

CAUSALMTA: Eliminating the User Confounding Bias for Causal Multi-touch Attribution

Di Yao¹, Chang Gong¹, Lei Zhang², Sheng Chen², Jingping Bi¹

¹ Institute of Computing Technology, Chinese Academy of Sciences

² Alibaba Inc.

Abstract

Multi-touch attribution (MTA), aiming to estimate the contribution of each advertisement touchpoint in conversion journeys, is essential for budget allocation and automatically advertising. Existing methods first train a model to predict the conversion probability of the advertisement journeys with historical data and calculate the attribution of each touchpoint using counterfactual predictions. An assumption of these works is the conversion prediction model is unbiased, *i.e.*, it can give accurate predictions on any randomly assigned journey, including both the factual and counterfactual ones. Nevertheless, this assumption does not always hold as the exposed advertisements are recommended according to user preferences. This confounding bias of users would lead to an out-of-distribution (OOD) problem in the counterfactual prediction and cause concept drift in attribution. In this paper, we define the causal MTA task and propose CAUSALMTA to eliminate the influence of user preferences. It systematically eliminates the confounding bias from both static and dynamic preferences to learn the conversion prediction model using historical data. We also provide a theoretical analysis to prove CAUSALMTA can learn an unbiased prediction model with sufficient data. Extensive experiments on both public datasets and the impression data in an e-commerce company show that CAUSALMTA not only achieves better prediction performance than the state-of-the-art method but also generates meaningful attribution credits across different advertising channels.

Introduction

Online advertising platforms have been widely deployed to help advertisers launch their advertisements (ads) across multiple marketing channels, such as social media, feed stream, and paid search. During the usage, the ad exposure sequences and conversion feedbacks of all customers are collected. Multi-touch attribution, short for MTA, aims to estimate each ad touchpoint's relative contribution in user conversion journeys. The attribution results will shed light on the budget allocation and automatically advertising.

Nowadays, instead of attributing the ad touchpoints by heuristic rules (Berman 2018), data-driven methods (Shao and Li 2011; Dalessandro et al. 2012; Ji and Wang 2017; Ren and etc. 2018; Arava et al. 2018; Yang, Dyer, and Wang

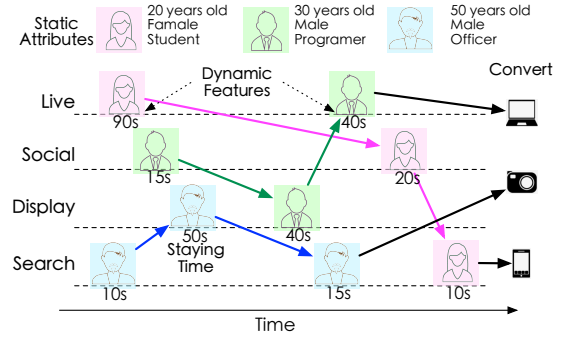


Figure 1: The influence of user preferences for MTA.

2020) which estimate the attribution credits according to the historical data have become the mainstream techniques. These methods learn a conversion prediction model with all observed historical data and then generate the counterfactual ad journeys by removing or replacing some touchpoints. Basing on some criteria, *e.g.*, Shapley value (Shapley 1953), the attribution credits can be estimated by using the prediction results of these counterfactual journeys. One essential assumption of these methods is the conversion prediction model should be unbiased, which means the model can give fair predictions on any randomly assigned journeys, including the factual and counterfactual ones. Unfortunately, this assumption does not hold in online advertising. The ad exposures are recommended according to the user preferences, leading the learned conversion prediction model to be biased. The discrepancy between observed training data and counterfactual data causes an out-of-distribution (OOD) problem in counterfactual prediction, which would harm the fairness of attribution. We define the attribution of the ad journeys with an unbiased prediction model as causal MTA.

Nevertheless, it is no trivial to eliminate the confounding bias of user preferences in MTA. The reasons are two folds: (1) Multiple confounders. The confounders in ad exposure generation consist of the static user attributes, such as genders, ages and education background, and dynamic features, *e.g.*, previously viewed ads and favorite items. Both the static and dynamic features should be taken into account for unbiased causal MTA. Existing works either focus on the static settings (Austin 2011; Johansson, Shalit, and Sontag 2016; Johansson et al. 2018; Zou et al. 2020) using IPW and

propensity score matching method for deconfounding, or are dedicated to the dynamic confounders (Lim 2018; Bica et al. 2020) learning an unbiased representation for prediction at each time step. All these works rely on the strong ignorability assumption (Pearl 2009), *i.e.*, no hidden confounders. In their settings, the static and dynamic features are hidden confounders mutually that disable the usage. (2) Delay feedback. The conversion results are observed at the end of the journey. Moreover, there is no explicit feedback available at each touchpoint. Existing sequential deconfounding methods (Xu, Xu, and Saria 2016; Roy, Lum, and Daniels 2017; Lim 2018; Bica et al. 2020) are designed for instant feedbacks, *e.g.*, the blood pressure, which can be observed immediately after taking the hypotensor. CAMTA (Kumar et al. 2020) is the most related method of our work. However, it takes the click action as the "pseudo" feedback at each touchpoint, which would involve other confounders. Above all, due to the peculiarities of advertising, there are no existed methods that can be used for unbiased causal MTA.

In this paper, we propose a novel method, namely CAUSALMTA, to mitigate the effect of user preferences-related confoundedness and achieve causal MTA. It learns an unbiased counterfactual prediction model which systematically eliminates the confounding bias from both static user attributes and dynamic features. One fundamental assumption of CAUSALMTA is that the influence of static user attributes and dynamic features are independent. This assumption is reasonable in online advertising because user attributes usually determine their item interests, and dynamic features determine how likely the users want to buy. As shown in Figure1, twenty years old students tend to be attracted by fancy phones and cosmetics, whereas the middle age guys usually like high cost-performance phones and anti-bald goods. Dynamic features, such as previously visited ads and staying time, reflect the purchase intention. The main contributions can be summarized as follows:

- We decomposed the confounding bias of user preferences into static user attributes and dynamic features, and defined the causal MTA problem.
- We propose the first method CAUSALMTA for causal MTA, which is provable for eliminating the confounding bias of user preferences in counterfactual prediction.
- Experiments on both opensource datasets and real-world impressing data of mobile phones shops from an e-commerce company show the superior of CAUSALMTA.

Related Work

Data-driven multi-touch attribution. Previously, marketers have applied simple rules, *e.g.*, the last touch, to attribute the influence of touched ads (Berman 2018), which either ignore the effects of other channels or neglect the channel difference. To overcome these drawbacks, researchers proposed data-driven attribution methods. The data-driven MTA model was first proposed in (Shao and Li 2011), and has been combined with survival analysis (Zhang, Wei, and Ren 2014) and hazard rate (Ji and Wang 2017) to reflect the influence of ad exposure. However,

the data-driven methods mentioned above neglect the customers' features and cannot directly allocate personalized attribution. Besides, the temporal dependency and dynamic interaction between channels need to be modeled. Recently, many DNN-based data-driven MTA methods have been proposed to address the issues, such as channel interaction, time dependency, user characteristics. Some literature (Arava et al. 2018; Ren and etc. 2018; Du and etc. 2019; Kumar et al. 2020; Yang, Dyer, and Wang 2020) leverage RNNs to model longitudinal data. DNAMTA (Arava et al. 2018) is an LSTM based deep sequential model which captures the touchpoint contextual dependency via attention mechanism and incorporates user context information and survival time-decay functions. DARNN *et al.* (Ren and etc. 2018) is a dual attention model that combines post-view and post-click attribution patterns for final conversion estimation.

Counterfactual Prediction. Positioned as a causal estimation problem by (Dalessandro et al. 2012), the calculation of attribution credits is actually based on counterfactual estimation (Zhang, Wei, and Ren 2014; Du and etc. 2019; Singal and etc. 2019; Shender et al. 2020). A limitation of the models mentioned above is the lack of exogenous variation in user exposure to advertising, which hazards the reliability of the attribution results as the training data of the counterfactual predictor is biased by confounders. Albeit extant papers (Sahni 2015; Barajas et al. 2016; Nair et al. 2017; Zantedeschi, Feit, and Bradlow 2017) mitigate the issue with full or quasi-randomization, the cost and complexity of such randomization restrict the number of users and ad-types. The idea of calibrating the conversion prediction model in MTA by removing confounding bias is inspired by works in counterfactual prediction. There is a large number of methods for counterfactual prediction using observational data in the static setting, involving utilizing propensity score matching (Austin 2011), learning unbiased representation for prediction (Johansson, Shalit, and Sontag 2016; Johansson et al. 2018; Zou et al. 2020), conducting propensity-aware hyperparameter tuning (Alaa and van der Schaar 2017, 2018). For estimating the effects of time-varying treatments in the area such as epidemiology where the treatments have instant feedback, many approaches (Xu, Xu, and Saria 2016; Roy, Lum, and Daniels 2017; Lim 2018; Bica et al. 2020) addressing the longitudinal setting are proposed. Because of the gap between the longitudinal data in epidemiology and ad journeys in MTA, those methods cannot be directly used in our task.

Preliminary

Problem Definition. We consider an ad exposure dataset \mathcal{D} which consists of N conversion journeys of U users. Each journey can be formulated as a triplet, *i.e.*, $(\mathbf{u}^i, \mathbf{J}^i, \mathbf{y}^i)$. \mathbf{u}^i stands for the static user attributes, which is unlikely to be changed during the user conversion journey. \mathbf{J}^i is a sequence of touchpoints, *i.e.*, $\{\mathbf{p}_t^i\}_{t=1}^{T^i}$. Each touchpoint $\mathbf{p}_t^i = (\mathbf{c}_t^i, \mathbf{f}_t^i)$ contains a channel index \mathbf{c}_t^i and a feature vector \mathbf{z}_t^i . Specifically, $\mathbf{c}_t^i \in \{\mathbf{c}_1, \dots, \mathbf{c}_k, \dots, \mathbf{c}_K\}$ indicates the exposed channels, where K is the number of ad channels. The feature vector \mathbf{f}_t^i includes the dynamic side information of this touch-

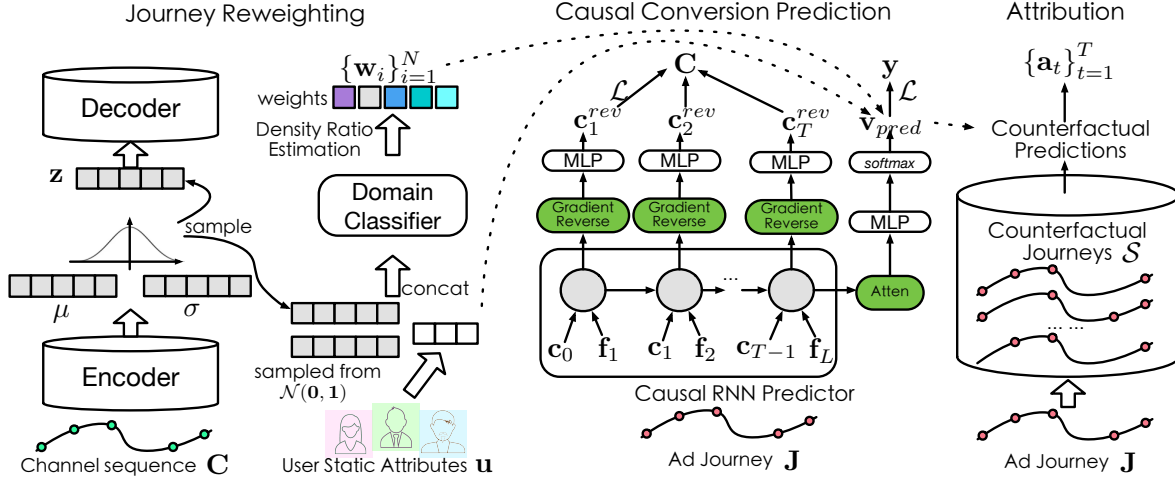


Figure 2: Architecture of CAUSALMTA.

point, *e.g.*, advertising form and staying time. y^i is a binary indicator that records whether the journey leads to a conversion event or not. The goal of MTA is to model the sequential pattern and assign the attribution credits to all the touch-points \mathbf{p}_t^i according to the whole information in \mathcal{D} . Nevertheless, historical data in \mathcal{D} often exhibits confounding bias due to user preferences, which could be a fatal challenge for estimating the attribution credits. The choice of the channel at a touch-point is likely to be influenced by multiple factors like user attributes and previously visited goods. Causal multi-touch attribution aims to estimate trustworthy attribution credits $\{\mathbf{p}_t^i\}_{t=1}^{T^i}$ of all touchpoints in \mathbf{J}^i .

Method Overview. As shown in Figure 2, CAUSALMTA is a novel model-agnostic framework consisting of two key modules, *i.e.*, journey reweighting and causal conversion prediction, which mitigate the confounding bias of user static attributes and dynamic features respectively. In journey reweighting, we employ the Variational Recurrent Auto-encoders (VRAE) to learn the generation probabilities of pure channels journeys, and conduct user demographic-based density estimation to calculate the likelihoods of the channels being randomly assigned that is used for weights computation. For causal conversion prediction, CAUSALMTA utilizes RNNs to model the dynamic features of journeys. A gradient reverse layer is building upon the outputs of each time step to ensure the model is unable to predict the next ad channel. It derives balancing representation, which removes the association between dynamic features and the ad exposure. Basing on the learned weights of journey reweighting, the last hidden output is trained to estimate the conversion probability. After that, we can obtain an unbiased prediction model. Lastly, with the constructed counterfactual journeys, the attribution credits can be estimated under Shapley value measure.

Methodology

In this section, we first specify the journey reweighting and causal conversion prediction respectively. After that, the calculation of attribution credits is detailed. In the end, we pro-

vide the theoretical analysis of CAUSALMTA.

Journey Reweighting

To mitigate the bias of static user attributes, the journey reweighting module takes pure channel sequences in \mathcal{D} as the input and estimates the sample weights of the prediction model according to how likely the journey be generated randomly. It consists of two procedures, *i.e.*, generation model for channel sequences and weights estimation of journeys.

Generation Model for Channel Sequences. We utilize VRAE (Variational Recurrent Auto-encoders) (Fabius, van Amersfoort, and Kingma 2015) to model the generation of channel sequences. When training data is large enough, the distribution of pure channel sequences tends to be random without considering user preferences. In this setting, the learned VRAE is capable of generating unbiased predictions of observed channel sequences.

For each ad journey $(\mathbf{u}, \mathbf{J}, y)$ in \mathcal{D} , we only concern with the channel information in this procedure and extract the pure channel sequence $\mathbf{C} = \{\mathbf{c}_t\}_{t=1}^T$. Taking \mathbf{C} as the input, CAUSALMTA employs the channel embedding affiliated with LSTM as the encoder and utilizes the final hidden state to generate the distribution over latent representation:

$$\begin{aligned} \{\mathbf{h}_t\}_{t=1}^T &= \text{LSTM}_{\text{enc}}(\mathbf{C}, \mathbf{h}_0) \\ \mu_z &= W_\mu \mathbf{h}_T + \mathbf{b}_\mu \\ \log(\sigma_z) &= W_\sigma \mathbf{h}_T + \mathbf{b}_\sigma \end{aligned}$$

where \mathbf{h}_0 is the initial hidden state of the encoder. Leveraging the reparametrization trick, we sample a vector \mathbf{z} from the distribution to initialize the hidden state of the decoder:

$$\begin{aligned} \mathbf{h}'_0 &= \tanh(W_z^T \mathbf{z} + \mathbf{b}_z) \\ \{\mathbf{h}'_t\}_{t=1}^T &= \text{LSTM}_{\text{dec}}(\mathbf{C}_{\text{out}}, \mathbf{h}'_0) \\ \mathbf{c}'_t &= \text{softmax}(W_o \mathbf{h}'_t + \mathbf{b}_o) \end{aligned}$$

where \mathbf{h}'_0 is the initial hidden state of the decoder; \mathbf{C}_{out} is the feed previous input which takes the output of previous step as the input; \mathbf{c}'_t is the decoded channel sequence.

The loss function is composed of two parts: 1) the reconstruction loss which is defined as the cross-entropy between \mathbf{c}_t and \mathbf{c}'_t . 2) the KL divergence between the posterior and prior distribution over the latent variable:

$$\mathcal{L}_w = \alpha \sum_{i=1}^N \sum_{t=1}^{T_i} CE(\mathbf{c}_t, \mathbf{c}'_t) + \beta D_{KL}(q_\phi(\mathbf{z}) || p_\theta(\mathbf{z})) \quad (1)$$

where $p_\theta(\mathbf{z})$ is the prior distribution usually assumed to be a standard normal distribution $\mathcal{N}(0, \mathbf{I})$; $q_\phi(\mathbf{z}|\mathbf{c}^i)$ is the posterior approximation ($\mathcal{N}(\mu^i, (\sigma^i)^2)$); α and β are hyperparameters that control the importance each parts.

Weights Estimation for Ad Journeys. To eliminate the bias of user static features, we estimate the weights of observed journeys. The journeys approximating to randomly assigned have higher weights in conversion prediction training than those are severely affected by user preferences. Formally, the learned weights $W_T(\mathbf{u}, \mathbf{C})$ should be subject to $W_T(\mathbf{u}, \mathbf{C}) = p(\mathbf{C})/p(\mathbf{C}|\mathbf{u})$ (Fong et al. 2018; Zou et al. 2020). When we learn a variational distribution $q_\phi(\mathbf{z}|\mathbf{c})$, the variational sample weights can be computed as follows:

$$w^i = W_T(\mathbf{u}^i, \mathbf{c}^i) = \{\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{c}^i)} [\frac{1}{W_z(\mathbf{u}^i, \mathbf{z})}]\}^{-1} \quad (2)$$

where $W_z(\mathbf{u}, \mathbf{z})$ can be viewed as the density ratio estimation to decorrelate \mathbf{u} and \mathbf{z} for points in space $\square \times \mathcal{Z}$. The detailed proof can be found in the appendix.

In CAUSALMTA, we design a binary domain classifier to help estimate $W_z(\mathbf{u}, \mathbf{z})$. Training data of the classifier is generated cooperating with the encoder of VRAE. We label static user attributes with real latent representation $\{(\mathbf{u}^i, \mathbf{z})\}_{1 \leq i \leq N}$, $\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{c}^i)$ as positive ones, whereas samples with latent representation sampled from standard normal distribution $\{(\mathbf{u}^i, \mathbf{z})\}_{1 \leq i \leq N}$, $\mathbf{z} \sim p_\theta(\mathbf{z})$ as negative ones. We first embed the user attributes into latent vectors and train a domain classifier to fit these samples:

$$\begin{aligned} \mathbf{e}_u &= \text{Embedding}(\mathbf{u}) \\ \mathbf{x} &= \text{concat}(\mathbf{e}_u, \mathbf{z}) \\ p_{\theta_d}(L|\mathbf{u}, \mathbf{z}) &= \text{sigmoid}(\text{MLP}(\mathbf{x})) \end{aligned}$$

Now that we have $p(L=0) = p(L=1)$, the density ratio estimation $W_z(\mathbf{u}, \mathbf{z})$ can be conducted as follows:

$$W_z(\mathbf{u}, \mathbf{z}) = \frac{p(\mathbf{u}, \mathbf{z}|L=0)}{p(\mathbf{u}, \mathbf{z}|L=1)} = \frac{p(L=0|\mathbf{u}, \mathbf{z})}{p(L=1|\mathbf{u}, \mathbf{z})} \quad (3)$$

Using this formula, we can obtain the weights of all journeys, i.e., $\{\mathbf{w}_i\}_{i=1}^N$.

Causal Conversion Prediction

After the generation of sample weights, CAUSALMTA utilizes them to train a trustworthy conversion prediction model. Besides the static attributes, the biases of prediction are also caused by dynamic user features. To mitigate them, we borrow the idea from CRN (Bica et al. 2020) and involve a gradient reverse layer to learn balancing representation. Due to the delay feedback problem, we refine the structure of CRN to make it suitable for MTA.

Formally, we first reorganize the dataset. For each journey $(\mathbf{u}, \mathbf{J}, y)$, we adopt one step offset on the channel sequence and fill the blank position with a unified placeholder, i.e., $\mathbf{C}_+ = \{\mathbf{c}_0, \mathbf{c}_t\}_{t=1}^{T-1}$. CAUSALMTA takes \mathbf{C}_+ along with other dynamic features $\mathbf{F} = \{\mathbf{f}_t\}_{t=1}^T$ as the input and employs LSTM with the attention mechanism to obtain the trustworthy prediction:

$$\begin{aligned} \mathbf{e}_u, \mathbf{e}_{C_+}, \mathbf{e}_F &= \text{Embedding}(\mathbf{u}, \mathbf{C}_+, \mathbf{F}) \\ \mathbf{v}_{in} &= \text{concat}(\mathbf{e}_{C_+}, \mathbf{e}_F) \\ \{\mathbf{out}_t\}_{t=1}^T &= \text{LSTM}_{\text{pred}}(\mathbf{v}_{in}, \mathbf{h}_0) \end{aligned}$$

where $\mathbf{e}_{C_+}, \mathbf{e}_F, \mathbf{v}_{in}$ are the sequences of latent vectors. Once the output vectors are generated, we adopt them for two parallel processes. One for eliminating the bias of dynamic features, and the other for conversion prediction.

$$\begin{aligned} \{\mathbf{v}_t^{rev}\}_{t=1}^T &= \text{MLP}(\text{GRL}(\{\mathbf{out}_t\}_{t=1}^T)) \\ \{\mathbf{c}_t^{rev}\}_{t=1}^T &= \text{softmax}(\{\mathbf{v}_t^{rev}\}_{t=1}^T) \\ \mathbf{v}_{attn} &= \text{Attention}(\mathbf{out}_T, \{\mathbf{out}_t\}_{t=1}^T) \\ \mathbf{v}_{pred} &= \text{softmax}(\text{MLP}(\mathbf{v}_{attn})) \end{aligned}$$

where GRL is the gradient reverse layer that ensures \mathbf{out}_t can not predict \mathbf{c}_t . The loss function of causal conversion prediction consists of two parts, i.e., reverse channel prediction and conversion prediction:

$$\mathcal{L}_p = \gamma \sum_{i=1}^N \sum_{t=1}^{T_i} CE(\mathbf{c}_t^{rev}, \mathbf{c}_t) + \delta \sum_{i=1}^N \mathbf{w}_i \cdot CE(\mathbf{v}_{pred}^i, y^i) \quad (4)$$

where CE is the cross-entropy loss; γ and δ are hyperparameters; \mathbf{w}_i is the learned journey weights. With the well-trained conversion prediction model, we can calculate the attribution credits of each touchpoint by constructing some counterfactual journeys.

Attribution Credits Calculation

CAUSALMTA computes Shapley values (Shapley 1953) for ad credits allocation. Based on assessing the marginal contribution of each player in the game, the Shapley value method is a general credit distribution method, and it has been widely used in MTA (Singal and etc. 2019; Yang, Dyer, and Wang 2020) due to its advantage of having an axiomatic foundation and catering to fairness consideration.

Formally, let $\mathbf{J}^i \setminus \{\mathbf{p}_t^i\}$ denote the counterfactual ad journey by removing touchpoint \mathbf{p}_t^i . \mathcal{S} can be viewed as a subsequence of the counterfactual ad journey $\mathbf{J}^i \setminus \{\mathbf{p}_t^i\}$. If we denote the result of causal conversion prediction for channel sequence \mathbf{J}^i as $p(\mathbf{J}^i)$, the Shapley values for ad exposure $\{\mathbf{c}_t^i\}$ can be defined as $SV_t^i = \sum_{\mathcal{S} \subseteq \mathbf{J}^i \setminus \{\mathbf{p}_t^i\}} \frac{|\mathcal{S}|!(|\mathbf{J}^i| - |\mathcal{S}| - 1)!}{|\mathbf{J}^i|!} [p(\mathcal{S} \cup \{\mathbf{p}_t^i\}) - p(\mathcal{S})]$ where $|\mathbf{C}^i|, |\mathbf{S}|$ are the cardinalities of these sets. If SV_t^i is negative, we set it zero. Then we normalize all incremental scores for each ad exposure as follows,

$$\mathbf{a}_t^i = \sigma(SV_t^i) / \sum_{t=1}^{T^i} \sigma(SV_t^i) \quad (5)$$

Algorithm 1: Learning procedure of CAUSALMTA.

Input:
The ad exposure dataset \mathcal{D} , i.e., $\{(\mathbf{u}^i, \mathbf{J}^i, y^i)\}_{i=1}^N$;

Output:
Attribution credits $\{\alpha_t^i\}_{t=1}^{T_i}$ for touchpoints $\{\mathbf{a}_t^i\}_{t=1}^{T_i}$.

- 1: # Generation model for channel sequences
- 2: **for** each journey $(\mathbf{u}^i, \mathbf{J}^i, y^i)$ in \mathcal{D} **do**
- 3: Evaluate \mathcal{L}_w according to Eq.(1) and update the parameters of VRAE model.
- 4: **end for**
- 5: # Weights estimation for ad journeys
- 6: **for** each ad journey $(\mathbf{u}^i, \mathbf{J}^i, y^i)$ in \mathcal{D} **do**
- 7: Generate latent representation for positive and negative samples respectively.
- 8: Optimize the parameters of domain classifier.
- 9: **end for**
- 10: Conduct density ratio estimation and calculate sample weights \mathbf{w}^i according to Eq.(2) and Eq.(3).
- 11: # Causal conversion prediction
- 12: **for** each ad journey $(\mathbf{u}^i, \mathbf{J}^i, y^i)$ in \mathcal{D} **do**
- 13: Evaluate \mathcal{L}_p according to Eq.(4) and update the parameters of conversion prediction model.
- 14: **end for**
- 15: # Calculation of Attribution Credits
- 16: **for** each ad journey $(\mathbf{u}^i, \mathbf{J}^i, y^i)$ in \mathcal{D} **do**
- 17: **for** each touchpoint \mathbf{c}_t^i in channel sequence \mathbf{C}^i **do**
- 18: Compute SV_t^i according to its definition.
- 19: Calculate attribution credits \mathbf{a}_t^i according to Eq.(5).
- 20: **end for**
- 21: **end for**
- 22: **return** $\{\{\mathbf{a}_t^i\}_{t=1}^{T_i}\}_{i=1}^N$

where $\sigma(x) = \max(0, x)$ and \mathbf{a}_t^i are the attribution credits of the corresponding ad exposures. The pseudo-code of CAUSALMTA is shown in Algorithm 1.

Theoretical Analysis of CAUSALMTA

Under the assumption of independence, we can decompose the overall confounding bias \mathcal{B} into the bias introduced by user static attributes \mathcal{B}_u and the bias introduced by dynamic user features \mathcal{B}_F , i.e., $\mathcal{B} = \mathcal{B}_u + \mathcal{B}_F$. CAUSALMTA aims to obtain an unbiased prediction model and achieve $\mathcal{B} = 0$.

We prove that the confounding bias from static user attributes \mathcal{B}_u can be mitigated by estimating sample weights \mathbf{w}^i for ad journeys. Formally, let \mathcal{E}_{cf} denote the counterfactual prediction error, which is the target to be minimized. Unfortunately, \mathcal{E}_{cf} can not be directly measured on the observational dataset. We can derive the upper bound of \mathcal{E}_{cf} , which is given by

$$\mathcal{B}_u = \mathcal{E}_{cf} - \mathcal{E}_f^w \leq IPM_G(W_T(\mathbf{u}, \mathbf{C})p(\mathbf{u}, \mathbf{C}), p(\mathbf{u})p(\mathbf{C}))$$

where \mathcal{E}_f^w is the prediction error on the re-weighted data and IPM denotes Integral Probability Metric. When $W_T(\mathbf{u}, \mathbf{C}) = \frac{p(\mathbf{C})}{p(\mathbf{C}|\mathbf{u})}$, the equation $\mathcal{E}_{cf} = \mathcal{E}_f^w$ can be proved. More details of the proof are available in the appendix.

In dynamic settings, \mathcal{B}_F equals zero if we can prove that the learned representation removes the association between dynamic features and the ad exposure. We build

Table 1: The overview of the Criteo dataset

Statistics	Raw	Processed
No. of users	6,142,256	157,331
No. of campaigns	675	10
No. of journeys	6,514,319	196,560
No. of convert journeys	435,810	19,890
No. of touchpoints	16,468,027	787,483

the representation \mathbf{v}_t^{rev} invariant across different ad channels to eliminate biases caused by dynamic user features. We achieve this by minimizing the formula $\mathcal{L}_{rev} = \sum_{i=1}^N \sum_{t=1}^{T_i} CE(\mathbf{c}_t^{rev}, \mathbf{c}_t)$ in Eq.(4). We can prove that

$$\mathcal{L}_{rev} = K \cdot JSD(p(\mathbf{v}_t^{rev}|\mathbf{c}_1), \dots, p(\mathbf{v}_t^{rev}|\mathbf{c}_K)) - K \log K$$

where $K \log K$ is a constant, and $JSD(\cdot, \dots, \cdot)$ denotes the multi-distribution Jensen-Shannon Divergence (Li et al. 2018), which is non-negative and 0 if and only if all distributions are equal. To minimize \mathcal{L}_{rev} , we derive $p(\mathbf{v}_t^{rev}|\mathbf{c}_1) = \dots = p(\mathbf{v}_t^{rev}|\mathbf{c}_K)$, where \mathbf{v}_t^{rev} is the learned representation invariant across different ad channels. For details, see the appendix.

Experiment

In this section, we evaluate the performance of CAUSALMTA and answer the following questions:

- **Q1:** Does CAUSALMTA perform better than the state-of-the-art MTA methods on conversion prediction?
- **Q2:** What are the capabilities of the journey reweighting module and the causal conversion prediction module?
- **Q3:** Does CAUSALMTA work well on real ad impression dataset?

Experimental Settings

A conversion prediction task is employed to examine the performance of CAUSALMTA. In this section, we briefly introduce the data description, evaluation metrics, compared baselines, and hyperparameter settings. More details of this part can be found in the appendix.

Data Descriptions. The performance of CAUSALMTA is evaluated on two datasets, i.e., **Criteo** and **Ad Impression of an E-commerce company**. **Criteo**¹ is a public dataset on ad bidding and be widely used in MTA (Diemert et al. 2017; Ren and etc. 2018; Kumar et al. 2020). Following the same experimental setting of CAMTA(Kumar et al. 2020), we choose top-10 highly exposed channels and journeys containing more than 3 touchpoints. **Ad Impression of an E-commerce Company** contains the ad impression data of mobile phone shops in 30 days. These touchpoints are categorized into 40 channels, including interact, feed, display, search, live show, etc.

Evaluation Metrics. For conversion prediction, we evaluate the performance in terms of **log-loss**, **AUC**. For fairness, the log-loss only contains the conversion prediction

¹<http://ailab.criteo.com/criteo-attribution-modeling-bidding-dataset/>

Table 2: Results of conversion prediction on Criteo dataset. SL, DL, CL in the left column indicate the statistical learning-based methods, deep learning-based methods and causal learning-based methods, respectively.

	Method	AUC	Log-loss
SL	LR	0.8370 ± 0.03	0.191 ± 0.01
	SP	0.5637 ± 0.02	0.249 ± 0.015
	AH	0.5203 ± 0.03	0.262 ± 0.01
DL	DNAMTA	0.9127 ± 0.02	0.1360 ± 0.013
	DARNN	0.8726 ± 0.02	0.165 ± 0.006
	DeepMTA	0.9104 ± 0.03	0.112 ± 0.012
CL	JDMTA	0.9127 ± 0.01	0.0838 ± 0.007
	CAMTA	0.9347 ± 0.02	0.0715 ± 0.007
Ours	CAUSALMTA	0.9659 ± 0.01	0.0517 ± 0.003
	CM-RW	0.9539 ± 0.01	0.0560 ± 0.003
	CM-CAUSAL	0.9517 ± 0.01	0.0534 ± 0.002

part of Equation 4. It can be reckoned as a standard measurement to estimate the classification performance. AUC can be a metric reflecting the pairwise ranking performance of the estimation between converted and non-converted ad impression sequences.

Compared Methods. In our experiments, CAUSALMTA is compared with 8 baseline methods which can be divided into three categories, *i.e.*, statistical learning-based methods, deep learning-based methods, and causal learning-based methods. The statistical learning-based methods consist of three methods, *i.e.*, Logistic Regression (Shao and Li 2011)(**LR**), Simple Probabilistic (Dalessandro et al. 2012)(**SP**), and Additive Hazard (Zhang, Wei, and Ren 2014)(**AH**). Deep learning-based methods contain three methods, *i.e.*, **DNAMTA** (Arava et al. 2018), **DARNN** (Ren and etc. 2018), and **DeepMTA** (Yang, Dyer, and Wang 2020). The causal learning-based methods also have two works, *i.e.*, **JDMTA** (Du and etc. 2019) and **CAMTA** (Kumar et al. 2020). Besides, we also compare CAUSALMTA with two ablation methods, *i.e.*, CM-RW and CM-CAUSAL. Detailed descriptions of these methods are available in the appendix.

Parameter Settings. For LSTMs in CAUSALMTA, we stack three 3 layer LSTMs as the encoder, decoder, and the predictor respectively. MLP models in CAUSALMTA are composed of 4 fully connected layers with *ELU* as the activate function. CAUSALMTA has 4 hyperparameters *i.e.*, α, β, γ and δ , we empirically set $\alpha = \beta = 0.5$, and $\gamma = \delta = 0.5$. All the experiments are conducted on a high-end server with $2 \times$ NVIDIA GTX3090 GPUs. All the compared baselines are trained in 30 epochs and the best model is chosen to report.

Performance Comparison

The detailed evaluation results on different baselines are presented in Table 2. As shown, CAUSALMTA continuously outperforms all the compared baselines, which proves the validity of eliminating the confounding bias on static user attributes and dynamic features. CAMTA is the strongest baseline but also inferior to CAUSALMTA. It utilizes click labels as the auxiliary information, which probably involves

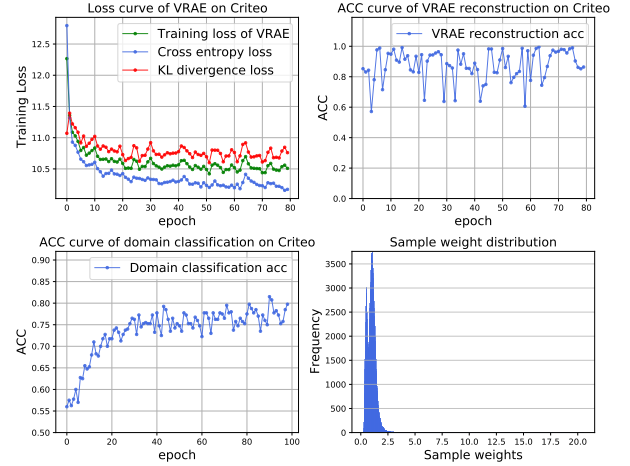


Figure 3: Learning curves and sample weight distribution on Criteo dataset.

additional confounding bias. Moreover, CAMTA does not consider the difference between static and dynamic confounders, which would also harm the performance. One interesting phenomenon is CAUSALMTA has a more stable confidence interval compared to other baselines. It indicates that the parameters in CAUSALMTA tend to converge to similar values with different initialization. To a certain extent, CAUSALMTA is more robust than other baselines.

Comparing the performance of different categories of methods, we can observe that SL methods are inferior to the other two categories. SL methods either use statistical laws or employ logistic regression to predict the conversion probability, which can not well model the conversion process. DL methods perform better than the SL methods but are also inferior to the CL methods. It proves that the deep learning techniques are more suitable for conversion prediction due to their large parameter space and high ability to model complex tasks. However, these methods have poor performance compared to the CL methods. It is because deep learning methods directly use the observed data to train the prediction models, which are incapable of handling confounding bias and would suffer from the out-of-distribution problem. CL methods outperform other baselines with a large margin, which demonstrates the prediction performance highly increased by eliminating the confounding bias.

Ablation Studies

To explore the effectiveness of the journey reweighting and causal conversion prediction, we compare CAUSALMTA with two ablation methods, *i.e.*, CM-RW and CM-CAUSAL, which remove the reweighting procedure and the gradient reverse layer respectively. We first show the intermediate results of journey reweighting. The reconstruction accuracy of VRAE and the classification accuracy of the domain classifier directly influence the performance of CAUSALMTA. As shown in Figure 3, the reconstruction accuracy is approximate to 98%, and the classification accuracy is approximate to 82%, which indicates that the VRAE and domain classifier are well trained, and the results of journey reweighting are significant. We also witness the reconstruction accuracy

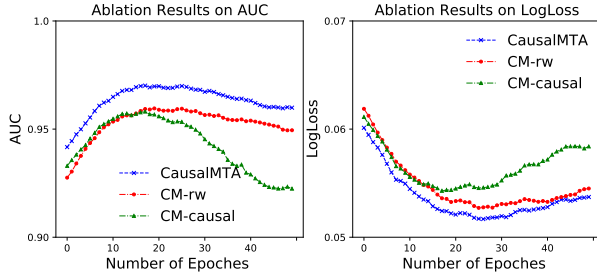


Figure 4: Performance comparison of ablation methods.

fluctuates at a high level as KL divergence dominates the training loss when the cross-entropy loss is small enough.

We summarize the metrics of ablations in Table 2 and illustrate the training procedure in Figure 5. As shown, the AUCs of CM-rw and CM-causal are inferior to CAUSALMTA, which proves that both the journey reweighting and causal conversion prediction help improve the performance. By removing the journey reweighting model, the performance of CAUSALMTA decreases from 0.9659 to 0.9539. By removing the gradient reverse layer, the performance of CAUSALMTA decreases from 0.9659 to 0.9617. We can observe the improvement of gradient reverse layer(causal prediction) is more significant than journey reweighting, which indicates the confounding bias of dynamic feature are more obvious than the static user attributes. This result is consistent with our cognition. Moreover, as illustrated in Figure 5, the convergence speed of CAUSALMTA is faster than CM-rw and CM-causal, which shows the superiority of CAUSALMTA in eliminating the confounding bias of user preferences.

Empirical Applications in an E-commerce Company

We evaluate the performance of CAUSALMTA on real applications in this section. In application, channel attributions of each shop are more meaningful to guide the budget allocation. We train the attribution models in the first 15 days and use the last data for testing. We choose all of its converted journeys in the test set for one specific shop and compute the mean credits of 40 channels. After that, we employ two experiments, *i.e.*, attribution improvement and offline data replay, to evaluate the performance of CAUSALMTA.

Attribution Improvement. We compute the attribution credits of two representative cellphone shops utilizing CAUSALMTA and compare them with the credits calculated by an LSTM-based conversion prediction model. The comparison results are illustrated in Figure 5. As shown, the credits of search channels on both shops are decreased, indicating that the estimated contribution of search ads is reduced after eliminating the confounding bias of user preferences. This is because user tends to search the goods before paying. The attributions of the search channel are usually overestimated, and CAUSALMTA can mitigate this kind of bias. For different shops, the improvements are consistent with its budget allocation. After examining the total budget on each channel, we found Shop 1 spent more money on Live and Shop 2 spent more money on Display, which coincides with our attribution result.

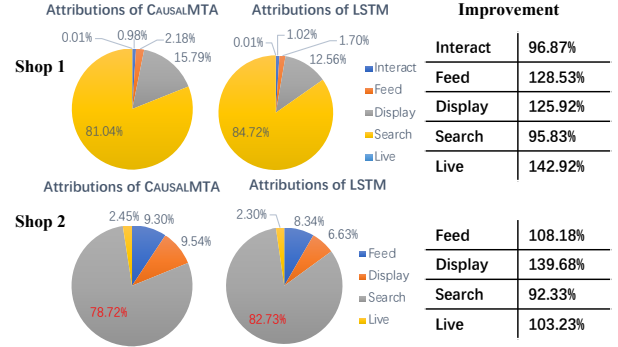


Figure 5: The comparison of attribution changes for two cellphone shops.

Table 3: Profit comparison of data replay.

Method	1/2	1/4	1/8	1/16
Shop 1				
LSTM	69.8	63.7	56.2	58.3
CAUSALMTA	72.3	70.2	57.1	59.2
Improvement	+3.58%	+10.20%	+1.60%	+1.54%
Shop 2				
LSTM	20.9	18.3	17.5	15.2
CAUSALMTA	21.1	19.1	17.8	15.8
Improvement	+0.96%	+4.37%	+1.71%	+3.95%

Offline Data Replay. In this experiment, we employ the attribution credits to guide the budget allocation. Based on the attribution credits, we first compute the return-on-investment (ROI) of each channel and utilize the normalized weights of ROI as budget proportion (Ren and etc. 2018). Assuming that the total cost of the test set is $cost_t$, we set the evaluation budgets as $1/2, 1/4, 1/8, 1/16$ of $cost_t$ and replay the historical data to select journeys satisfying evaluation budgets. Table 3 shows the comparison results of the profit in each evaluation budget. We can observe that the profit CAUSALMTA consistently outperforms the LSTM-based predictor on all evaluation budgets, which indicates that the attribution credits of CAUSALMTA reflect the causal relationships in advertising. It can be used to guide budget allocation and achieve better profit.

Conclusion

In this paper, we define the problem of causal MTA, which eliminates the confounding bias introduced by user preferences and assigns the attribution credits fairly over all touchpoints. To this end, we propose CAUSALMTA, which composes the confounding bias of user preferences into two independent parts, *i.e.*, the static user attributes and the dynamic features. CAUSALMTA employs journey reweighting and causal conversion prediction to solve these two kinds of confounding bias respectively. We prove that CAUSALMTA is capable of generating unbiased conversion predictions of ad journey. Extensive experiments on the public dataset and real commercial data from an e-commerce company show that CAUSALMTA outperforms all the compared baselines and works well in the real-world application.

References

- Alaa, A. M.; and van der Schaar, M. 2017. Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 3424–3432.
- Alaa, A. M.; and van der Schaar, M. 2018. Limits of Estimating Heterogeneous Treatment Effects: Guidelines for Practical Algorithm Design. In Dy, J. G.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, 129–138. PMLR.
- Arava, S. K.; Dong, C.; Yan, Z.; Pani, A.; et al. 2018. Deep neural net with attention for multi-channel multi-touch attribution. *arXiv preprint arXiv:1809.02230*.
- Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate behavioral research*, 46(3): 399–424.
- Barajas, J.; Akella, R.; Holtan, M.; and Flores, A. 2016. Experimental designs and estimation for online display advertising attribution in marketplaces. *Marketing Science*, 35(3): 465–483.
- Berman, R. 2018. Beyond the last touch: Attribution in online advertising. *Marketing Science*, 37(5): 771–792.
- Bica, I.; Alaa, A. M.; Jordon, J.; and van der Schaar, M. 2020. Estimating counterfactual treatment outcomes over time through adversarially balanced representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Dalessandro, B.; Perlich, C.; Stitelman, O.; and Provost, F. 2012. Causally motivated attribution for online advertising. In *Proceedings of the sixth international workshop on data mining for online advertising and internet economy*, 1–9.
- Diemert, E.; Meynet, J.; Galland, P.; and Lefortier, D. 2017. Attribution Modeling Increases Efficiency of Bidding in Display Advertising. In *Proceedings of the ADKDD’17, Halifax, NS, Canada, August 13 - 17, 2017*, 2:1–2:6. ACM.
- Du, R.; and etc. 2019. Causally driven incremental multi touch attribution using a recurrent neural network. *arXiv:1902.00215*.
- Fabius, O.; van Amersfoort, J. R.; and Kingma, D. P. 2015. Variational Recurrent Auto-Encoders. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*.
- Fong, C.; Hazlett, C.; Imai, K.; et al. 2018. Covariate balancing propensity score for a continuous treatment: application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12(1): 156–177.
- Ji, W.; and Wang, X. 2017. Additional multi-touch attribution for online advertising. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.
- Johansson, F.; Shalit, U.; and Sontag, D. 2016. Learning representations for counterfactual inference. In *International conference on machine learning*, 3020–3029. PMLR.
- Johansson, F. D.; Kallus, N.; Shalit, U.; and Sontag, D. 2018. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*.
- Kumar, S.; Gupta, G.; Prasad, R.; Chatterjee, A.; Vig, L.; and Shroff, G. 2020. CAMTA: Causal Attention Model for Multi-touch Attribution. In Fatta, G. D.; Sheng, V. S.; Cuzocrea, A.; Zaniolo, C.; and Wu, X., eds., *20th International Conference on Data Mining Workshops, ICDM Workshops 2020, Sorrento, Italy, November 17-20, 2020*, 79–86. IEEE.
- Li, Y.; Tian, X.; Gong, M.; Liu, Y.; Liu, T.; Zhang, K.; and Tao, D. 2018. Deep Domain Generalization via Conditional Invariant Adversarial Networks. In Ferrari, V.; Hebert, M.; Sminchisescu, C.; and Weiss, Y., eds., *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV*, volume 11219 of *Lecture Notes in Computer Science*, 647–663. Springer.
- Lim, B. 2018. Forecasting Treatment Responses Over Time Using Recurrent Marginal Structural Networks. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 7494–7504.
- Nair, H. S.; Misra, S.; Hornbuckle IV, W. J.; Mishra, R.; and Acharya, A. 2017. Big data and marketing analytics in gaming: Combining empirical models and field experimentation. *Marketing Science*, 36(5): 699–725.
- Pearl, J. 2009. *Causality*. Cambridge university press.
- Ren, K.; and etc. 2018. Learning Multi-touch Conversion Attribution with Dual-attention Mechanisms for Online Advertising. In *CIKM 2018*, 1433–1442. ACM.
- Roy, J.; Lum, K. J.; and Daniels, M. J. 2017. A Bayesian nonparametric approach to marginal structural models for point treatments and a continuous or survival outcome. *Biostatistics*, 18(1): 32–47.
- Sahni, N. S. 2015. Effect of temporal spacing between advertising exposures: Evidence from online field experiments. *Quantitative Marketing and Economics*, 13(3): 203–247.
- Shao, X.; and Li, L. 2011. Data-driven multi-touch attribution models. In Apté, C.; Ghosh, J.; and Smyth, P., eds., *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*, 258–264. ACM.
- Shapley, L. S. 1953. A value for n-person games. *Contributions to the Theory of Games*, 2(28): 307–317.
- Shender, D.; Amini, A. N.; Bao, X.; Dikmen, M.; Richardson, A.; and Wang, J. 2020. A Time To Event Framework For Multi-touch Attribution. *arXiv preprint arXiv:2009.08432*.

- Singal, R.; and etc. 2019. Shapley meets uniform: An axiomatic framework for attribution in online advertising. In *WWW*, 1713–1723.
- Xu, Y.; Xu, Y.; and Saria, S. 2016. A Bayesian nonparametric approach for estimating individualized treatment-response curves. In *Machine Learning for Healthcare Conference*, 282–300. PMLR.
- Yang, D.; Dyer, K.; and Wang, S. 2020. Interpretable Deep Learning Model for Online Multi-touch Attribution. *arXiv:2004.00384*.
- Zantedeschi, D.; Feit, E. M.; and Bradlow, E. T. 2017. Measuring multichannel advertising response. *Management Science*, 63(8): 2706–2728.
- Zhang, Y.; Wei, Y.; and Ren, J. 2014. Multi-touch attribution in online advertising with survival theory. In *2014 IEEE International Conference on Data Mining*, 687–696. IEEE.
- Zou, H.; Cui, P.; Li, B.; Shen, Z.; Ma, J.; Yang, H.; and He, Y. 2020. Counterfactual Prediction for Bundle Treatment. *Advances in Neural Information Processing Systems*, 33.

Appendix

Theoretical Analysis of CAUSALMTA

Proof of weights calculation. In order to create a pseudo-population which debiases by means of sample re-weighting, the weights should cater to the equation $W_T(\mathbf{U}, \mathbf{C}) = p(\mathbf{C})/p(\mathbf{C}|\mathbf{U})$ (Fong et al. 2018; Zou et al. 2020). When we learn a variational distribution $q_\phi(\mathbf{z}|\mathbf{c})$ of the original channel assignment of touch-points, the variational sample weights can be computed as follows:

$$\begin{aligned} w^i &= W_T(\mathbf{u}^i, \mathbf{c}^i) = \frac{p(\mathbf{c}^i)}{p(\mathbf{c}^i|\mathbf{u}^i)} \\ &= \frac{p(\mathbf{c}^i)}{\int_{\mathbf{z}} p(\mathbf{c}^i|\mathbf{z}) p(\mathbf{z}|\mathbf{u}^i) d\mathbf{z}} = \frac{1}{\int_{\mathbf{z}} \frac{p(\mathbf{c}^i|\mathbf{z})}{p(\mathbf{c}^i)} p(\mathbf{z}|\mathbf{u}^i) d\mathbf{z}} \\ &= \frac{1}{\int_{\mathbf{z}} \frac{p(\mathbf{z}|\mathbf{c}^i)}{p(\mathbf{z})} p(\mathbf{z}|\mathbf{u}^i) d\mathbf{z}} = \frac{1}{\int_{\mathbf{z}} \frac{p(\mathbf{z}|\mathbf{u}^i)}{p(\mathbf{z})} p(\mathbf{z}|\mathbf{c}^i) d\mathbf{z}} \\ &= \frac{1}{\int_{\mathbf{z}} \frac{p(\mathbf{z}, \mathbf{u}^i)}{p(\mathbf{z}) p(\mathbf{u}^i)} p(\mathbf{z}|\mathbf{c}^i) d\mathbf{z}} = \frac{1}{\int_{\mathbf{z}} \frac{1}{W_Z(\mathbf{u}^i, \mathbf{z})} p(\mathbf{z}|\mathbf{c}^i) d\mathbf{z}} \\ &= \frac{1}{\mathbb{E}_{\mathbf{z} \sim q_\phi(\mathbf{z}|\mathbf{c}^i)} \left[\frac{1}{W_Z(\mathbf{u}^i, \mathbf{z})} \right]} \end{aligned}$$

Where $W_Z(\mathbf{U}, \mathbf{Z})$ can be viewed as the density ratio estimation to decorrelate \mathbf{U} and \mathbf{Z} for points in space $\mathcal{U} \times \mathcal{Z}$.

Proof of the journal reweighting module. In the task of multi-touch attribution, provided with observational data, we hope to learn a hypothesis $f_{\theta_p} : \mathcal{U} \times \mathcal{C} \mapsto \mathbb{R}$ with model parameters θ_p , which predicts the conversion rate based on the confounders and touch-points. In this setting, the concept of counterfactual is to guarantee the learned hypothesis to predict accurate outcome when the assignment of touch-point (e.g., the channel preference) is random. For the individual \mathbf{U} , when $\mathcal{L}()$ denotes the error function and $y()$ denotes the ground-truth outcome, the prediction error can be formed as:

$$\mathcal{E}(\mathbf{U}) = \mathbb{E}_{\mathbf{C} \sim p(\mathbf{C})} [\mathcal{L}(f_{\theta_p}(\mathbf{U}, \mathbf{C}), y(\mathbf{U}, \mathbf{C}))]$$

The target of the counterfactual prediction error to be minimized is $\mathcal{E}_{cf} = \mathbb{E}_{\mathbf{U} \sim p(\mathbf{U})} [\mathcal{E}(\mathbf{U})]$. But in the observational dataset, the touch-points are assigned based on confounders, i.e., $\mathbf{C} \sim p(\mathbf{C}|\mathbf{U})$. Instead of directly using supervised learning, optimizing the prediction error on the re-weighted data

$$\mathcal{E}_f^w = \mathbb{E}_{\mathbf{U}, \mathbf{C} \sim p(\mathbf{U}, \mathbf{C})} [\mathcal{L}(f_{\theta_p}(\mathbf{U}, \mathbf{C}), y(\mathbf{U}, \mathbf{C})) W_T(\mathbf{U}, \mathbf{C})]$$

can lead to a more accuracy counterfactual prediction.

Assuming a family G of functions $g : \mathcal{U} \times \mathcal{C} \mapsto \mathbb{R}$, and we have $\mathcal{L}(f(\mathbf{U}, \mathbf{C}), y(\mathbf{U}, \mathbf{C})) = l(\mathbf{U}, \mathbf{C}) \in G$. We can therefore bridge the gap between the counterfactual loss and

the re-weighted loss under observational data.

$$\begin{aligned} &\mathcal{E}_{cf} - \mathcal{E}_f^w \\ &= \int_{\mathbf{U}} \int_{\mathbf{C}} (p(\mathbf{U})p(\mathbf{C}) - W_T(\mathbf{U}, \mathbf{C})p(\mathbf{U}, \mathbf{C})) \\ &\quad \cdot \mathcal{L}(f(\mathbf{U}, \mathbf{C}), y(\mathbf{U}, \mathbf{C})) d\mathbf{U} d\mathbf{C} \\ &\leq \left| \int_{\mathbf{U}} \int_{\mathbf{C}} (p(\mathbf{U})p(\mathbf{C}) - W_T(\mathbf{U}, \mathbf{C})p(\mathbf{U}, \mathbf{C})) \right. \\ &\quad \cdot \mathcal{L}(f(\mathbf{U}, \mathbf{C}), y(\mathbf{U}, \mathbf{C})) d\mathbf{U} d\mathbf{C} \left. \right| \\ &\leq \sup_{g \in G} \left| \int_{\mathbf{U}} \int_{\mathbf{C}} (p(\mathbf{U})p(\mathbf{C}) - W_T(\mathbf{U}, \mathbf{C})p(\mathbf{U}, \mathbf{C})) \right. \\ &\quad \cdot g(\mathbf{U}, \mathbf{C}) d\mathbf{U} d\mathbf{C} \left. \right| \\ &= IPM_G(W_T(\mathbf{U}, \mathbf{C})p(\mathbf{U}, \mathbf{C}), p(\mathbf{U})p(\mathbf{C})) \end{aligned}$$

When $W_T(\mathbf{U}, \mathbf{C}) = \frac{p(\mathbf{C})}{p(\mathbf{C}|\mathbf{U})}$, we have:

$$\begin{aligned} IPM_G(W_T(\mathbf{U}, \mathbf{C})p(\mathbf{U}, \mathbf{C}), p(\mathbf{U})p(\mathbf{C})) &= 0 \\ \mathcal{E}_f^w &= \mathcal{E}_{cf} \end{aligned}$$

Proof of the causal prediction module. The optimal prediction probabilities of ad exposure are given by

$$\mathbf{c}^{rev*} = \arg \max_{\mathbf{c}^{rev}} \sum_{k=1}^K \int_{\mathbf{v}^{rev}} p(\mathbf{v}^{rev}|\mathbf{c}^k) \log(\mathbf{c}_k^{rev}) d\mathbf{v}^{rev}$$

By maximizing value function and leveraging Lagrange multiplies, we can derive \mathbf{c}^{rev*} by the following form

$$\arg \max_{\mathbf{c}^{rev}} \sum_{k=1}^K (p(\mathbf{v}^{rev}|\mathbf{c}^k) \log(\mathbf{c}_k^{rev})) + \chi \left(\sum_{k=1}^K \mathbf{c}_k^{rev} - 1 \right)$$

We have $\mathbf{c}_k^{rev} = -\frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\chi} = \frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m)}$ by setting the above derivative to zero and solving \mathbf{c}^{rev*} .

Therefore, the objective $\min_{\mathbf{v}^{rev}} \mathcal{L}_{rev}$ for the learned representation \mathbf{v}^{rev} becomes

$$\min_{\mathbf{v}^{rev}} \sum_{k=1}^K \mathbb{E}_{\mathbf{v}^{rev} \sim p(\mathbf{v}^{rev}|\mathbf{c}_k)} \left[\log \frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m)} \right]$$

We can derive that

$$\begin{aligned} &\sum_{k=1}^K \mathbb{E}_{\mathbf{v}^{rev} \sim p(\mathbf{v}^{rev}|\mathbf{c}_k)} \left[\log \frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m)} \right] + K \log K \\ &= \sum_{k=1}^K \left(\mathbb{E}_{\mathbf{v}^{rev} \sim p(\mathbf{v}^{rev}|\mathbf{c}_k)} \left[\log \frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m)} \right] + \log K \right) \\ &= \sum_{k=1}^K \mathbb{E}_{\mathbf{v}^{rev} \sim p(\mathbf{v}^{rev}|\mathbf{c}_k)} \left[\log \frac{p(\mathbf{v}^{rev}|\mathbf{c}^k)}{\frac{1}{K} \sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m)} \right] \\ &= \sum_{k=1}^K KL \left(p(\mathbf{v}^{rev}|\mathbf{c}^k) \parallel \frac{1}{K} \sum_{m=1}^K p(\mathbf{v}^{rev}|\mathbf{c}^m) \right) \\ &= K \cdot JSD(p(\mathbf{v}^{rev}|\mathbf{c}^1), \dots, p(\mathbf{v}^{rev}|\mathbf{c}^K)) \end{aligned}$$

where $KL(\cdot||\cdot)$ is the KL divergence and $JSD(\cdot, \dots, \cdot)$ is the Jensen-Shannon Divergence (Li et al. 2018; Bica et al. 2020) in the multi-distribution form. Because $JSD(\cdot, \dots, \cdot)$ is non-negative and equals zero when all distributions are equal and $K \log K$ is a constant, we have that $p(\mathbf{v}_t^{rev} | c_1) = \dots = p(\mathbf{v}_t^{rev} | c_K)$ by minimizing \mathcal{L}_{rev} .

Experiments

Details of data preprocessing. The attribution modeling for bidding dataset published by **Criteo** company is widely deployed in the research of modeling user behavior and ad attribution (Diemert et al. 2017; Ren and etc. 2018; Kumar et al. 2020). As a sample of 30 days of Criteo live traffic data, this dataset has more than 16 million ad impression records and 45 thousand conversions over 700 ad campaigns. Each ad impression record contains items such as timestamp, user id, ad campaign, and side information. There is also a label denotes whether a click action has occurred, and the corresponding conversion ID if this sequence of ad impressions finally leads to a conversion. We preprocess the raw Criteo dataset in the following procedures: (i) we count the top 10 campaigns with the largest number of ad impression records and filter out the ad impression records corresponding to other campaigns; (ii) we group the ad impression entries, which have the same user id and conversion id, into the same sequence and sort each sequence by timestamp; (iii) for a conversion id of -1, i.e., for a specific user, a group of ad impression records that did not cause the user to convert, the original group is divided at a time interval of 3 days based on timestamp; (iv) we filter out ad sequences that are less than 3 in length; (v) we divide it into the train set and the test set, and ensure the set of user id in the test set is a subset of user id in the train set.

Details of compared baselines. We compare CAUSALMTA with four kinds of baselines, i.e., statistical learning-based methods (SL), deep learning-based methods (DL), causal learning-based methods (CL) and ablations.

The statistical learning-based methods consist of three works:

- **LR** (Logistic Regression) model for ad attribution is proposed by Shao and Li (Shao and Li 2011), in which channel’s attribution values are calculated as the learned coefficients.
- **SP** (Simple Probabilistic) model calculates the conversion rate taking into the conversion probability from the observed data into account. As in (Dalessandro et al. 2012), the conversion rate is

$$p(y = 1 | \{c_j\}_{j=1}^{m_i}) = 1 - \prod_j^{m_i} (1 - \Pr(y = 1 | c_j = k))$$

- **AH** (Additive Hazard) proposed by Zhang *et al.* (Zhang, Wei, and Ren 2014) is the first user conversion estimation model based on survival analysis and additive hazard function.

The deep learning-based methods also consist of three works:

- **DNAMTA** is the Deep Neural Net with Attention Multi-touch Attribution model proposed by Arava *et al.* (Arava et al. 2018). It leverages LSTM and attention mechanism to model the dynamic interaction between ad channels, and incorporates user-context information to reduce estimation bias.
- **DARNN** is the Dual-Attention Recurrent Neural Network proposed in (Ren and etc. 2018) which uses dual-attention RNNs to combine both post-view and post-click attribution patterns together for the user conversion estimation.
- **DeepMTA** is a phased-LSTM based model (Yang, Dyer, and Wang 2020) which combines deep neural networks and additive feature explanation model for interpretable online multi-touch attribution. For fair comparison, we replace the phased-LSTM with vanilla LSTM.

The causal learning-based methods consist of two works:

- **JDMTA** is a causal-inspired model (Du and etc. 2019) which employs Shapley Value to compute the attribution credits for touchpoints.
- **CAMTA** is the Causal Attention Model for Multi-touch Attribution proposed by Kumar *et al.* (Kumar et al. 2020). This model leverages counterfactual recurrent network to minimize selection bias in channel assignment while conducting conversion estimation.

We also compare CAUSALMTA with its two ablations:

- **CM-RW** removes the journey reweighting module and treats all journeys equally. It only employs the proposed causal conversion prediction model for MTA.
- **CM-CAUSAL** replaces the causal RNN predictor with traditional RNN and only utilize the reweighting mechanism to eliminate the confounding bias of static user attributes.