beneficial.
- We can explore why the ESN module in DECSN can reduce the treatment bias from the ATE perspective. Firstly, according to the IPW(inverse probability weighting)[12] theory we know that $ATE = \mathbb{E}_{\mathbb{P}}\{\frac{W_i \cdot Y_i}{\pi(X_i)}\} - \mathbb{E}_{\mathbb{P}}\{\frac{(1-W_i)\cdot Y_i}{1-\pi(X_i)}\}$. Combining the Eq.(3) in section 3.2, we can expand the equation as:

$$
\begin{aligned}
ATE &= \frac{1}{N}\sum_i \frac{\mathbb{E}_{\mathbb{P}}\{W_i \cdot Y_i | X_i\}}{\pi(X_i)} - \frac{1}{N}\sum_i \frac{\mathbb{E}_{\mathbb{P}}\{(1-W_i)\cdot Y_i | X_i\}}{\pi(X_i)} \\
&= \frac{1}{N}\sum_i \frac{ESTR_i}{\pi(X_i)} - \frac{1}{N}\sum_i \frac{ESCR_i}{1-\pi(X_i)} \\
&= \frac{1}{N}\sum_i \mu_1(X_i) - \frac{1}{N}\sum_i \mu_0(X_i)
\end{aligned}
$$

  At this point, we can see its mechanism to reduce the treatment bias.

5.3 Cross Network (X-network) Analysis (Q2)

Recall that the design of the cross network (X-network) utilizes a specially designed intermediate Pseudo Treatment Effect (PTE) variable to connect with both Treat Response (TR) and Control Response (CR) to balance the learned response functions. As the X-network structure only differs with TARNet in this additional PTE Network, we compare the performance of X-network and TARNet in Table 2.

We can see that X-network improves the ϵ_{PEHE} +9.6% on the Epilepsy dataset and AUUC +10.0% on the production dataset. Furthermore, there also exists significant improvement in ϵ_{ATE} and ϵ_{ATT}. All these prove that the designed X-network is capable of learning each response function more balanced, and can well simulate the hidden treatment effect.

6 CONCLUSION AND FUTURE WORK

In this paper, we propose an integrated multi-task model for Individual Treatment Effect (ITE) estimation, Deep Entire Space Cross Networks (DESCN). It introduces an Entire Space Network (ESN) to jointly learn the distribution of treatment and response functions in the entire sample space to address the *treatment bias* issue. Moreover, a novel cross network (X-network) is presented which is capable of directly modeling hidden treatment effects, and also has the advantage of alleviating distribution *sample imbalance*. Finally, DESCN combines ESN and X-network in a multi-task learning manner, which integrates information of the treatment propensity, the response, and the hidden treatment effect through a cross network. Extensive experiments in both synthetic dataset adapted from the ACIC data challenge and a large-scale voucher distribution production dataset are conducted. The results show that DESCN achieves over +4% to +7% improvement in ϵ_{PEHE} and AUUC on the synthetic dataset and production dataset respectively. This demonstrates that DESCN is effective in ITE estimation accuracy and uplift ranking.

There are several potential future works for exploration. In industrial applications, multiple treatments could be given at the same time. An extension of DESCN to the multi-treatment setting could be studied by extending the X-network to multiple response functions. Another direction is to extend DESCN to fit continuous treatment outcomes, which also has wide applications.

REFERENCES

[1] Ralph G. Andrzejak, Klaus Lehnertz, Florian Mormann, Christoph Rieke, Peter David, and Christian E. Elger. 2001. Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Physical Review E* 64, 6 (Nov. 2001). https://doi.org/10.1103/physreve.64.061907

[2] Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* 4, 1 (2010), 266–298. https://doi.org/10.1214/09-aoas285

[3] Marco Cuturi and Arnaud Doucet. 2014. Fast Computation of Wasserstein Barycenters. arXiv:1310.4375 [stat.ML]

[4] Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A Kernel Two-Sample Test. *Journal of Machine Learning Research* 13, 25 (2012), 723–773. http://jmlr.org/papers/v13/gretton12a.html

[5] Leo Guelman, Montserrat Guillén, and Ana M. Pérez-Marín. 2015. A decision support framework to implement optimal personalized marketing intervention. *Decision Support Systems (Elsevier)* 72 (april 2015), 24–32. https://doi.org/10.1016/j.dss.2015.01.010

[6] Pierre Gutierrez and Jean-Yves Gerardy. 2016. Causal Inference and Uplift Modeling A review of the literature. *JMLR: Workshop and Conference Proceedings 67* (2016).

[7] Pierre Gutierrez and Jean-Yves Gérardy. 2016. Causal Inference and Uplift Modelling: A Review of the Literature. *PAPIs* (July 2016).

[8] JMaciej Ja´skowski and Szymon Jaroszewicz. 2012. Uplift modeling for clinical trial data. *ICML 2012 Workshop on clinical data analysis* (2012).

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

[10] Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. Meta-learners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences* 116, 10 (Feb. 2019), 4156–4165. https://doi.org/10.1073/pnas.1804597116

[11] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire Space Multi-Task Model: An Effective Approach for Estimating Post-Click Conversion Rate. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval* (2018).

[12] Paul R Rosenbaum and Donald B Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 1 (1983), 41–55.

[13] Donald B. Rubin. 2005. Causal Inference Using Potential Outcomes. *J. Amer. Statist. Assoc.* 100 (2005), 322 – 331.

[14] Piotr Rzepakowski and Szymon Jaroszewicz. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowl. Inf. Syst.* 32, 2 (Aug. 2012), 303–327. https://doi.org/10.1007/s10115-011-0434-0

[15] Piotr Rzepakowski and Szymon Jaroszewicz. 2012. Decision trees for uplift modeling with single and multiple treatments. *Knowl. Inf. Syst.* 32, 2 (Aug. 2012), 303–327.

[16] Patrick Schwab, Lorenz Linhardt, Stefan Bauer, Joachim M. Buhmann, and Walter Karlen. 2020. Learning Counterfactual Representations for Estimating Individual Dose-Response Curves. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (April 2020), 5612–5619. https://doi.org/10.1609/aaai.v34i04.6014

[17] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. 2019. Perfect Match: A Simple Method for Learning Representations For Counterfactual Inference With Neural Networks. (2019).

[18] Uri Shalit, Fredrik D. Johansson, and David Sonta. 2017. Estimating individual treatment effect: generalization bounds and algorithms. *Proceedings of the 34th International Conference on Machine Learning* V1 (Aug. 2017), 3076–3085.

[19] Claudia Shi, David M. Blei, and Victor Veitch. 2019. Adapting Neural Networks for the Estimation of Treatment Effects. *Advances in Neural Information Processing Systems* (2019).

[20] Michał Sołtys, Szymon Jaroszewicz, and Piotr Rzepakowski. 2015. Ensemble methods for uplift modeling. *Data Min. Knowl. Discov.* 29, 6 (Nov. 2015), 1531–1559.

[21] Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Statist* (1996), 267–288.

[22] Cédric Villani. 2008. Optimal Transport: Old and New.

[23] Stefan Wager and Susan Athey. 2018. Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* 113, 523 (June 2018), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

[24] Liuyi Yao, Sheng Li, Yaliang Li, Mengdi Huai, Jing Gao, and Aidong Zhang. 2018. Representation Learning for Treatment Effect Estimation from Observational Data. *Advances in Neural Information Processing Systems* (2018), 2638–2648.

[25] Yan Zhao, Xiao Fang, and David Simchi-Levi. 2017. Uplift Modeling with Multiple Treatments and General Response Types. (May 2017). arXiv:1705.08492 [cs.AI]

[26] Zhenyu Zhao and Totte Harinen. 2019. Uplift Modeling for Multiple Treatments with Cost Optimization. *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)* (Jan. 2019), 422–431. https://doi.org/10.1109/DSAA.2019.00057

[27] Zhenyu Zhao, Yumin Zhang, Totte Harinen, and Mike Yung. 2020. Feature Selection Methods for Uplift Modeling. *Computer Science, Mathematics* (May 2020). https://arxiv.org/abs/2005.03447