arXiv:1808.03737v2 [cs.IR] 30 Aug 2018

Learning Multi-touch Conversion Attribution with Dual-attention Mechanisms for Online Advertising

Kan Ren, Yuchen Fang, Weinan Zhang, Shuhao Liu,
Jiajun Li, Ya Zhang, Yong Yu, Jun Wang
Shanghai Jiao Tong University, University College London
{kren, arthur_fyc, wnzhang, yyu}@apex.sjtu.edu.cn, j.wang@cs.ucl.ac.uk

ABSTRACT

In online advertising, the Internet users may be exposed to a sequence of different ad campaigns, i.e., display ads, search, or referrals from multiple channels, before led up to any final sales conversion and transaction. For both campaigners and publishers, it is fundamentally critical to estimate the contribution from ad campaign touch-points during the customer journey (conversion funnel) and assign the right credit to the right ad exposure accordingly. However, the existing research on the multi-touch attribution problem lacks a principled way of utilizing the users' pre-conversion actions (i.e., clicks), and quite often fails to model the sequential patterns among the touch points from a user's behavior data. To make it worse, the current industry practice is merely employing a set of arbitrary rules as the attribution model, e.g., the popular *last-touch* model assigns 100% credit to the final touch-point regardless of actual attributions. In this paper, we propose a Dual-attention Recurrent Neural Network (DARNN) for the multi-touch attribution problem. It *learns* the attribution values through an attention mechanism directly from the conversion estimation objective. To achieve this, we utilize sequence-to-sequence prediction for user clicks, and combine both post-view and post-click attribution patterns together for the final conversion estimation. To quantitatively benchmark attribution models, we also propose a novel yet practical attribution evaluation scheme through the proxy of budget allocation (under the estimated attributions) over ad channels. The experimental results on two real datasets demonstrate the significant performance gains of our attribution model against the state of the art.

KEYWORDS

Conversion Attribution, Multi-Touch Attribution, Computational Advertising, Attention Mechanism, Deep Learning

ACM Reference Format:
Kan Ren, Yuchen Fang, Weinan Zhang, Shuhao Liu, Jiajun Li, Ya Zhang, Yong Yu, Jun Wang . 2018. Learning Multi-touch Conversion Attribution with Dual-attention Mechanisms for Online Advertising. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271677

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CIKM '18, October 22–26, 2018, Torino, Italy
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271677

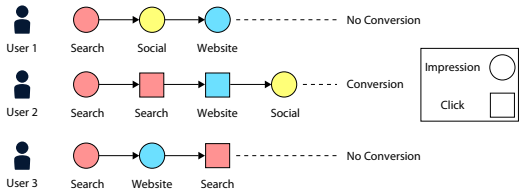

Figure 1: An illustration of different user activity sequences over multiple channels. Here we illustrate two user interactions with the ad contents, i.e., impressions and clicks, over three typical ad channels. Each user would gain an impression and then would probably click on that. After a sequence of user actions, the final conversion may be drawn according to the comprehensive experiments.

1 INTRODUCTION

A benefit for online advertising is that advertisers would be able to get a significant amount of user feedbacks to measure the successfulness of their ad campaigns and optimize them accordingly. Aiming at delivering the optimization above, computational advertising has gained a large attraction and achieved great progress in many technical fields, including user targeting [17, 22], bidding strategy [18, 21, 35] and budget pacing [1, 2, 14].

As illustrated in Figure 1, with online advertising, an Internet user may be exposed to a sequence of ad campaigns from multiple channels, such as search engines, social media, mobile platforms, before reaching to any final conversion and transaction. It is crucial for advertisers to attribute the right conversion credit onto each *touch point* (i.e., the interaction between the user and the ad content, along the customer journey). The reasons are threefold. First, advertisers should know the contribution of each single touch point to the final conversion so as to make informed impression-level ad buying decisions [15]. Second, if the attribution of each conversion over multiple ad exposures can be accurately and reliably estimated, a more quantitative credit-based ad pricing scheme can be established between advertisers and publishers (and ad tech providers). Last but not least, the attribution aggregated over ad channels may provide useful guidance for advertisers to allocate their budgets over these ad channels so as to acquire more positive user actions with lower cost in the next-round campaigns [10].

Traditionally, the attribution problem is addressed simply by a rule-based approach among most advertisers, such as first touch, last touch and other simple mechanisms [27]. Specifically, the first (last) touch method generally attributes the conversion credits to the first (last) user interaction with the ad content. While these methods are easy to deploy, they obviously lack adequate capability of useful pattern recognition to support the higher-level budget

allocation [6]. Shao and Li [24] proposed the first data-driven multi-touch attribution model to allocate the credits to all the user touch points. Whereafter, many works have been published including probabilistic models using some distributional assumptions [7, 29] and additive exciting process [12, 30, 37]. Despite of the claimed advantages, there are three important factors missing in the above solutions.

Sequential Pattern Modeling. These methods are all based on the assumption that the user conversion would be driven by the individual advertising touch point of positive influence [24, 37], which may not be realistic for the user journey (conversion funnel). In fact, sequential patterns within the user browsing behavior are of great value for response prediction or decision making in many fields such as recommender systems [23], information retrieval [26] and search advertising [36].

Data-driven Credit Assignment. The attribution credits obtained in these models are heuristically assigned to each user interaction with the advertiser's contents, rather than statistically learned from the data. For example, Ji and Wang [12] proposed that the final conversion would be driven by the additive hazard rate of being converted at the time of each previous touch point, which pre-assumes that the more user exposures, the higher probability of the final user action. As is illustrated in Figure 2 of one real-world dataset used in our experiments, the conversion rate does not necessarily increase w.r.t. the user action sequence length. This assumption may cause unconscionable ad exposures and may destruct user experience in online service.

Different Pre-conversion Behaviors. Almost all the related works ignore the difference between various types of user behaviors. Specifically, they assume the attribution are solely based on post-view or post-click, or even simply treat these behavior types equally for conversion attribution, where the credits are placed solely on impressions or clicks, or even discard the difference between them. These treatments are not effective since the user shows apparently different preferences behind the different interactive actions, which may (not) lead to the final conversion in different degrees.

To address the above limitations, in this paper, we propose a Dual-attention Recurrent Neural Network (DARNN) to capture the sequential user behavior patterns and learn the optimal attentions as the conversion attributions. Specificall, our model has two learning objectives. On one hand, we utilize sequence-to-sequence architecture to model the relationship between the impressions and the click actions, where the click behavior modeling is handled in this procedure. On the other hand, the final sequence prediction is the probability of the user conversion with the attention learned from the sequential modeling. The advantage of the attention mechanism is that it not only contributes to the sequence prediction accuracy, but also naturally *learns* the attribution of the conversion action over the whole sequence of the touch points. Moreover, DARNN applies the attention mechanism not merely on the features of the original touch point, but additionally over the learned hidden states of click actions, and then dynamically combines both attentions to predict the final conversion, which is the reason of *dual* attention. By this means the conversion estimation has captured both impression-level and click-level patterns. We also note that both dual-attention and the dynamic combination for the final conversion prediction are statistically learned from the data.

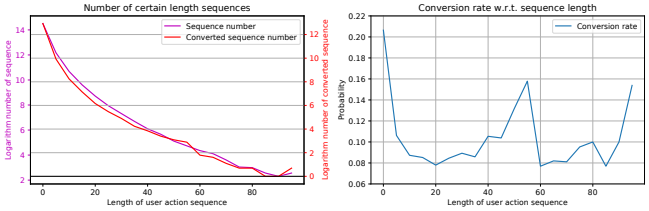

Figure 2: The statistics of action sequence lengths and the conversion rates over Criteo dataset. The left plot shows the number of sequences and the converted sequences w.r.t. sequence length; the right plot presents the user conversion rate w.r.t. the user action sequence length.

In addition, we also propose an offline evaluation framework for conversion attribution mechanisms. Since the obtained attribution credits over different channels could direct the budget allocation for the subsequent ad campaigns. However, none of the related work has empirically shown the effectiveness of the calculated attributions [7, 12, 13, 24, 30, 37] unless spending a huge budget to conduct online A/B test [10, 29]. Since budget allocation over different ad channels is always an important decision for advertisers to make, it is crucial for them to evaluate the performance of their multi-touch attribution methods, before the online A/B testing phase.

To sum up, the novelty of our work is fourfold. (i) We build sequential pattern learning models for user behavior sequence. (ii) Our model learns the attribution from the final conversion estimation rather than heuristically assigning credits. (iii) We combine different attributions on various user action types. (iv) We also propose an offline evaluation protocol to measure the effectiveness of the attribution model through ad budget allocation and campaign data replay. Our experimental results prove the significant improvement (over 5.5%) of our DARNN model on conversion estimation performance against state-of-the-art baselines. The back replay evaluation also illustrates that the proposed model achieves the best cost effective performance.

2 RELATED WORKS

In online advertising, conversion attribution is commonly calculated by some rule-based methods, such as first-touch and last-touch, whereafter the return-on-investment (ROI) is gained based on the achieved attribution results which may result in some bias [6]. In recent years, many works based on multi-touch attribution (MTA) have been proposed for modeling the attribution for the sequential touch points over various channels [4, 25]. Shao and Li [24] proposed the first work for data-driven multi-touch attribution model, which estimates the conversion rate based on the viewed ads of the user by the bagged logistic regression model. Some other works are mainly based on the probabilistic models with some distributional assumptions. Dalessandro et al. [7] proposed a causally motivated methodology that the conversion credits should be assigned by a causal cooperation model such as Shapley value. Xu et al. [29] argued that the user behavior has different additional effect on the final user decision of conversion and proposed a lift-based prediction model for real-time ad delivery. However, these methods did not take all the touch points of a user into the whole consideration so that the temporal and sequential factors

were ignored [12]. Moreover, these models did either not consider much of sequential pattern modeling, which has been shown great effectiveness of user modeling [36].

As for the multiple interactions between advertisers and users, many works proposed the exciting point process methods for user behavior modeling. Yan et al. [31] developed a two-dimensional Hawkes process model to capture the influences from sellers' activities on their contributions to the winning outcome in sales pipeline analytics. Xu et al. [30] presented an MTA model based on mutually exciting point process which independently considered the impressions and clicks as random events along the continuous time. These exciting point process methods only considered the occurrence of the event which ignored the data of non-conversion cases. For analyzing the cumulative effects of the touch points, many works [12, 25, 29, 37] made an assumption that the final conversion was influenced by the additive contributions from the touch points along the user browsing history. However, it might result in a trend that more ad exposures were better which severely destroyed the user experience [32]. The reason is that the attribution of each touch points along the user behavior sequence may positively contribute or counteract the final conversion. Thus, it is more reasonable to dynamically calculate and assign the attribution credits over the user behaviors.

Another school of MTA modeling is based on the survival theory [12, 13, 37], which models the conversion event as the predictive goal and estimates the probability for the event occurrence at the specific time while considering the censored data, i.e., the true occurrence time is later than the observation time. Nevertheless, these methodologies focus more on single point prediction and fail to consider the sequential patterns embedded in the user browsing history. Moreover, the obtained attribution credits are mainly calculated based on heuristic additive assumptions, which may not be effective in practice. They also made assumptions about the survival function such as exponential hazard function [12, 37] and Weibull distribution for hazard rate estimation [13] to make their model parameterized and thus optimizable. However, such parameterization could severely constrain the capacity of the model to fit various real-world data.

Considering all the limitations above, we propose a dual-attention recurrent neural network for both conversion estimation and attribution. Attention mechanism is originally proposed for machine translation tasks, where a sequence-to-sequence model samples the next output word by attend each word of the input sentence [3]. In our problem, the attention is modeled as the attribution which may dynamically *learn* to assign the credits over all the historical touch points for a specific user. The sequential patterns have been efficiently captured by the recurrent mechanism.

Moreover, few works have discussed the budget allocation from the obtained attribution model. Diemert et al. [8] proposed a bidding strategy based on the attribution credits for each real-time auction, which is not appropriate in general applications of online advertising. Geyik et al. [10] presented a method for online budget allocation based on the obtained ROI from conversion attribution. We borrow the idea of the ROI calculation from [10] and devise an offline evaluation framework for multi-touch conversion attribution, which is the first offline experimental evaluation methodology for attribution models.

3 METHODOLOGY

In this section, we firstly formulate the problem of the multi-touch conversion attribution, and then propose our sequential behavior modeling with dual-attention mechanism. Finally we present our evaluation protocol for conversion attribution guided budget allocation.

3.1 Problem Definition

Without loss of generality, let us focus our study on the advertiser side. When a user u_i is taking Internet activities, e.g., browsing online contents, querying search engines or playing on social media, etc., there would be many sequential interactions between this user and the ad contents of an advertiser, which are called touch points for the ad campaign. The observations are the user browsing sequences $\{u_i, \{q_j\}_{j=1}^{m_i}, y_i, T_i\}_{i=1}^{n}$ for each user u_i who generates totally m_i browsing activities with the ad of the advertiser. y_i is the indicator of whether the user converts and T_i is the conversion time if the conversion occurs, otherwise null. Each touch point $q_j = (\boldsymbol{x}_j, z_j)$ contains the feature vector $\boldsymbol{x}_j$ of the this touch point and the binary action type z_j, i.e., non-click impression or click. Among them, the feature $\boldsymbol{x}_j$ includes the side information of the user and the ad contents, e.g., user ID, advertising form, website, the operation systems and browser information, also with the channel ID feature c_j over which this touch point is delivered and the time t_j of the interaction occurrence.

The goal is to model the sequential user patterns and derive efficient conversion attribution credits for all the touch points $\{q_j\}_{j=1}^{m_i}$ along the user browsing sequence. In return, the better conversion attribution obtained, the higher accuracy of the user conversion estimation for each browsing sequence. Similar formulation has been adopted in many literatures [12, 13, 37]. In Sec. 3.2, we present a recurrent neural network to model the sequential patterns and the final conversion rate. We also apply sequence-to-sequence modeling for user click pattern mining and jointly learn impression and click patterns for conversion estimation. The key component in this sequence modeling methodology is the dual-attention mechanism which takes two types of the user actions (i.e., impressions and clicks) into a unified comprehensive framework and facilitate the conversion modeling, which is described in Sec. 3.3. As a result, the obtained attention from the sequence modeling is naturally the conversion attribution over the whole user browsing history.

Interestingly, the derived conversion attribution also contributes to budget allocation for the subsequent ad delivery [10]. In Sec. 3.4, we propose an evaluation protocol for budget allocation with offline campaign data.

3.2 Sequential Modeling

We utilize recurrent neural network (RNN) for sequential user modeling, as illustrated in Figure 3. Leveraging RNN for sequential modeling and time-series prediction has been widely applied in information retrieval systems [19, 26, 33]. Note that our methodology aims at final conversion estimation rather than sequential prediction for click at each touch point.

The whole structure can be divided as three separate parts that (i) the encoder for the impression-level behavior modeling; (ii) the decoder and sequential prediction for click probability; (iii) taking the above modeling output we implement dual-attention for jointly modeling impression and click behavior and produce the

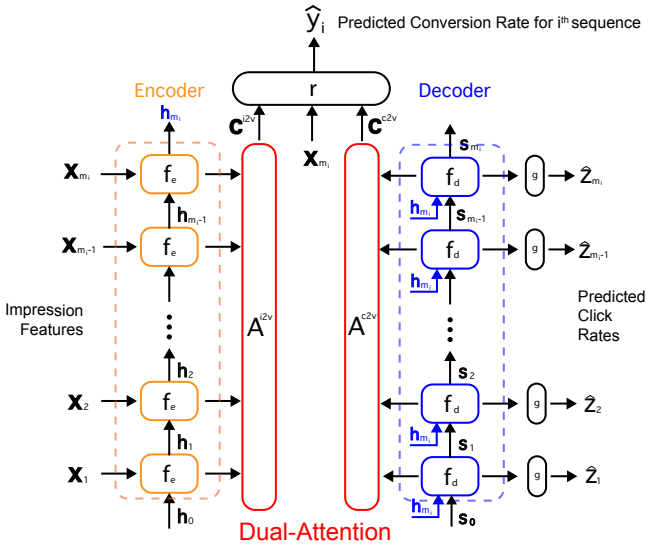

Figure 3: Sequential modeling with dual-attention.

final conversion estimation. We will clarify the first two parts in this section and discuss the attention mechanism later.

Impression-level Behavior Modeling. For the i^{th} user behavior sequence $\{u_i, \{q_j\}_{j=1}^{m_i}, y_i, T_i\}$ where $q_j = (\boldsymbol{x}_j, z_j)$, the input feature sequence to the RNN model is $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_j, \ldots, \boldsymbol{x}_{m_i})$. Since the side-information feature vector is mostly categorical [34], we firstly utilize an embedding layer to transform the sparse input feature into dense representation vector for subsequent training, which has been widely used in the related literatures [20, 28].

Then we feed the embedded feature vectors through the encoder RNN function f_e approach as

$$\boldsymbol{h}_j = f_e(\boldsymbol{x}_j, \boldsymbol{h}_{j-1}) , \tag{1}$$

where $\boldsymbol{h}$ is the hidden vector of each time step j. We implement f_e as a standard long short-term memory (LSTM) model described in [11], which has been widely used in natural language processing fields.

Click-level Sequential Prediction. In this part, our goal is to model the click action at each time when each ad is shown to the user.

In the sequence-to-sequence model, the decoder defines a probability over the click outcomes $\mathbf{z}$ by decomposing the *joint probability* into the ordered conditionals as

$$p(\mathbf{z}) = \prod_{j=1}^{m_i} p(z_j = 1 \mid \{z_1, \ldots, z_{j-1}\}, \mathbf{x}) . \tag{2}$$

Note that $\mathbf{z} = (z_1, \ldots, z_{m_i})$ and $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{m_i})$. With this decoder each conditional probability is modeled as

$$\hat{z}_j = p(z_j = 1 \mid \{z_1, \ldots, z_{j-1}\}, \mathbf{x}) = g(z_{j-1}, \boldsymbol{s}_j) . \tag{3}$$

Here g is the output function which is a multi-layer fully connected perceptron with sigmoid activation function $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$ that outputs the probability of $z_j = 1$. And $\boldsymbol{s}_j$ is the hidden vector at click-level of the j^{th} touch point, calculated by

$$\boldsymbol{s}_j = f_d(\boldsymbol{s}_{j-1}, z_{j-1}, \boldsymbol{h}_{m_i}) , \tag{4}$$

where f_d is the nonlinear decoder RNN function, potentially multi-layered, that models the sequential click patterns for user behavior sequence. We utilize the same structure of LSTM model as the encoder f_e. Each hidden state $\boldsymbol{s}_j$ in the decoder uses the last hidden state $\boldsymbol{h}_{m_i}$ from the encoder. Our first loss is based on the sequential prediction for click probabilities as

$$L^c = \sum_{i=1}^{n} \sum_{j=1}^{m_i} -z_j \log \hat{z}_j - (1 - z_j) \log(1 - \hat{z}_j) . \tag{5}$$

There are two rationales for the sequence-to-sequence click prediction in this work. The first is to some extent similar with the idea of multi-task learning to alleviate the data sparsity problem and conduct a shared base representation of user behavior features. As is known that the users follow a pattern of actions that they may click after impression of the ad and after a sequence of ad delivery they may (not) drive the final conversion, which derives the data sparsity problem behind the "impression-click-conversion" action pattern [16]. Specifically, clicks are less frequent events than impressions and conversions are much rarer than clicks. It is necessary to conduct a methodology to tackle with the data sparsity challenge. Our intuition is to utilize the signal of click behavior to boost the estimation capacity for the sparse conversion behaviors. Another reason for the sequential click pattern mining is to obtain the click-level attribution modeling for multiple pre-conversion behavior modeling, which has shown statistically more important attribution credits than impression-level behaviors in our experimental results.

3.3 Learning Attribution with Dual-attention

Our final goal is to model the sequential user patterns and predict the conversion probability. The final output is calculated as

$$\begin{aligned} \boldsymbol{c}^{i2v} &= A^{i2v}(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_j, \ldots, \boldsymbol{h}_{m_i}) , \\ \boldsymbol{c}^{c2v} &= A^{c2v}(\boldsymbol{s}_1, \ldots, \boldsymbol{s}_j, \ldots, \boldsymbol{s}_{m_i}) , \\ \hat{y}_i &= p(y = 1 \mid \mathbf{x}, \mathbf{z}) = r(\boldsymbol{x}_{m_i}, \boldsymbol{c}^{i2v}, \boldsymbol{c}^{c2v}) , \end{aligned} \tag{6}$$

where r contains a weighting function for balancing impression-level and click-level attribution, which will be described in detail later in this section, and a dense multi-layer neural network for the final conversion prediction. $\boldsymbol{x}_{m_i}$ is the feature vector of the last touch point which would be fed through the same embedding layer as that in Sec. 3.2. $\boldsymbol{c}^{i2v}$ is the context vectors of all the input user behavior vectors capturing impression patterns and $\boldsymbol{c}^{c2v}$ is the context vector from modeling the click patterns for conversion estimation.

Learning Attention through Conversion Estimation. The loss is calculated by the cross entropy for the conversion estimation that

$$L^v = \sum_{i=1}^{n} -y_i \log \hat{y}_i - (1 - y_i) \log(1 - \hat{y}_i) . \tag{7}$$

The key component is the attention input $\boldsymbol{c}^{i2v}$ and $\boldsymbol{c}^{c2v}$ from impression-level and click-level, respectively. The mechanism of attention function A is illustrated in Figure 4.

To calculate the impression-to-conversion attention $\boldsymbol{c}^{i2v}$ and the click-to-conversion attention $\boldsymbol{c}^{c2v}$, we propose a unified energy-based function as

$$\boldsymbol{c} = l(\boldsymbol{h}_1, \ldots, \boldsymbol{h}_j, \ldots, \boldsymbol{h}_{m_i}) = \sum_{j=1}^{m_i} a_j \boldsymbol{h}_j . \tag{8}$$

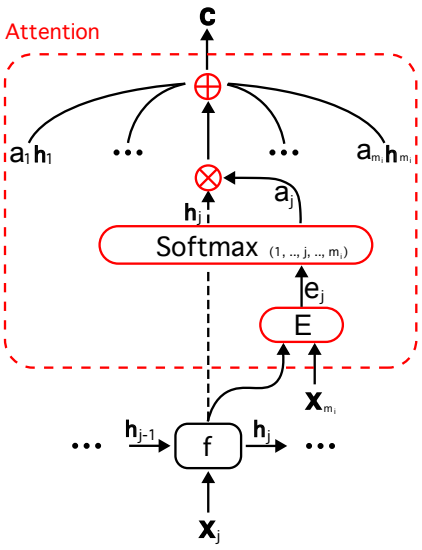

Figure 4: Attention calculation mechanism.

Note that this formulation is expressed to calculate $\boldsymbol{c}^{i2v}$ while without losing generality by replacing $\boldsymbol{x}$ and $\boldsymbol{h}$ in Eq. (1) with $\boldsymbol{s}$, z and $\boldsymbol{h}_{m_i}$ in Eq. (4) for $\boldsymbol{c}^{c2v}$ calculation.

And the weight a_j is calculated based on the softmax operated energy value e_j as

$$a_j = \frac{\exp(e_j)}{\sum_{k=1}^{m_i} \exp(e_k)} , \tag{9}$$

where

$$e_j = E(\boldsymbol{h}_j, \boldsymbol{x}_{m_i}) \tag{10}$$

is an energy model which scores the credit of each touch point to the final conversion. Note that the energy function E is a nonlinear multi-layer deep neural network with tanh activation function $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$. The way we calculate the attention through the energy function E is similar to that in the natural language processing field [3, 9].

As a result, the dual-attention mechanism is expressed as

$$\begin{aligned} \boldsymbol{c}^{i2v} &= \sum_{j=1}^{m_i} a_j^{i2v} \boldsymbol{h}_j , \\ \boldsymbol{c}^{c2v} &= \sum_{j=1}^{m_i} a_j^{c2v} \boldsymbol{s}_j , \end{aligned} \tag{11}$$

and the values of a in both attention calculation are obtained through Eqs. (9) and (10).

Attribution Calculation with Dual-attention. Till now, we have obtained the estimated conversion probability $p(y|\mathbf{x}, \mathbf{z})$ and the attention results a_j^{i2v} and a_j^{c2v} for each touch point, based on which we can naturally assign the credits for each touch point j.

Recall that the final conversion estimation is based on the learned dual-attention vector, i.e., $\boldsymbol{c}_j^{i2v}$ and $\boldsymbol{c}_j^{c2v}$, and the final touch point feature vector $\boldsymbol{x}_{m_i}$, here we adopt a dynamic weighting function f_λ to balance the effcts of the two attentions that

$$\lambda = \frac{\exp\left[f_\lambda(\boldsymbol{x}_{m_i}, \boldsymbol{c}^{c2v})\right]}{\exp\left[f_\lambda(\boldsymbol{x}_{m_i}, \boldsymbol{c}^{i2v})\right] + \exp\left[f_\lambda(\boldsymbol{x}_{m_i}, \boldsymbol{c}^{c2v})\right]} , \tag{12}$$

where λ measures the importance of the click-level attention w.r.t. that from the impression-level and f_λ is a multi-layer perceptron whose goal is to learn the weight of two attention results for the final conversion estimation.

Thus, the estimation function r mentioned first in Eq. (6) is that

$$r(\boldsymbol{x}_{m_i}, \boldsymbol{c}^{i2v}, \boldsymbol{c}^{c2v}) = r_{\text{conv}}\left((1-\lambda)\cdot \boldsymbol{c}^{i2v} + \lambda \cdot \boldsymbol{c}^{c2v}\right) . \tag{13}$$

Here $r_{\text{conv}}(\cdot)$ is a multi-layer neural network for conversion estimation with sigmoid activation function.

Now that we have weighted the contribution of impression-level and click-level attentions for final conversion estimation, we can naturally obtain the attribution for each touch point through these learned patterns as

$$\text{Attr}_j = (1-\lambda)\cdot a_j^{i2v} + \lambda \cdot a_j^{c2v} . \tag{14}$$

The motivation for building such a dual-attention mechanism is that we care both the impression-level and the click-level user behavior patterns to facilitate conversion estimation and the subsequent attribution results.

3.4 Evaluation Protocol

With the attribution credits allocated to the touch points along the user behavior sequences, our focus moves onto the efficiency of budget allocation based on the calculated attribution credits. Note that almost all the related works report only the conversion estimation performance; few of them test the budget allocation under the obtained attribution credits, except online A/B testing which is expensive and risky. Here we propose a framework to offline evaluate the conversion attribution model based on the historic data of a campaign.

In online advertising, the guideline of the advertiser to allocate budgets for the subsequent ad delivery on different channels of the ad campaign is intuitively based on the past performance. The performance here means the effectiveness, i.e., return on investment (ROI), of the ad delivery onto each channel, which is measured as the obtained positive user conversions w.r.t. the delivered ad costs. The most intuitive idea is to allocate more budgets for the channels or sub-campaigns with higher ROI than others, to gain more user conversions. However, different attribution methods substantially influence the ROI calculation results [10]. Specifically, the idea of our evaluation protocol is to first calculate the ROI performance results for each channel under different attribution models, and then utilize the offline replay of ad delivery history to measure the performance of the obtained fresh conversions and, considering the costs in the offline replay, calculate the effectiveness results of the ad delievery for different evaluation baselines. So that, under this evaluation, the more proper attribution credits one model proposed, the better performance it would obtain in the subsequent budget allocation for different ad channels and naturally obtain better performance through the ad replay evaluation.

Next we will first present the budget allocation method based on attribution-guided ROI results. Then we illustate our back evaluation algorithm w.r.t. the allocated budget scheme for later performance comparison.

ROI-based Budget Allocation. In this stage, the first problem is to allocate the budget $\{b_1, \ldots, b_K\}$ across K channels according to the obtained attribution credits. Here we follow the idea presented in [10] that

$$\text{ROI}_{c_k} = \frac{\sum_{\forall y_i=1} \text{Attr}(c_k|y_i)\, V(y_i)}{\text{Money spent on channel } c_k}, \tag{15}$$

where

$$\text{Attr}(c_k|y_i) = \sum_{j=1}^{m_i} \text{Attr}_j \cdot 1(c_j = c_k) \tag{16}$$

is the overall credit attributed on the channel c_k by aggregating the credit Attr_j of all touch points j's within this channel, $1(\cdot)$ is the indicator function, and $V(y_i)$ is the value of the conversion. After the ROI calculation we allocate budgets for different channels w.r.t. the obtained ROI proportion as $b_k = \frac{\text{ROI}_{c_k}}{\sum_{v=1}^{K} ROI_{c_v}} \times B$ for channel c_k, where B is the total budget.

Back Evaluation under Reallocated Budgets. The historic data is a series of event sequences $\{\text{seq}^s\}$ and each sequence is represented as $\{(q_i, \boldsymbol{x}_i, t_i, y_i, c_i, o_i)\}$ where each user interaction identified by q_i is on the specific channel c_i at time t_i with cost o_i and the feature vector $\boldsymbol{x}_i$ includes the click label information. Each event is either an ad serving event without conversion ($y_i = 0$) or a user conversion event ($y_i = 1$).

In addition, we introduce the concept of *conversion blacklist*. If the budget of one channel c_k is exhausted at moment t, then the conversion events of all the unfinished sequence with ad serving event after t on channel c_k become invalid. These conversions should be put into the conversion blacklist. This is reasonable because if the user cannot observe the ad touch point, it is no guarantee that she will finally convert at the end of the sequence. Such a back test result serves as a lower bound estimation of the true but unknown performance.

Given the budget allocation $(b_1, b_2, \ldots, b_K)$ across K channels, we can make the following back test as presented in Algorithm 1. Specifically, the back test goes over the historic events by their recording time t_i. If there is no budget left for the channel c_k for the back playing event $(q_i, \boldsymbol{x}_i, t_i, y_i, c_i = c_k, o_i)$, then put the sequence indicator seq^s into conversion blacklist. After the back test, we can evaluate the attribution models by the cost O and the obtained valid conversion number Y.

Algorithm 1 Back Evaluation for Budget Allocation

Input: The events $\{(q_i, \boldsymbol{x}_i, t_i, y_i, c_i, o_i)\}$ ordered by the serving time t_i and the budget allocation $\{b_1, \ldots, b_K\}$.
Output: The total conversion number Y and the total cost O.

```
Initially set the blacklist of sequence list B = {} and the obtained conversion number Y = 0, total cost O = 0.
for each event (q_i, x_i, t_i, y_i, c_i, o_i) in the data do
    if seq^s not in B then
        if the budget for c_i channel b_{c_i} > o_i then
            O = O + o_i
            Y = Y + y_i
            b_{c_i} = b_{c_i} - o_i
        else
            Put seq^s into B
        end if
    end if
end for
```

4 EXPERIMENTS

In this section, we firstly present the experiment setup including the description of two real-world datasets, the evaluation measurements and the compared models used in our experiments. Then we illustrate the corresponding results for the two-staged experiment settings. The first stage is for the conversion estimation accuracy while the second one is for the attribution guided budget allocation performance over history data. In addition, we have published our code[1] for repeatable experiments.

4.1 Datasets

In our experiments, we apply our model and the compared baselines over two real-world datasets.

Miaozhen is a leading marketing technique company in China. This dataset [37] includes almost 1.24 billion advertising logs from May 1 to June 30 and April 4 to June 9 in 2013. Specifically it contains about 59 million users and 1044 conversions. These ad contents have been exposed over 2498 channels with 40 advertising forms, such as button ads and social ads. In the dataset, every time a user is exposed to the ad or click on the ad contents, the exact time of the user action with the side information will be recorded. Moreover, it also contains the purchasing information as the conversion of the user with the corresponding timestamp. The user is tracked according to the user cookie identifier which is anonymized in the dataset. With these logs, we are able to reconstruct the time line of the user action sequence including impression and click information, the exposure ad channels and the conversion labels for each sequence.

Criteo is a pioneering company in online advertising research. They have published this dataset[2] for attribution modeling in real-time auction based advertising [8]. This dataset is formed of Criteo live traffic data in a period of 30 days. It has more than 16 million impressions and 45 thousand conversions over 700 campaigns. The impressions in this dataset may derive click actions so each touch point along the user action sequence has a label of whether a click has occurred, and the corresponding conversion ID if this sequence of touch points leads to a conversion event. Each impression log also contains the cost information, which will be used in our second state experiment for attribution effectiveness evaluation. Since the channel data are missing so we take campaign as the budget allocation targets.

Data Preprocessing and Sampling. Since the user conversion is a rare event, we perform negative sampling in data preprocessing. Following [12, 37], the sequence preparation and sampling rules are that (i) if the user has multiple conversion events, her action sequence will be split according to the conversion time to guarantee that each sequence has at most one conversion; (ii) we extract the user action sequences with the minimal length of 3 and maximal length of 20 with the sequence duration within 14 days; (iii) all of the user sequences leading to conversion events have been retained

[1] Repeatable experiment code link: https://github.com/rk2900/deep-conv-attr.

[2] Processed dataset link: http://apex.sjtu.edu.cn/datasets/13.

and we uniformly sample the sequences without conversions to 20 times of the number of converted sequences.

4.2 Evaluation Pipeline and Metrics

Here we present the evaluation pipeline and the measurements over the compared settings. Overall, we have two stages of the experiments.

The first stage focuses on the conversion estimation performance which has been widely adopted in the conversion attribution task [12, 13, 37]. Specifically, given the evaluation samples $\{u_i, \{q_j\}_{j=1}^{m_i}, y_i, T_i\}_{i=1}^{n}$ in the test dataset, the model predicts the output of conversion probability $\hat{y}_i$ after the user going through each sequence of touch points. There are two evaluation metrics for measuring the performance of each model. **Log-loss** is the common measurement to estimate the classification performance for the event probability prediction which is the cross entropy as is expressed in Eq. (7). The other metric is **AUC** (area under ROC curve) which measures the pairwise ranking performance of the classification results between the converted and nonconverted sequence samples.

The second stage aims at the performance of budget allocation, with the calculated conversion attributions, for various channels or sub-campaigns. According to Algorithm 1, we replay all the test campaign data w.r.t. the recorded timestamp, and calculate the performance for the below metrics. Note that we set the evaluation budgets for each model as 1/2, 1/4, 1/8, 1/16, 1/32 of the total budget in the whole test dataset. The similar evaluation setting has been widely adopted in online advertising researches [21, 22, 35]. The **number of conversions** is the total number of the achieved conversions. **Profit** is the total gains, i.e., total value of the obtained conversions. The other two metrics are **CVR** (conversion rate) and **CPA** (cost per conversion action). CVR is the ratio of the converted sequences among all the touched user impression sequences which reflects the ratio of gain for the ad delivery. And CPA is the cost averaged by the obtained conversion numbers which mesures the efficiency for the ad campaign. Note that only Criteo dataset contains the cost information so that our second stage experiments is conducted on Criteo dataset.

4.3 Compared Settings

In this section, we discuss the compared baselines and our model settings. We compare four baseline models with our dual-attention model. We also discuss the advantages of our dual-attention mechanism against the normal RNN model with single attention mechanism. Note that the click label ground truth z has been included in the input feature $\boldsymbol{x}$ in the other baseline models, except for our proposed model which utilizes click as the prediction label, for equally comparison.

- **LR** is the Logistic Regression model proposed in [24] and the attribution is calculated as the learned coefficient values of the regression model parameter for each channel.
- **SP** is a Simple Probabilistic model whose idea is derived from [7] and the conversion rate of each user action sequence is calculated as in [37] that

$$p(y = 1|\{c_j\}_{j=1}^{m_i}) = 1 - \prod_{j}^{m_i}(1 - \Pr(y = 1|c_j = k)), \qquad (17)$$

  where $\Pr(y = 1|c_j = k)$ is the conversion probability from the observed data w.r.t. the kth channel.
- **AH** (AdditiveHazard) model is the first work [37] using survival analysis and additive hazard function of conversion with the consideration of the touch point time to predict the final conversion rate. More details could be found in the paper.
- **AMTA** is the Additional Multi-touch Attribution model proposed in [12] which was state-of-the-art for this conversion attribution problem. It applies survival analysis to model the conversion estimation and utilizes the hazard rate of conversion at the specific time to model the conversion attribution.
- **ARNN** is the normal Recurrent Neural Network (i.e., only encoder part) method with the single Attention merely based on impression-level patterns to model the conversion attribution that $\hat{y}_i = r'(\boldsymbol{c}^{i2v})$, rather than sequence-to-sequence modeling in Eq. (4). This model is to illustrate the advantage of our dual-attention mechanism for data sparsity problem and multi-view learning schema.
- **DARNN** is our proposed model with dual-attention mechanism, which has been described in Section 3.

All the deep models are trained separately over one NVIDIA GeForce GTX 1080 Ti with Intel Core i7 processor for five hours. The detailed hyperparameter settings have been described in our published code, including learning rate, feature embedding size, hidden state size in RNN cell, etc.

Table 1: Conversion estimation results on two datasets. AUC: the higher the better; Log-loss: the lower the better.

	Miaozhen		Criteo	
Models	AUC	Log-loss	AUC	Log-loss
LR	0.8418	0.3496	0.9286	0.3981
SP	0.7739	0.5617	0.6718	0.5535
AH	0.8693	0.6791	0.6791	0.5067
AMTA	0.8357	0.1636	0.8465	0.3897
ARNN	0.8914	0.1610	0.9793	0.1850
DARNN	**0.9123**	**0.1095**	**0.9799**	**0.1591**

4.4 Conversion Estimation Performance

Our first evaluation is to measure the performance of user conversion estimation. Table 1 presents the detailed evaluation results under different models. From the statistics in the table, our model outperforms other baselines under both evaluation metrics. The results also reflect the other findings as below. (i) Both of the attention-based methods, i.e., DARNN and ARNN, achieve much better performance for sequential prediction than other compared models, which reflects the great pattern mining capability of deep neural networks. (ii) The exciting point process based methods AH and AMTA has poor classification performance for the conversion estimation. The reason is that they are designed to model the additive hazard ratio of conversion for each touch point. Though they learn the conversion prediction for the whole sequence, they do not consider much of the sequential patterns within the user behavior sequence. (iii) For the log-loss metric, the baselines get relatively higher (i.e., poorer) values than the deep models, which reflects that these baselines predict the conversion probability with totally large or small absolute values. Note that, however, AUC has no relationship to the direct output value of the model but considers the pairwise ranking performance. So almost all the baselines get considerably acceptable AUC results.

As for the learning procedure, since our proposed DARNN model captures both impression-level and click-level patterns and optimizes under two types of losses as that in Eqs. (5) and (7), the

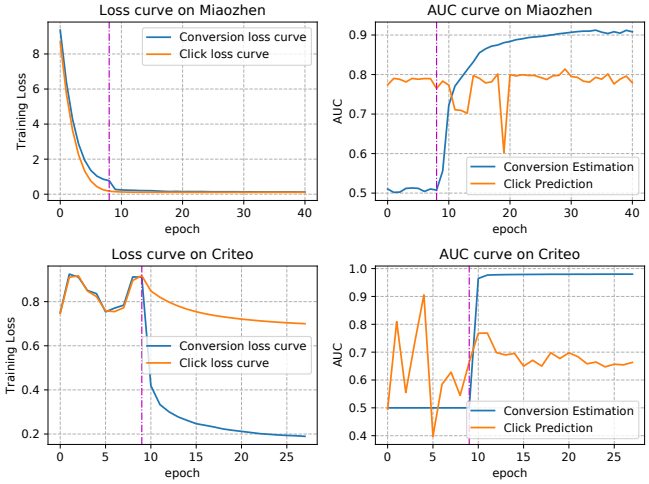

Figure 5: Learning curves on two datasets. Here one "epoch" means one whole iteration over the train dataset. The vertical purple line means the conversion estimation optimization starts in the second training stage.

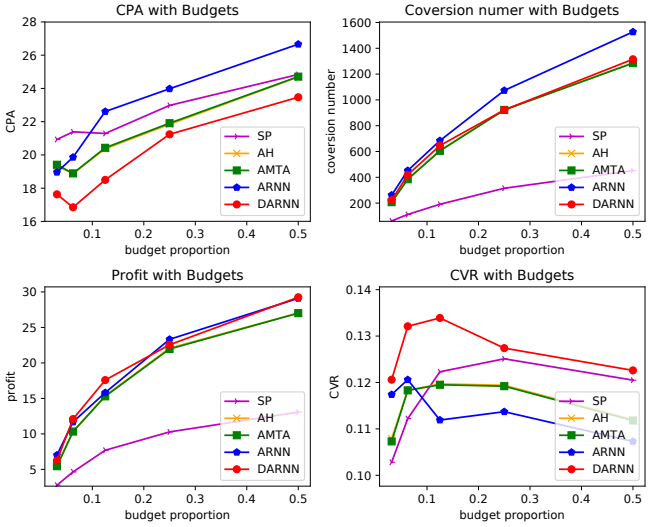

Figure 6: Performance with budgets on Criteo.

training procedure of our DARNN model generates two learning curves in Figure 5. In model training, we firstly make the model learn the click patterns, i.e., only optimize under the sequential click prediction loss as that in Eq. (5) and then, after the convergence of the first objective, we turn on the conversion estimation training, i.e., optimize under both two losses till convergence of the conversion loss. The convergence of each objective is defined as two successive rising of the optimization loss. The reason of two-stage training is to stabilize the model optimization under these two learning objectives. The similar training procedure has been studied in multi-task learning for recommender systems [5].

We may easily find from Figure 5 that our model not only optimizes the sequential click prediction, but also learns the conversion estimation. Moreover, the two learning objectives have been alternatively optimized to convergence at the second stage. Both the click prediction and the conversion estimation achieve excellent prediction performance.

4.5 Attribution Guided Budget Allocation

In the second stage of the experiments, we evaluate the effectiveness of different conversion attribution models for budget allocation. After replaying the historic touch points along the ordered timestamps, we calculated the total costs and the obtained conversion numbers of the compared model settings. To calculate the obtained profits for each model, we make the conversion value $V(y_i = 1)$ in Eq. (15) as eCPA (effective cost-per-action) which is constant for each model and calculated as $V(y_i = 1) = \text{eCPA}_{\text{train}} =$ (total cost/conversion number) in the training data. The detailed results are presented in Table 2 and Figure 6. As is presented in the table, since LR performs quite poor, we eliminate LR results in the figure for better illustration. Moreover, note that, we also compared simple *last-touch* attribution method in the second-stage experiment. We did not report this heuristic method in our experiments since the result showed that AH baseline model performed almost the same as the last-touch attribution method which is quite interesting and needs further investigations in the future work.

From Table 2 and Figure 6, we may find that: (i) As the budget increases, all the models spend more to earn each user action, i.e., the CPA value of each model is increasing, which is reasonable. (ii) Both of the attention-based neural network models, i.e., ARNN and DARNNs, achieve relatively better performance compared with the other models over all the evaluation metrics. The reason is probably the sequential pattern mining of these two models. (iii) The two baselines AMTA and AH achieve very similar performance, which is probably accounted for the similar idea of the additional conversion probability modeling within their models. (iv) DARNN model achieves the best performance under CPA and CVR, which reflects the effectiveness our learned attribution values of dual-attention mechanism. Moreover, this result also shows the advantage of the dual attention mechanism over single attention model ARNN. (v) ARNN spends money more aggressive than other models thus getting poor CPA result. The reason may be that its attribution is based merely on impression-level and the pattern captured tends to long-term investment on the user behavior. However, our DARNN model spends the budget more economically which leads to more efficient budget pacing, i.e., lower CPA. This indicates that combining both impression-level and click-level attention will take advantages of both long-term (impression to conversion) and short-term (click to conversion) behavior patterns.

4.6 Comprehensive Analysis

In this part, we look deeper into the learned attribution model. We first discuss the calculated attribute credits over both touch point level and channel level, and then analyze the results of the learned weighting parameter λ according to Eq. (12) which controls the influence of the two types of user actions, i.e., impressions and clicks.

First, we illustrate the touch point attribution in Miaozhen dataset which calculates the averaged attribution credits over all the sequence samples with fixed sequence length, on each touch point. Specifically, here the credits on the j^{th} touch point is averaged over all the converted sequences with the fixed sequence length m as $\overline{\text{Attr}}_j = \frac{1}{N_m} \sum_{i=1}^{N_m} y_i \cdot \text{Attr}_{ij}$, where N_m is the total number of the

Table 2: Budget allocation evaluation results. CPA: the lower the better; Profit & CVR: the higher the better.

Models	CPA					Profit					Conversion Num.					CVR				
	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{16}$	$\frac{1}{32}$
LR	31.79	29.47	29.77	27.83	27.46	8.022	6.938	4.386	3.238	1.954	576	427	275	181	107	0.0928	0.0910	0.0873	0.0827	0.0748
SP	24.84	22.98	21.29	21.39	20.93	13.07	10.28	7.694	4.648	2.776	452	315	191	112,	62	0.1205	0.1251	0.1223	0.1122	0.1028
AH	24.69	21.84	20.37	18.89	19.32	27.03	22.08	15.38	10.32	5.491	1286	925	607	385	208	0.1120	0.1194	0.1197	0.1183	0.1079
AMTA	24.71	21.91	20.43	18.89	19.41	27.01	21.96	15.29	10.33	5.446	1285	922	605	385	207	0.1118	0.1192	0.1195	0.1183	0.1073
ARNN	26.66	23.98	22.61	19.86	18.96	29.10	**23.32**	15.81	11.68	**7.010**	**1527**	**1073**	**684**	**452**	**262**	0.1073	0.1137	0.1119	0.1206	0.1174
DARNN	**23.47**	**21.24**	**18.50**	**16.85**	**17.63**	**29.25**	22.56	**17.58**	**12.09**	6.26	1315	922	646	419	223	**0.1226**	**0.1274**	**0.1339**	**0.1321**	**0.1206**

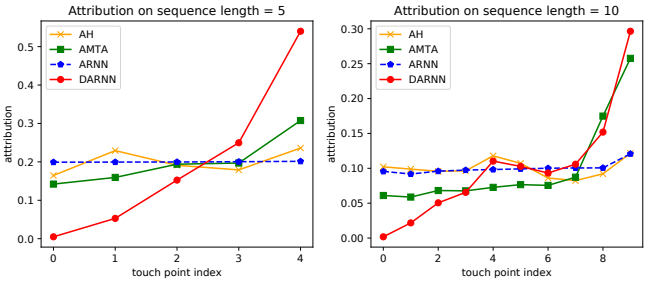

Figure 7: Touch point level attribution statistics (Miaozhen).

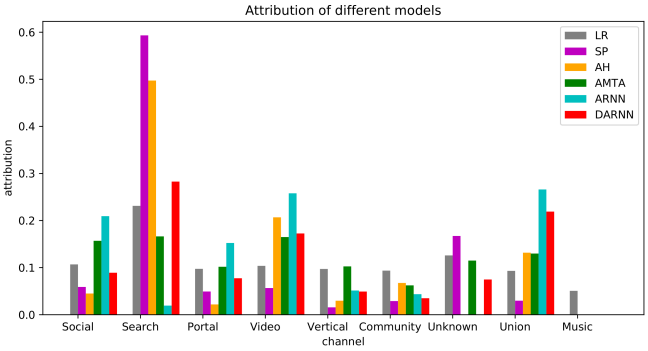

Figure 8: Attribution of different channels on Miaozhen.

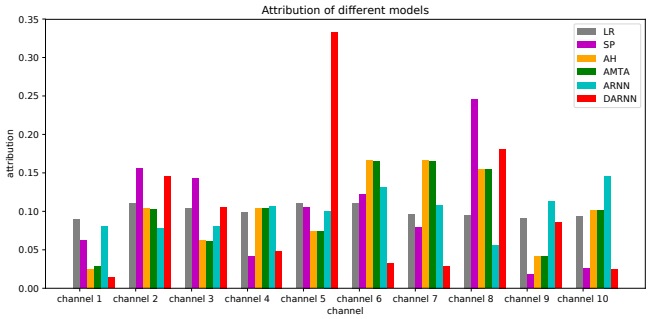

Figure 9: Attribution distribution over channels on Criteo.

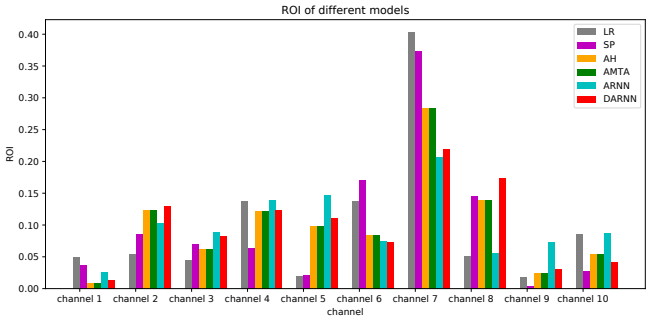

Figure 10: ROI distribution over channels on Criteo.

sequences with length m and y_i is the conversion indicator of the sequence sample.

Figure 7 illustrates the touch point conversion attribution results on the sequences with length of 5 and 10 respectively. Since LR and SP calculate the attribution based on different channels rather than each touch point, so we cannot get the specific result of these two models at the touch point level. From the figure we may find that the credits attributed on each touch point varies over different models. When sequence length is relatively short, DARNN learns that the touch point closer to the final touch may more likely derive the final conversion. In longer sequences (with length of 10), our DARNN model place higher credits for the touch points in the middle process while the attribution drops a little later and consequently rises to much higher when final conversion approaches. This phenomenon is reasonable since it is not always correct about endless ad delivery for the user and, moreover, it reflects the tradeoff between the ad effectiveness and the user experience of the Internet service. However, ARNN seems to "average" the credits over all the touch points within the sequence. Note that ARNN only concerns impression-level contributions, which in contrast shows the great effects of click-level patterns in our proposed dual-attention mechanism on the final conversion attribution.

The next analysis is based on the channel level attribution distribution. We illustrate the calculated attribution credits of converted sequences over multiple channels in Miaozhen dataset as that in Figure 8. The horizontal axis is the channel information varying from social media to music platforms. Since there is no conversions on music channel, no credit has been assigned by the models except LR which takes the learned parameter coefficient of the channel feature for conversion attribution. From the illustration we can find that (i) LR, SP, AH and DARNN models assign the highest attribution credits to search channel, while ARNN and AMTA attribute the most onto video channel. (ii) SP and AH assign relatively much higher credits on search channel, while the other models distribute attribution more smoothly. (iii) Vertical and community channels have low credits while union channel has much higher attribution under attention-based models. From these findings we find the significance of the replay evaluation for attribution guided budget allocation in the second stage of experiments, since the calculated conversion attribution credits over different channels vary from different models.

In addition, we illustrate the attribution credits on Criteo dataset in Figure 9 and the corresponding ROI_{c_k} of each channel c_k in Figure 10 which is calculated according to Eq. (15) under different

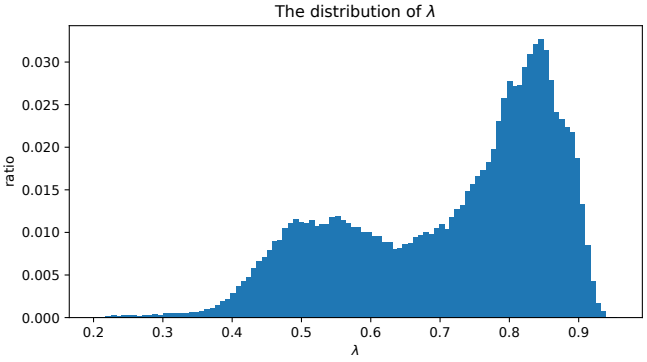

Figure 11: The distribution of λ over Criteo dataset.

models. From channel level, our DARNN model assigns the highest credits onto channel 5. However, the ROI calculation derives that all models allocate the most budget credits onto channel 7. This is reasonable since the ROI is based on both channel level and touch point level information as that in Eq. (15).

Finally, in Figure 11, we statistically visualize the value distribution of λ which controls the impact from impression-level and click-level patterns, respectively. As is calculated in Eq. (12), note that, when λ gets larger the click-level patterns get higher impact on the conversion attribution. We may find from the figure that the click-level patterns relatively contribute more to the final conversion estimation, which reflects that the effectiveness of our dual-attention mechanism for different action pattern mining, as is described in Sec. 3.3. Generally speaking, the results illustrate the importance of combining both impression pattern and click pattern through dual-attention mechanism, especially that the click-level patterns contribute better under tight budget cases.

5 CONCLUSION AND FUTURE WORK

In this paper, we proposed a dual-attention recurrent neural network model for learning to assign conversion credits over the ad touch point sequences. Our model not only captures sequential user patterns, but also pays attention to both impression-level and click-level user actions and derives an effective conversion attribution methodology. The experiments show the significant improvement over the other state-of-the-art baselines.

One of the limitations of this work is that we have not taken the cost of ad impressions into account in the attention mechanism. It is of great interest to take the cost factor into modeling and improve the cost-effectiveness performance in the future as that in the works [21, 22] of real-time auction advertising.

REFERENCES

[1] Deepak Agarwal, Souvik Ghosh, Kai Wei, and Siyu You. 2014. Budget pacing for targeted online advertisements at LinkedIn. In *KDD*. ACM, 1613–1619.

[2] Kareem Amin, Michael Kearns, Peter Key, and Anton Schwaighofer. 2012. Budget optimization for sponsored search: Censored learning in MDPs. *UAI* (2012).

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[4] Ron Berman. 2017. Beyond the last touch: Attribution in online advertising. (2017).

[5] Xuezhi Cao, Haokun Chen, Xuejian Wang, Weinan Zhang, and Yong Yu. 2018. Neural Link Prediction over Aligned Networks. In *AAAI*.

[6] John Chandler-Pepelnjak. 2009. Measuring roi beyond the last ad. *Atlas Institute Digital Marketing Insight* (2009), 1–6.

[7] Brian Dalessandro, Claudia Perlich, Ori Stitelman, and Foster Provost. 2012. Causally motivated attribution for online advertising. In *ADKDD*. ACM, 7.

[8] Eustache Diemert, Julien Meynet, Pierre Galland, and Damien Lefortier. 2017. Attribution Modeling Increases Efficiency of Bidding in Display Advertising. In *ADKDD*. ACM.

[9] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122* (2017).

[10] Sahin Cem Geyik, Abhishek Saxena, and Ali Dasdan. 2014. Multi-touch attribution based budget allocation in online advertising. In *ADKDD*. ACM, 1–9.

[11] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* (1997).

[12] Wendi Ji and Xiaoling Wang. 2017. Additional Multi-Touch Attribution for Online Advertising. In *AAAI*.

[13] Wendi Ji, Xiaoling Wang, and Dell Zhang. 2016. A probabilistic multi-touch attribution model for online advertising. In *CIKM*. ACM.

[14] Kuang-Chih Lee, Ali Jalali, and Ali Dasdan. 2013. Real time bid optimization with smooth budget delivery in online advertising. In *ADKDD*. ACM, 1.

[15] Kuang-chih Lee, Burkay Orten, Ali Dasdan, and Wentong Li. 2012. Estimating conversion rate in display advertising from past performance data. In *KDD*. ACM, 768–776.

[16] 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. *SIGIR* (2018).

[17] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *KDD*.

[18] Claudia Perlich, Brian Dalessandro, Rod Hook, Ori Stitelman, Troy Raeder, and Foster Provost. 2012. Bid optimizing and inventory scoring in targeted online advertising. In *KDD*. ACM, 804–812.

[19] Yao Qin, Dongjin Song, Haifeng Cheng, Wei Cheng, Guofei Jiang, and Garrison Cottrell. 2017. A dual-stage attention-based recurrent neural network for time series prediction. *arXiv preprint arXiv:1704.02971* (2017).

[20] Yanru Qu, Han Cai, Kan Ren, Weinan Zhang, Yong Yu, Ying Wen, and Jun Wang. 2016. Product-based neural networks for user response prediction. In *ICDM*.

[21] Kan Ren, Weinan Zhang, Ke Chang, Yifei Rong, Yong Yu, and Jun Wang. 2018. Bidding Machine: Learning to Bid for Directly Optimizing Profits in Display Advertising. *TKDE* (2018).

[22] Kan Ren, Weinan Zhang, Yifei Rong, Haifeng Zhang, Yong Yu, and Jun Wang. 2016. User response learning for directly optimizing campaign performance in display advertising. In *CIKM*. ACM, 679–688.

[23] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *WWW*.

[24] Xuhui Shao and Lexin Li. 2011. Data-driven multi-touch attribution models. In *KDD*. ACM, 258–264.

[25] Ritwik Sinha, Shiv Saini, and N Anadhavelu. 2014. Estimating the incremental effects of interactions for marketing attribution. In *Behavior, Economic and Social Computing (BESC), 2014 International Conference on*. IEEE, 1–6.

[26] Jun Song, Jun Xiao, Fei Wu, Haishan Wu, Tong Zhang, Zhongfei Zhang, and Wenwu Zhu. 2017. Hierarchical Contextual Attention Recurrent Neural Network for Map Query Suggestion. *TKDE* (2017).

[27] Jun Wang, Weinan Zhang, and Shuai Yuan. 2017. Display Advertising with Real-Time Bidding (RTB) and Behavioural Targeting. *Now Publisher* (2017).

[28] Xuejian Wang, Lantao Yu, Kan Ren, Guanyu Tao, Weinan Zhang, Yong Yu, and Jun Wang. 2017. Dynamic attention deep model for article recommendation by learning human editors' demonstration. In *KDD*. ACM.

[29] Jian Xu, Xuhui Shao, Jianjie Ma, Kuang-chih Lee, Hang Qi, and Quan Lu. 2016. Lift-based bidding in ad selection. In *AAAI*.

[30] Lizhen Xu, Jason A Duan, and Andrew Whinston. 2014. Path to purchase: A mutually exciting point process model for online advertising and conversion. *Management Science* 60, 6 (2014), 1392–1412.

[31] Junchi Yan, Chao Zhang, Hongyuan Zha, Min Gong, Changhua Sun, Jin Huang, Stephen Chu, and Xiaokang Yang. 2015. On machine learning towards predictive sales pipeline analytics. In *AAAI*.

[32] Shuai Yuan. 2015. *Supply side optimisation in online display advertising (Chapter 3)*. Ph.D. Dissertation. UCL (University College London).

[33] Shuangfei Zhai, Keng-hao Chang, Ruofei Zhang, and Zhongfei Mark Zhang. 2016. Deepintent: Learning attentions for online advertising with recurrent neural networks. In *KDD*. ACM.

[34] Weinan Zhang, Tianming Du, and Jun Wang. 2016. Deep Learning over Multi-field Categorical Data: A Case Study on User Response Prediction. *ECIR* (2016).

[35] Weinan Zhang, Shuai Yuan, and Jun Wang. 2014. Optimal real-time bidding for display advertising. In *KDD*. ACM, 1077–1086.

[36] Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. 2014. Sequential click prediction for sponsored search with recurrent neural networks. *arXiv preprint arXiv:1404.5772* (2014).

[37] Ya Zhang, Yi Wei, and Jianbiao Ren. 2014. Multi-touch attribution in online advertising with survival theory. In *ICDM*. IEEE, 687–696.