Understanding Consumer Journey using Attention based Recurrent Neural Networks

Yichao Zhou*
University of California, Los Angeles
yz@cs.ucla.edu

Shaunak Mishra
Yahoo Research
shaunakm@verizonmedia.com

Jelena Gligorijevic
Yahoo Research
jelenas@verizonmedia.com

Tarun Bhatia
Yahoo Research
tarunb@verizonmedia.com

Narayan Bhamidipati
Yahoo Research
narayanb@verizonmedia.com

ABSTRACT

Paths of online users towards a purchase event (conversion) can be very complex, and guiding them through their journey is an integral part of online advertising. Studies in marketing indicate that a conversion event is typically preceded by one or more purchase funnel stages, viz., unaware, aware, interest, consideration, and intent. Intuitively, some online activities, including web searches, site visits and ad interactions, can serve as markers for the user's funnel stage. Identifying such markers can potentially refine conversion prediction, guide the design of ad creatives (text and images), and lead to higher ad effectiveness. We explore this hypothesis through a set of experiments designed for two tasks: (i) conversion prediction given a user's activity trail, and (ii) funnel stage specific targeting and creatives. To address challenges in the two tasks, we propose an attention based recurrent neural network (RNN) which ingests a user activity trail, and predicts the user's conversion probability along with attention weights for each activity (analogous to its position in the funnel). Specifically, we propose novel attention mechanisms, which maintain a global weight for each activity across all user trails, and also indicate the activity's funnel stage. Use of the proposed attention mechanisms for the first task of conversion prediction shows significant AUC lifts of 0.9% on a public dataset (RecSys 2015 challenge), and up to 3.6% on three proprietary datasets from a major advertising platform (Yahoo Gemini). To address the second task, the activity weights from the proposed mechanisms are used to automatically assign users to funnel stages via a scalable scoring method. Offline evaluation shows that such activity weights are more aligned with editorially tagged activity-funnel stages compared to weights from existing attention mechanisms and simpler conversion models like logistic regression. In addition, results of online ad campaigns in Yahoo Gemini with funnel specific user targeting and ad creatives show strong performance lifts further validating the connection across online activities, purchase funnel stages, stage-specific custom creatives, and conversions.

*Work done while the author was at Yahoo Research.

KDD '19, August 4–8, 2019, Anchorage, AK, USA

ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330753

CCS CONCEPTS

• **Information systems → Online advertising**;

ACM Reference Format:
Yichao Zhou, Shaunak Mishra, Jelena Gligorijevic, Tarun Bhatia, and Narayan Bhamidipati. 2019. Understanding Consumer Journey using Attention based Recurrent Neural Networks. In *The 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19), August 4–8, 2019, Anchorage, AK, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3292500.3330753

1 INTRODUCTION

Studies in marketing [7, 13] strongly indicate the existence of a purchase funnel, *i.e.*, a consumer (user) may go through multiple stages before finalizing a purchase (conversion). A typical purchase funnel has the following stages before a purchase event: unaware, aware, interest, consideration and intent. Users in some of these funnel stages (*i.e.*, consideration, and intent) may have a stronger likelihood of conversion than others (*i.e.*, unaware, aware, and interest). To influence users towards conversion, online advertising platforms need to: (i) predict a user's conversion probability [3, 16], and (ii) expose users to appropriate ads. Intuitively, understanding the purchase funnel for an advertiser, *i.e.*, the kind of online activities users perform in each funnel stage, can not only improve conversion prediction, but also guide the design of custom ad creatives for users in each funnel stage.

Given the intuitive benefits of understanding the purchase funnel, we consider observable online activities (*i.e.*, search queries, site visits, online article views, and ad interactions) which online users perform before converting on an advertiser. In particular, given a trail (sequence) of relevant online activities that a user has performed, we want to estimate the user's funnel stage for a given advertiser in an interpretable and scalable manner (vis-a-vis activities, trail length, and advertisers). Since there is no available groundtruth in this context[1], and the activity trails could be arbitrarily complicated (*e.g.*, having loops), we simplify our task to understanding a single activity's position in the funnel. We explain and justify this simplification in further detail below using the example of a theme park.

Illustrative example: Consider a theme park, and a trail of a (relevant) user activities as shown in the left half of Figure 1. In general, users might perform a subset of activities listed on the left half,

[1]To the best of our knowledge, there are no datasets which have funnel stage labels for a trail of online activities including search, site visits, content views and ad interactions relevant to a given advertiser.

and they might do it in any arbitrary order with repetitions. A human editor can parse the activity trail in Figure 1, and assign (tag) activities to the theme park's purchase funnel stages as shown in Figure 1. The editor may have the following thought process while tagging activities with funnel stages.

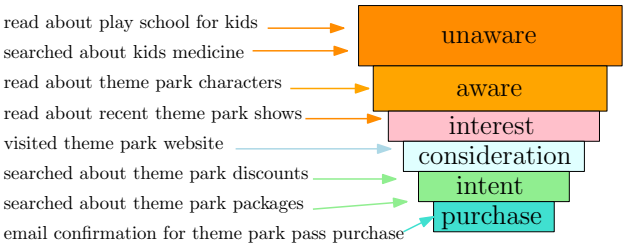

Figure 1: Example showing an assignment of user activities to funnel stages for a theme park visit.

- Reading and searching about kids does not indicate the user's awareness about the theme park, but having kids indicates a higher propensity to purchase theme park tickets; hence such activities may be used to identify *unaware* users. In comparison, reading about theme park characters shows that the user is *aware* about the theme park.
- Reading about recent theme park shows indicates the user's potential *interest* in visiting the theme park. Visiting the theme park website indicates further *consideration*, and searching about discounts indicates stronger *intent*.

The above example illustrates that activities can serve as markers for a user's funnel stage. Intuitively, such markers can help towards both conversion prediction (*e.g.*, users doing intent stage activities have a higher probability of conversion), and guiding custom creative design. As an example of funnel specific creatives, users in the *unaware* stage may be shown an ad to increase their awareness about the park (with clicks landing them on introductory videos and blogs). Whereas, users in the intent stage may be shown deals landing them directly on the park's purchase webpage. Given such plausible benefits, in this paper, we focus on leveraging the activity-funnel connection for two tasks: (i) improving upon state-of-the-art conversion prediction models, and (ii) testing funnel specific creatives for an advertiser.

In the case of (i), existing state-of-the-art conversion models for online advertising are either sequential (*e.g.*, RNN with attention [23]) or non-sequential (*e.g.*, deep residual networks [18]). In principle, sequential models are better suited for predicting outcomes (*e.g.*, conversion) given a sequence of user activities [23]. However, existing sequential models miss out on an explicit mechanism to identify activities representative of funnel stages[2], and refine their predictions accordingly. We address this deficiency, by proposing two *global* attention mechanisms for RNN based conversion prediction; in particular, the proposed attention mechanisms learn a weight corresponding to the activity's position in the purchase funnel.

[2]In the conversion prediction task, there is no existing data set to infer funnel tags for activities relevant to an advertiser. The conversion model is trained on trails of user activities with a binary label indicating conversion or no-conversion.

In the case of (ii), although there is no prior work (to the best of our knowledge) on data driven design of funnel stage specific creatives, there has been prior work on assigning users directly to a purchase funnel stage using hidden Markov models (HMMs) ingesting user activity trails [1, 9]. Such HMMs do not require any activity-funnel understanding, but at the same time it is non-trivial to infer (from the trained HMM) which activities are tied to a particular funnel stage; hence, they miss out on valuable (activity) insights for the advertiser as well as guidelines for creative design (we describe additional shortcomings of HMM approaches in Section 3 on our proposed architecture). Therefore, we take the route of inferring activity-funnel tags automatically from a trained sequential conversion model (from RNN attention weights in particular). Using such funnel tags, we identify users in various (user) funnel stages via a deterministic (scoring) logic applied to the user activity trails with activity-funnel tags for each activity in their trail. The identified users can then be exposed to (user) funnel specific creatives based on activities assigned to each funnel stage. Our user scoring logic is scalable in terms of the number of activities, trail length and the number of advertisers.

In summary, our main contributions can be listed as follows:

(1) We propose two global attention mechanisms for RNN based conversion prediction which significantly outperform existing attention based baselines on a public data set (RecSys 2015 challenge, 0.9% AUC lift) as well as datasets for three advertisers from a major advertising platform, *i.e.*, Yahoo Gemini (up to 3.6% AUC lift).
(2) We use the RNN activity attention scores (from the conversion prediction model) to assign funnel stages to activities. Although a heuristic, the automatically assigned funnel tags are very close to editorial tags. The proposed attention mechanisms achieve the best accuracy in terms of matching editorial tags (measured by RMSE and NDCG).
(3) Automatically tagged activities are consumed in a deterministic user scoring logic (user funnel assignment for tagged activity trails) to identify a user's funnel stage, and to show the user funnel specific creatives. Offline user scoring metrics also show the superiority of the funnel tags inferred from the proposed attention mechanisms. In addition, our experiments involving online ad campaigns in Yahoo Gemini with funnel specific ads show significant incremental conversion rate (3%-6%), as well as lift in click-through-rate (CTR) and reduced user acquisition costs (cost-per-action) compared to conventional campaigns. We also share a few non-trivial activity insights (anonymized by advertiser domain) which we discovered during our experiments.

The remainder of this paper is organized as follows. In Section 2, we describe related work, followed by Section 3 on our overall architecture. We then cover the conversion prediction setup, and proposed attention mechanisms in Section 4. This is followed by Section 5 on user scoring (*i.e.*, assigning funnel stages to users) and Section 6 on creative design. The experimental results on both conversion prediction and funnel specific ad targeting are covered in Section 7. Section 8 covers the conclusion, and we provide a reproducibility supplement at the end (Section 9).

2 BACKGROUND

In this section, we first give a brief overview of online advertising. This is followed by related work on ad click and conversion prediction models, purchase funnel modeling in advertising, and attention mechanism in RNNs.

Online advertising. In a regular online advertising setup [3, 16], advertisers sign up with ad platforms (*e.g.*, Google Ads, Facebook Ads, Yahoo Gemini) and launch campaigns to show ads on online properties associated with the ad platform. Advertisers typically create one or more creatives to target relevant audience (*i.e.*, ad groups) and for each ad group they specify a bid. The bid is essentially the maximum amount they are willing to pay for a certain action (*e.g.*, ad click). At each ad serving opportunity, an auction is run by the ad platform as follows. Bids from multiple advertisers (campaigns) are ranked, and the one with the highest bid is chosen for the ad serving opportunity, eventually leading to an ad impression. Based on the impressions that an advertiser wins in the auction, the advertiser typically cares about include click-through-rate (CTR), conversion rate (adCVR), and cost-per-action CPA [3] where:

$$CTR = \frac{total\ clicks}{total\ impressions}, \quad adCVR = \frac{total\ conversions}{total\ clicks}, \quad (1)$$

$$CPA = \frac{total\ spend}{total\ conversions}. \quad (2)$$

Click and conversion prediction models: Click and conversion prediction models play an important role in auctions for ads, and are crucial for advertisers to target relevant users (*i.e.*, users who are more likely to click and convert) [3, 11]. In large scale advertising setups (*e.g.*, platforms owned by Google, Facebook, and Yahoo), logistic regression (LR) models have been successfully used [3, 16]. Recently, more sophisticated deep learning models have also been used for CTR and CVR prediction, *e.g.*, deep residual networks [18] and deep sequential models [8, 10, 23]; the sequential models perform significantly better than their non-sequential counterparts.

Purchase funnel modeling: Prior work on purchase funnel modeling in the context of online advertising focuses on assigning users directly to a purchase funnel stage using hidden Markov models (HMMs) ingesting user activity trails [1, 9]. In [1], the focus is on multi-touch ad attribution using the purchase funnel stages, while the work in [9] is mainly on the negative impact of ads (annoyance) vis-a-vis user funnel stages. In theory, it is also possible to use conversion models to cluster users [20] (*e.g.,* by RNN hidden layer representations), and assign funnel stages to user clusters. However, such approaches (HMM and user clustering in general) miss out on explicitly infering the connection between activities and funnel stages, which is a major focus in this paper.

Attention Mechanism: Attention mechanism for RNNs was first introduced in the context of neural machine translation [2]. The mechanism automatically detected a linguistically plausible alignment between a source sentence and the corresponding target sentence. Following this, [15] proposed global and local attention mechanisms. The global approach looked at the entire sentence, but the local one focused on a subset of the sequence when performing the neural translation. The ensemble attention model in [15] yielded the state-of-the-art result in the WMT'15 English-to-German translation task. However, both the local and global approaches ignored the overall significance of each element in the sequence (which we focus on in this paper). In [21], a hierarchical attention model for RNN was proposed, which enhanced document classification performance by using both character-level and word-level features. [5] proposed a long short-term memory (LSTM) network to tackle the machine reading problem. In their work, a self-attention mechanism was proposed to draw the correlation between each word and the previous set of words. In a more recent work, [19] proposed a multi-head self-attention mechanism to get rid of recurrent neural units, and to allow self-attention encode the context information for each word of the sentence. Recently, the attention mechanism has also been applied in many other applications in advertising, such as conversion prediction [17], click prediction [24], and search advertising [22]. In this paper, we propose two variants of the attention mechanism for RNNs; they are inspired by the purchase funnel, and enable the inference of a user activity's position in the funnel.

3 AD TARGETING ARCHITECTURE

In this section, we give an overview of our proposed architecture for targeting users with funnel stage specific ads, and also justify the need for activity-funnel understanding. Given an advertiser, the high level goal is to understand which activities can serve as markers for a user's funnel stage, and leverage such understanding to show users ads customized for a funnel stage. We break down this goal into sub-tasks as follows.

(1) Activity selection: automatic selection of activities relevant to the advertiser's conversion event leading to *seed list* $\mathcal{A}_{ADV}$.
(2) Activity → funnel tagging: automatic assignment of a funnel stage to each activity in $\mathcal{A}_{ADV}$ without human labeling.
(3) Activity trail → funnel mapping (user scoring): automatic assignment of user trails (of tagged activities) to funnel stages.
(4) Creative design and targeting: editorial ad customization using funnel tags in the seed list, and targeting users in a funnel stage with such custom ads.

For the task of activity-funnel tagging listed above, we focus on using activity attention weights from a trained RNN based conversion model. The intuition here is that activities with higher attention weights (*i.e.*, with higher influence towards a conversion event) may be representative of funnel stages closer to conversion. Exploring this heuristic for multiple advertisers, and validating its accuracy via intrinsic evaluations (*i.e.*, alignment with editorial tags) as well as extrinsic evaluations (user scoring and ad campaign metrics) is one of the major contributions in this paper. Figure 2 gives a block level overview of our architecture, and shows how activity-funnel tagging fits in the system. We describe below the detailed

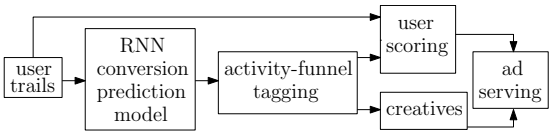

Figure 2: Overview of the entire ad targeting system.

justification behind focusing on activity-funnel tagging.

Justification for activity-funnel tagging: We introduce activity-funnel tagging as an intermediate sub-task, as opposed to directly mapping user trails to funnel stages (as in HMM based formulations [1, 9]). Our approach is motivated by the following reasons.

- It is relatively easier and faster to generate an editorially labeled data set (via domain experts) with activity-funnel tags than generating a data set with user trail-funnel tags. Such a dataset can be used to verify the accuracy of tags vis-a-vis human judgement.
- Training and online scoring (user stage inference) of HMMs may not be scalable with respect to number of activities, trail lengths, and the number of advertisers (since we need a model for each advertiser). Activity-funnel tags learnt offline (from a trained RNN) on the other hand, can be used with a scalable scoring logic suitable for low latency ads serving.
- Activity-funnel tags can lead to valuable insights for advertisers, and guidelines for creative design. In additon, insightful search queries (activities) with such funnel tags, can be directly used in conjunction with search retargeting ad campaigns for advertisers (who typically bid for obvious search keywords with their product/brand's name). In comparison, reverse engineering a trained HMM to infer such activity-funnel insights may not be straightforward, and is an independent exercise beyond the scope of this paper.

4 ACTIVITY-FUNNEL TAGGING VIA CONVERSION PREDICTION

In this section, we first explain the conversion prediction setup in Section 4.1 (including seed list selection for an advertiser). This is followed by Section 4.2 which formally describes gated recurrent unit (GRU) based RNNs with existing (local) attention mechanism; we explain how this existing mechanism can be used for conversion prediction and activity-funnel tagging. Next, in Sections 4.3 and 4.4, we introduce two (global) attention mechanisms, and explain how they are designed towards better conversion prediction, and activity-funnel tagging.

4.1 Conversion prediction setup

Consider a user activity trail (sequence) $\mathbf{a} = [a_1, a_2, \ldots, a_l]$, where the online activities $a_t \in \mathcal{A}_{ADV}$ are in chronological order, $\mathcal{A}_{ADV}$ is the set of relevant online activities (*i.e.*, the seed list) for advertiser ADV, and l is the length of the trail. Seed list selection is similar to filter based methods for feature selection; we select only those activities whose conversion rate [3] exceeds a pre-determined threshold. Given such an activity trail $\mathbf{a}$ corresponding to a user, we are interested in the probability of the user converting on the advertiser ADV. In other words, the goal of sequential conversion prediction is to estimate the following probability:

$$\mathbb{P}\left(\text{conversion on } ADV|\mathbf{a}\right) = \mathbb{P}\left(y_{ADV} = 1|\mathbf{a}\right),$$

where y_{ADV} is a binary variable indicating conversion ($y_{ADV} = 1$) or no conversion ($y_{ADV} = 0$). We consider a separate conversion model for each advertiser in our setup, and hence the task of conversion prediction boils down to solving the above binary classification problem. In addition, we are also interested in extracting (from the conversion prediction model) a set of weights for activities in $\mathcal{A}_{ADV}$ such that the weights are indicative of the position (stage) of the activities in the purchase funnel. We will refer to this secondary goal as the activity-funnel inference problem in the remainder of the paper, and denote the set of extracted weights by $\boldsymbol{\beta} = \left[\beta_1, \beta_2, \ldots, \beta_{|\mathcal{A}_{ADV}|}\right]$.

In this paper, to address both conversion prediction and activity-funnel tagging, we focus on the use of RNN based conversion prediction models. In particular, we leverage existing RNN attention mechanisms, and propose two new attention mechanisms aimed at better conversion prediction, as well as activity-funnel inference (covered in the following subsections).

4.2 Local attention in GRU based RNN (LATT)

Figure 3 shows a bi-directional GRU based RNN model with (local) attention mechanism. As shown, the activity embedding layer, bi-directional GRU layer, and the local attention layer are the major components of the model; we describe below the details for each these components.

Input layer and activity embeddings: The activity sequence $\mathbf{a} = [a_1, \ldots, a_l]$, is fed as input to the activity embedding block, which learns a low dimensional activity representation $\mathbf{x}_t$ for each activity in $a_t \in \mathcal{A}_{ADV}$ (*i.e.*, $\mathbf{x}_t$ is the embedding for activity a_t). The randomly initialized embeddings are learnt as a part of the RNN model training.

Bi-directional GRU layer: As shown in Figure 3, the activity embedding sequence $[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l]$ is fed to a bi-directional GRU layer [6][4]. The final sequence representation is obtained by aggregating (concatenating) representations from the forward and backward GRU cell as: $\mathbf{h}_t = \left[\mathbf{h}_t^{(f)}, \mathbf{h}_t^{(b)}\right]$.

Local attention mechanism: The attention mechanism shown in Figure 3 was first introduced in [2] for the task of neural machine translation. We briefly discuss below this (local) attention mechanism before describing our proposed attention mechanisms in the following subsections. In this mechanism, an attention layer is added on top of the GRU-based RNN module, in order to distinguish the contribution of each output from the RNN layer towards final prediction. In particular, it first transforms the RNN outputs $\mathbf{h}_t$ (from the GRU cells) to a low dimensional representation $\mathbf{u}_t$ as follows:

$$\mathbf{u}_t = \tanh\left(\mathbf{W}_s\mathbf{h}_t + \mathbf{b}_s\right). \tag{3}$$

Then, it introduces a context vector $\mathbf{u}_s$ to be learnt during the RNN training process. In a sense, it measures how much attention should be given to each input representation $\mathbf{u}_t$. The attention layer calculates the inner product between $\mathbf{u}_t$ and $\mathbf{u}_s$, and normalizes

[3] The conversion rate for an activity a_0 is defined as the ratio of count of users who did a_0, and then converted on the advertiser within a time window (*e.g.*, 3 months) over the total count of users who did a_0. In addition, all *purchase* activities for ADV are known a priori, and excluded from the tagging process; the trails of all converters end with the activity prior to the purchase (conversion) activity.

[4] We choose GRU in our setup because its performance is at par with LSTM for many applications [12], and the model is much simpler. In addition, our proposed attention mechanisms can be used for LSTM based RNNs without loss of generality.

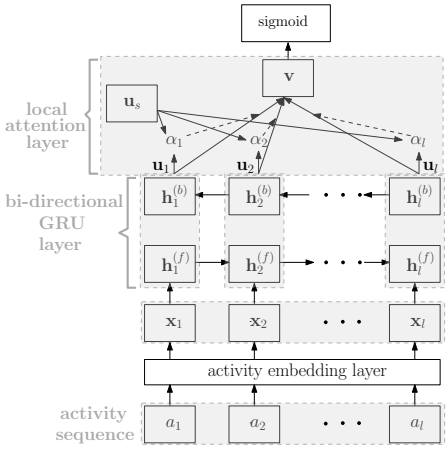

Figure 3: Bi-directional GRU based RNN model (with local attention mechanism) for conversion prediction.

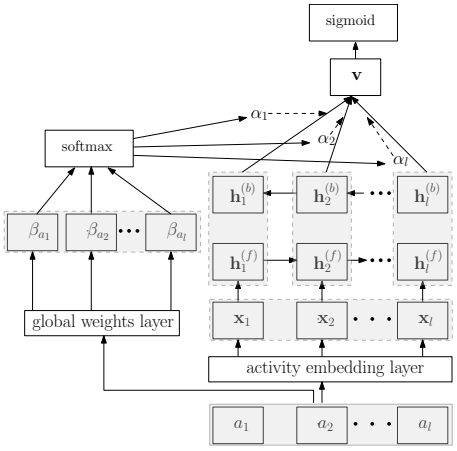

Figure 4: The global attention-based RNN model.

the attention weights through a softmax function as follows:

$$\alpha_t = \frac{\exp\left(\mathbf{u}_t^T \mathbf{u}_s\right)}{\sum_{t=1}^{l} \exp\left(\mathbf{u}_t^T \mathbf{u}_s\right)}, \tag{4}$$

where (normalized) attention weight $\alpha_t \in \mathbb{R}$. The attention layer outputs $\mathbf{v}$ which is the weighted summation of all the low-dimentional vectors $\mathbf{u}_t$'s as the latent representation of each activity sequence:

$$\mathbf{v} = \sum_{t=1}^{l} \alpha_t \mathbf{h}_t. \tag{5}$$

Since this attention mechanism only calculates the attention scores in terms of specific trails, we name this attention model as local attention mechanism (LATT), *i.e.*, it has attention *local* to a trail.

Activity-funnel tagging: With LATT, each activity may attain a wide range of attention weights across different user trails (since the weight for an activity in each user trail can be different). To get an aggregate LATT score (across all trails) for each activity, we consider two aggregation functions: (i) max (leading to $\boldsymbol{\beta}_{LATT-max}$), and (ii) average of all attention scores for an activity (leading to $\boldsymbol{\beta}_{LATT-avg}$). Once such a global activity weight vector is obtained ($\boldsymbol{\beta}_{LATT-max}$ or $\boldsymbol{\beta}_{LATT-avg}$), the activities are ranked by their global weight, they are clustered into five groups {0: unaware, 1: aware, . . . , 4: intent} using the Jenks Natural Breaks algorithm [14] (higher weighted activities are in stages closer to conversion).

4.3 Global attention mechanism (GATT)

The LATT attention mechanism in Section 4.2 introduces a context vector $\mathbf{u}_s$ to measure the local contribution of each input activity towards the combined sequence representation. This neglects the overall (global) importance of each activity in the purchase funnel. Therefore, we introduce the idea of a *global* attention mechanism to capture the global importance of each activity (similar to its position in the purchase funnel), and hence refine conversion prediction; this also enables the computation of funnel tags without the need for weight aggregations across user trails (as done for $\boldsymbol{\beta}_{LATT-max}$ and $\boldsymbol{\beta}_{LATT-avg}$). Figure 4 shows the graphical representation of the proposed global attention mechanism (GATT), and we describe the details below.

Similar to the activity embedding layer, the global weights layer (as shown in Figure 4) looks up the global weight $\beta_{a_t} \in \mathbb{R}$ for an activity a_t in the input user trail. The global weights (vector denoted by $\boldsymbol{\beta}_{GATT}$) are randomly initialized, and are learnt as part of the RNN training process. For computing global attention associated with each activity in the input trail, we further normalize each β_t, and compute the latent trail representation as shown below as shown below:

$$\alpha_t = \frac{\exp\left(\beta_t\right)}{\sum_{t=1}^{l} \exp\left(\beta_t\right)}, \quad \mathbf{v} = \sum_{t=1}^{l} \alpha_t \mathbf{h}_t. \tag{6}$$

where $\alpha_t \in \mathbb{R}$ is the normalized attention weight for activity a_t, and $\mathbf{v}$ is the latent representation of the user trail. The predicted conversion probability $\hat{p}$, and the loss function used for training can now be specified as:

$$\hat{p} = \sigma\left(\mathbf{w}_v^T \mathbf{v} + b_v\right), \tag{7}$$

$$loss_{GATT} = \sum_{i=1}^{|trails_{ADV}|} -y_i \log \hat{p}_i - (1 - y_i) \log(1 - \hat{p}_i), \tag{8}$$

where $\sigma(\cdot)$ is the sigmoid function, $loss_{GATT}$ is the loss function (aggregated over all user trails with seed list activities for advertiser ADV), and y_i is the true binary label for the trail i (conversion or no conversion). From the trained model, the activity-funnel tags are obtained as follows: the activities are ranked by their global weights (in $\boldsymbol{\beta}_{GATT}$), and then clustered into five groups (using Jenks Natural Breaks) as done in the case of LATT.

4.4 LR attention mechanism (LRATT)

In the GATT model, due to the dependence on $\mathbf{v}$, $\boldsymbol{\beta}_{GATT}$ has a weaker influence on the final conversion prediction. Inspired by logistic regression (LR) which assigns a unique weight to each activity, and has the ability to capture positive and negative influence of an activity on conversion, we propose the LR attention model (LRATT) as shown in Figure 5. The LRATT model can also be in-

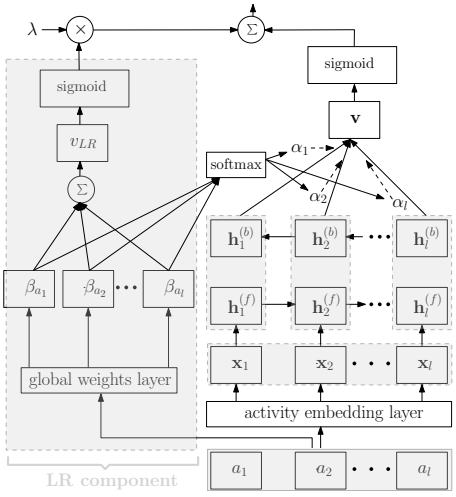

Figure 5: The LR attention-based RNN model.

terpreted as a wide and deep approach (as in [4] which employs a wide LR part, and a deep multilayer perceptron part for non-sequential prediction); LRATTT specifically draws motivation from the purchase funnel to refine sequential conversion prediction and activity-funnel tagging. The main difference with the GATT model, is the update for weights β_t which is influenced by both $\mathbf{v}$ from the RNN part, as well as an additional LR loss as described below:

$$\hat{p}_{LR} = \sigma\left(v_{LR} + b_{LR}\right), \tag{9}$$

$$loss_{LR} = \sum_{i=1}^{|trails_{ADV}|} -y_i \log\left(\hat{p}_{LR,i}\right) - (1-y_i)\log\left(1-\hat{p}_{LR,i}\right), \tag{10}$$

$$loss_{\text{LRATT}} = loss_{GATT} + \lambda loss_{LR}, \tag{11}$$

where $v_{LR} = \sum_{a_t \in \mathbf{a}} \beta_{a_t}$ (*i.e.*, the sum of global weights for activities in the sequence $\mathbf{a}$, and $loss_{GATT}$ is the original GATT loss function (8). The factor λ balances the influences from both the loss functions. The predicted conversion probability is still obtained as in GATT (*i.e.*, $\hat{p}$ in (8)), and the LR part serves as a regularizer for learning the global weights. In addition, while training, we initialize the global weights $\boldsymbol{\beta}_{LRATT}$ with weights from a pre-trained LR model for conversion prediction (over the same dataset). The activity-funnel tags from LRATT are obtained in the same manner as done for GATT (activities are ranked and clustered into 5 groups).

5 USER SCORING

After obtaining activity-funnel tags from a conversion model (as described in Section 4), we use such tags for assigning funnel stages to activity trails via an interpretable and scalable scoring logic (described below in Section 5.1). Following our scoring logic, we explain the metrics for scoring evaluation in Section 5.2.

5.1 MAX-FUNNEL scoring logic

We first setup additional notation for our scoring algorithm (described below). For the task of funnel stage assignment, the online activity trail of user i till (current) time t^* is denoted by:

$$Trail_i(t^*) = [(act_{i1},\ \Delta t_{i1}),\ (act_{i2},\ \Delta t_{i2}), \ldots,\ (act_{ij},\ \Delta t_{ij}) \ldots],$$

where Δt_{ij} is the time difference between $t*$ and the occurrence of act_{ij} in the trail. Also, if activity $act_{ij} \in \mathcal{A}_{ADV}$ (*i.e.*, the seed list for *ADV*), then there is a corresponding funnel tag $funnel(act_{ij}) \in \{0, 1, 2, 3, 4, 5\}$. The scoring logic is as described in Algorithm 1. In simple words, the scoring logic looks at a user's activity trail, finds the activity with the maximum funnel tag (= user's funnel stage); hence the name MAX-FUNNEL scoring. There is an additional modification for recency, where the funnel tag of activities (for stages > 1) beyond an expiry limit ($\Delta_{recency}$ set to 4 weeks) from current time is downgraded to 1 (*i.e.*, aware).

Algorithm 1: MAX-FUNNEL scoring

current time :: t^*
trail of user i :: $Trail_i(t^*)$
initialize tagged trail ::$Trail_{i,tagged} = [\ ]$
for $(act_{ij},\ \Delta t_{ij}) \in Trail_i$ **do**
  if $act_{ij} \in \mathcal{A}_{ADV}$ **then**
    $fun_{ij} = funnel(act_{ij})$
  if $\Delta t_{ij} > \Delta_{recency}$ and $fun_{ij} > 1$ **then**
    $fun_{ij} = 1$
  append (fun_{ij}) to $Trail_{i,tagged}$
$fun_i^* = \max_j(fun_{ij} \in Trail_{i,tagged})$

As shown above, the stage of user i after scoring is fun_i^*, which is basically the maximum funnel tag in the user's trail.

Remark 1. *The motivation for using a simple scoring scheme as described above comes from scalability (in terms of number of users, and trail length) as well as need for intuitive interpretations. The above scoring scheme can implemented efficiently in a large scale online setting since for each user we need to only store: (i) the current stage, and (ii) seed list activities within a recency window for enabling future funnel stage updates. The proposed scheme is much simpler compared to scoring an HMM or deep neural network for serving ads to online users under strict latency constraints (typically of the order of few milliseconds).*

5.2 Scoring evaluation metrics

In this section, we describe metrics for checking the quality of the seed list $\mathcal{A}_{ADV}$, and funnel tagging in the context of user scoring.

5.2.1 Coverage. To measure coverage, we score users based on their activity trails till current time t^*. Let the set of users assigned a funnel tag be $\mathbf{S}_{t^*}$; this is the set of users with at least one seed list activity in their trail. Let set of converters in the time window $(t^*, t^* + \Delta t)$ be $\mathbf{C}_{\Delta t}$, where Δt is set to 4 weeks. The coverage % is defined as $\frac{|\mathbf{C}_{\Delta t} \cap \mathbf{S}_{t^*}|}{|\mathbf{C}_{\Delta t}|} \times 100$. For example, say 8 million users were assigned funnel stages for an advertiser on January 1, 2019, and from January 1 to January 28, 2019, 50×10^3 users converted. If within the 8 million users, 25×10^3 converted between January 1 and January 1 2019, the coverage % is 50%. Higher coverage % indicates a better seed list which is able to capture users likely to convert in the next 4 weeks.

Conversion rates and validity: Similar to the notation introduced for defining coverage % above, let $S_{t^*,k}$ be the scored users in funnel stage $k \in \{0, 1, \ldots 5\}$ at time t^*. The conversion rate for funnel stage k is defined as $conv_k = \frac{|C_{\Delta t} \cap S_{t^*,k}|}{|S_{t^*,k}|}$. For example, if 1000 users were in funnel stage 3 at t^* and 100 out of those converted in the next 4 weeks, then 0.1 is the conversion rate for stage 3. Validity is computed from the funnel stage wise conversion rates in the following manner: $validity = \frac{\sum_{i=0}^{3}\sum_{j\geq i}^{4} \mathbb{1}_{conv_j > conv_i}}{10}$. In simpler words, it penalizes 1 point (out of $\binom{5}{2} = 10$ points) for every mismatched pair of stage wise conversion rates; a mismatch occurs when a higher numbered funnel stage has conversion rate less than or equal to the conversion rate of a lower numbered stage. A perfect score (= 1) for validity means that the conversion rates are in monotonically increasing order (from unaware to intent). This is related to the constraint in HMM formulations (for purchase funnel) which forces the conversion rates from hidden states to monotonically increase with closeness to the purchase stage [1].

6 CREATIVE DESIGN

Designing creatives with the right message for users in a funnel stage is an important part of our approach. To guide creative design, we summarize the seed list of activities (with funnel tags) in the following manner. First, we identify major themes for each funnel stage from the seed list. The themes are derived using the frequency and conversion rate of activities in that stage. Next, we use such themes as hints to guide creative strategists. For example, in the case of theme parks, intent stage ads can highlight deals while aware stage ads can have text like "ideas for holidays" with an image of the park landing users on introductory blogs. We share some of the insightful choices we made for running real campaigns in the supplemental section on reproducibility (Section 9.1).

7 RESULTS

In this section, we go over experimental results associated with conversion prediction, activity-funnel tags, and online ad campaigns with funnel specific ad targeting. The remainder of this section is organized as follows. In Section 7.1, we describe our datasets, followed by results on conversion prediction in Section 7.2. Finally, in Section 7.3 we go over offline and online (*i.e.*, from online ad campaigns) metrics associated with funnel specific ad targeting.

7.1 Evaluation datasets

RecSys 2015 challenge. We conducted conversion prediction experiments on publicly available dataset obtained from RecSys Challenge in 2015. This dataset contains a collection of sequences of click events with respective timesteps from Yoochoose website. Some of the click sessions ended with a purchase event (if so, label was set as positive, otherwise negative); we describe additional details in the supplemental reproducibility section (Section 9.2.1).

User activity trails from Yahoo! (Verizon Media). We also conducted experiments using user activity trails data from Verizon Media. This includes activities done in chronological order by a user; such activities are derived from heterogeneous sources, *e.g.*, Yahoo Search, Yahoo Gemini and viewing content on other publishers associated with Yahoo. The representation of an activity comprises of an activity ID, time stamp, the type (*e.g.*, content view), and a raw description of the activity (*e.g.*, the query for search activities). We obtain data from 3 advertisers (*e.g.*, ad clicks, conversions, and site visits). For describing results, we have anonymized the advertisers (as ADV_i for $i \in \{1, 2, 3\}$); ADV_1 is a mobile phone service provider, ADV_2 is an Internet and cable service provider, while ADV_3 is an employer for ride sharing drivers. In total, the data spanned about 2 billion unique activities, and over 40 million users.

Editorially prepared activity-funnel datasets. For each ADV_i ($i \in \{1, 2, 3\}$), we obtained an editorially prepared dataset of activity-funnel tags. Domain experts manually tagged the seed list for the 3 advertisers with funnel tags; such tags were only used for evaluation and were not a part of the proposed methods. The total number of activities for ADV_1, ADV_2, and ADV_3 was 5.8k, 4.3k, and 12.6k. These datasets were only used to evaluate the global attention weights vis-a-vis human annotated funnel stages (details in Section 9.1).

7.2 Conversion Prediction

7.2.1 Baseline models for conversion prediction.

- **Logistic regression (LR)** consuming the user activity trail as a one-hot encoded feature vector [3].
- **RNN** in its vanilla version with GRU units.
- **RNN + local attention mechanism (LATT)** as described in Section 4.2; also used in [17] for conversion prediction.
- **Multi-head self-attention (MH)** where the RNN layer is replaced with a self-attention mechanism [19].

7.2.2 Conversion model configuration and training. We describe the details in Section 9.3 (reproducibility supplement).

In the remainder of this section, we refer to our proposed models as GATT-{v1, v2} and LRATT-{v1, v2}. GATT-v1 denotes the global attention-based RNN model (as in Section 4.3), and GATT-v2 adds local attention to GATT-v1 (*i.e.*, we concatenate outputs from both the global and local attention layer before employing a sigmoid function for classification). LRATT-v1 is as described in Section 4.4, and LRATT-v2 is the combination of LATT and LRATT-v1, in the same spirit as done for GATT-v2.

Table 1: Test AUC lifts (%) for conversion prediction for ADV_1, ADV_2, and ADV_3, and absolute test AUC for RecSys2015.

model	ADV_1 lift	ADV_2 lift	ADV_3 lift	**RecSys2015** AUC
LR	-	-	-	0.669
MH	17.41%	5.42%	9.82%	0.719
RNN	19.18%	5.55%	11.20%	0.715
LATT	18.78%	5.55%	11.48%	0.730
GATT-v1	20.82%	7.48%	14.80%	**0.739**
GATT-v2	**20.95%**	7.48%	14.11%	**0.739**
LRATT-v1	20.68%	**7.74%**	14.66%	0.737
LRATT-v2	20.82%	7.61%	**15.08%**	0.736

7.2.3 Conversion prediction AUC results. The (offline) results for conversion prediction on all datasets are shown in Table 1. The proposed attention mechanisms dramatically improve the conversion

prediction (test) AUC by 2.17%, 2.19%, and 3.60% on the 3 advertiser datasets (compared to LATT). On the RecSys2015 dataset, our global attention-based models achieved an AUC of 0.739, which is a 7.0% lift above LR and 0.9% lift above LATT. The results show that identifying the global importance of activities improves conversion prediction, whereas considering only local contextual connections may lead to a partial understanding of the user's propensity to convert. We also notice that for different datasets, the variants of our global attention model show slightly different performance. There is no clear improvement when we concatenate local attention representations to the global ones. We conjecture that global attention for activities in some datasets, such as ADV_2, is more important, than the local attention contributions. In addition, the LR and MH models seem to be trading-off performance with efficiency.

Table 2: RMSE and NDCG for activity-funnel tagging.

model	ADV_1 RMSE	ADV_2 RMSE	ADV_3 RMSE	ADV_1 NCDG	ADV_2 NDCG	ADV_3 NDCG
LR	6.00	2.93	5.23	0.93	**0.93**	0.83
LATT-max	4.68	4.14	2.38	0.93	0.91	0.85
LATT-avg	5.11	5.55	2.25	0.92	0.91	0.84
GATT-v1	2.71	3.52	2.24	**0.94**	0.92	0.86
GATT-v2	2.60	2.98	**2.09**	**0.94**	0.92	**0.87**
LRATT-v1	**2.01**	**2.40**	3.18	**0.94**	**0.93**	0.85
LRATT-v2	2.14	2.45	3.17	**0.94**	**0.93**	0.85

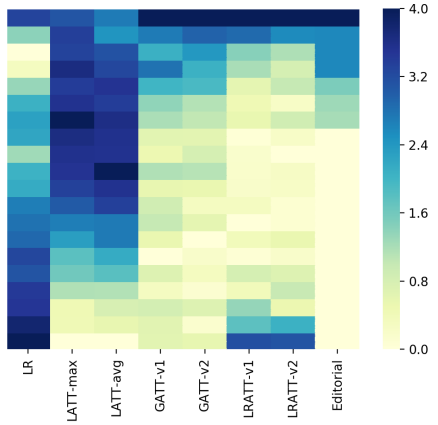

Figure 6: Heatmap showing the alignment of activity-funnel tags (activities sorted by increasing global weight for each model) with human annotated funnel tags for ADV_3 seed list; activities with higher editorial label are darker.

7.2.4 Activity-funnel tagging (intrinsic evaluation). We obtain the (global) attention weight for all seed list activities from a conversion model, and assign funnel tags $\{0, 1, \ldots, 4\}$ as described in Section 4, *i.e.*, we rank activities by weight, followed by clustering (using Jenks Natural Breaks algorithm) them into five groups such that activities with higher weights are assigned funnel stages closer to conversion. We use the editorially labeled tags (for seed list activities) to evaluate how the proposed (heuristic) activity-funnel tagging aligns with human judgement. We use two metrics for such intrinsic evaluation of funnel tagging: (i) Root Mean Square Error (RMSE), and (ii) Normalized Discounted Cumulative Gain (NDCG). For each choice of conversion model (or attention mechanism), the RMSE is calculated between the assigned funnel tags and the human annotated tags for all seed list activities. The smaller the RMSE is, the better the attention/global weights match human judgement. As shown in Table 2, GATT-v2 shows the best performance for ADV_3, while LRATT-v1 outperforms other models for ADV_1 and ADV_2. This shows that the proposed global attention mechanisms achieve superior funnel tagging performance, and LRATT works best in majority of the cases. We also use NDCG to evaluate the quality of activity-funnel tagging. Discounted Cumulative Gain (DCG) and NDCG at rank n can be calculated as follows:

$$\text{DCG}_n = \sum_{i=1}^{n} \frac{2^{rel_i}}{\log_2(1+i)}, \quad \text{NDCG}_n = \frac{\text{DCG}_n}{\text{IDCG}_n}, \tag{12}$$

where rel_i is the relevance value of ith activity (*i.e.*, its editorially annotated funnel stage), IDCG_n denotes the ideal DCG at rank n (when all activities are ranked according to the editorial funnel stages in a decreasing manner), and n is the size of the seed list, *i.e.*, $|\mathcal{A}_{ADV}|$. We present the NDCG results calculated on all sorted attention weights in Table 2. The larger the NDCG values are, the better the attention scores match human judgement. Similar to the RMSE evaluation, GATT and LRATT outperform the baselines. In addition to RMSE and NCDG, we plot a heatmap (Figure 6) to visualize the alignment between ADV_3's attention weights and human annotated funnel stages. For each model, we sort all activities in an increasing order of their global weight (bottom to top in each column in Figure 6), and divide each column into blocks of 20 activities. We then plot the heatmap of the average (editorial) funnel tag in each block. It shows that the GATT columns are closest (visually) to the editorial one, backing up the results in Table 2.

7.3 Funnel specific targeting and creatives

We first go over (offline) scoring results in Section 7.3.1, and results from online ad campaigns in Section 7.3.2.

7.3.1 Scoring coverage and validity. For each of the three advertisers, all users in our dataset from Verizon Media were scored (using the MAX-FUNNEL logic in Section 5.1) with activity trails spanning one year, and the conversion window was a 4 week period. The seed list coverage (as defined in Section 5.2.1, independent of the model used for funnel tagging) was: 81.9% for ADV_1, 76.4% for $ADV2$, and 50.72% for ADV_3. The model validity results (as defined in Section 5.2.1) across different choices of global weights (from conversion models) for activity-funnel tagging are shown in Table 3. Firstly, it is remarkable to see that conversion rates line up in mostly increasing order after our intuitive scoring method with the editorial (human) funnel tags. Secondly, the LRATT model achieves the same performance as the editorial version for ADV_1, and is the closest to human versions for ADV_2 and ADV_3 (0.8 vs. 0.9). It is promising to see the LRATT model match human performance.

7.3.2 Online performance of funnel specific creatives. Our online ad campaigns for ADV_1, ADV_2, and ADV_3 using funnel specific targeting were setup as independent campaigns in the Yahoo Gemini

Table 3: Validity metrics.

model	ADV_1	ADV_2	ADV_3
human	**1**	**0.9**	**0.9**
LR	0.8	0.5	0.5
LATT-avg	*0.9*	*0.7*	0.6
LATT-max	0.8	*0.7*	0.2
GATT-v1	0.8	*0.7*	0.5
GATT-v2	0.7	*0.7*	0.3
LRATT-v1	**1**	*0.7*	**0.8**
LRATT-v2	*0.9*	**0.8**	*0.7*

platform for a period of two months; each funnel stage was a separate ad campaign with its own creative. For reproducibility, we have covered details about the campaign setup and test-control split in Section 9.4). As a result of our campaigns, the (incremental) conversion rate lift in test versus control bucket (not exposed to our ad campaigns) was: 3%, 4% and 6% for ADV_1, ADV_2, and ADV_3 respectively; where the conversion rate in each bucket is the ratio of converters from the bucket to the total number of users in the bucket. Clearly, there is a significant increase in conversions when users are exposed to our campaigns. To study how funnel specific targeting performs vis-a-vis conventional campaigns (using CTR and CVR prediction models [3] with regular creatives), we compared the funnel stage-wise CTR and CPA performance for ADV_2 with a regular campaign for ADV_2 (details in Section 9.4). The results are shown in Table 4, and clearly the LRATT model has the highest number of funnel stages with better CTR and lower CPA. This is in line with LRATT doing well for ADV_2 in terms of RMSE, NDCG and validity. Most funnel stages see a significant CTR improvement compared to the regular campaign, validating custom creatives. We sees lower CPAs for the *unaware* funnel stage; we hypothesize that this might stem from lower competition for targeting such users (whom our custom ads enticed towards conversion).

Table 4: CTR and CPA lifts for funnel specific targeting versus a regular campaign for ADV_2.

		funnel stages					# stages with lift
	model	0	1	2	3	4	
CTR lift in %	human	13	9	-14	17	0	3
	LATT-avg	-44	6	-19	29	19	3
	LATT-max	-44	6	-19	28	19	3
	GATT-v1	43	6	-14	12	22	**4**
	GATT-v2	64	6	-19	28	22	**4**
	LRATT-v1	21	7	-14	27	7	**4**
	LRATT-v2	20	7	-12	15	8	**4**
CPA lift in %	human	-20	-33	-20	-44	-61	**5**
	LATT-avg	-	-31	-23	-50	-34	4
	LATT-max	-	-31	-24	-50	-32	4
	GATT-v1	-	-31	-20	-51	-55	4
	GATT-v2	-88	-30	-24	20	-54	4
	LRATT-v1	-64	-31	-22	-1	-53	**5**
	LRATT-v2	-63	-31	-21	-6	-53	**5**

8 CONCLUSION

The proposed global attention mechanisms not only outperform conversion prediction baselines in terms of AUC, but also produce activity attention weights which are closely aligned with human annotated funnel tags enabling automatic activity-funnel tagging for an arbitrary advertiser. In terms of funnel specific creatives, our approach is a stepping stone towards identifying online users in interpretable funnel stages, targeting them with custom ads, leading to higher CTR and valuable activity insights for an advertiser.

REFERENCES

[1] V. Abhishek, P. Fader, and K. Hosanagar. Media exposure through the funnel: A model of multi-stage attribution. *SSRN*, 2012.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[3] N. Bhamidipati, R. Kant, S. Mishra, and M. Zhu. A large scale prediction engine for app install clicks and conversions. In *CIKM 2017*.

[4] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, et al. Wide & deep learning for recommender systems. In *Workshop on Deep Learning for Recommender Systems*. ACM, 2016.

[5] J. Cheng, L. Dong, and M. Lapata. Long short-term memory-networks for machine reading. *arXiv preprint arXiv:1601.06733*, 2016.

[6] K. Cho and V. Merrienboer. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[7] E. Court, S. Mulder, and O. Vetvik. The Consumer Decision Journey. *McKinsey Quarterly*, 2009.

[8] Y. Cui, R. Tobossi, and O. Vigouroux. Modelling customer online behaviours with neural networks: applications to conversion prediction and advertising retargeting. *arXiv preprint arXiv:1804.07669*, 2018.

[9] A. Ghose, P. Singh, and V. Todri. Got annoyed? examining the advertising effectiveness and annoyance dynamics. *ICIS*, 2018.

[10] D. Gligorijevic, J. Gligorijevic, A. Raghuveer, M. Grbovic, and Z. Obradovic. Modeling mobile user actions for purchase recommendation using deep memory networks. SIGIR '18.

[11] J. Gligorijevic, D. Gligorijevic, I. Stojkovic, X. Bai, A. Goyal, and Z. Obradovic. Deeply supervised model for click-through rate prediction in sponsored search. *Data Mining and Knowledge Discovery*, 2019.

[12] K. Greff and Srivastava. Lstm: A search space odyssey. *IEEE transactions on neural networks and learning systems*, 28(10):2222–2232, 2017.

[13] B. J. Jansen and S. Schuster. Bidding on the buying funnel for sponsored search and keyword advertising. *Journal of Electronic Commerce Research*, 2011.

[14] G. F. Jenks. The data model concept in statistical mapping. *International yearbook of cartography*, 7:186–190, 1967.

[15] M.-T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[16] H. B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, J. Grady, L. Nie, T. Phillips, E. Davydov, D. Golovin, S. Chikkerur, D. Liu, M. Wattenberg, A. M. Hrafnkelsson, T. Boulos, and J. Kubica. Ad click prediction: a view from the trenches. KDD 2013.

[17] K. Ren, Y. Fang, W. Zhang, S. Liu, J. Li, Y. Zhang, Y. Yu, and J. Wang. Learning multi-touch conversion attribution with dual-attention mechanisms for online advertising. In *CIKM*, pages 1433–1442. ACM, 2018.

[18] Y. Shan, T. R. Hoens, J. Jiao, H. Wang, D. Yu, and J. C. Mao. Deep crossing: Web-scale modeling without manually crafted combinatorial features. KDD 2016.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[20] C. Yang, X. Shi, J. Luo, and J. Han. I know you?ll be back: Interpretable new user clustering and churn prediction on a mobile social app. *KDD*, 2018.

[21] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. Hierarchical attention networks for document classification. In *HLT*, 2016.

[22] S. Zhai, K.-h. Chang, R. Zhang, and Z. M. Zhang. Deepintent: Learning attentions for online advertising with recurrent neural networks. KDD '16.

[23] Y. Zhang, H. Dai, C. Xu, J. Feng, T. Wang, J. Bian, B. Wang, and T.-Y. Liu. Sequential click prediction for sponsored search with recurrent neural networks. In *AAAI*, 2014.

[24] G. Zhou, X. Zhu, C. Song, Y. Fan, H. Zhu, X. Ma, Y. Yan, J. Jin, H. Li, and K. Gai. Deep interest network for click-through rate prediction. In *KDD*. ACM, 2018.

9 APPENDIX (REPRODUCIBILITY NOTES)

In this Appendix (reproducibility supplement), we go over additional details required for reproducing the results presented in this paper.

9.1 Seed list generation, editorial tagging and creatives

In this section, we describe some insights, and the choices involved in the seed list generation and editorial funnel tagging of seed list activities (by domain experts) which might be helpful for reproducing the results presented in our paper.

To come up with a seed list, we select top k activities by conversion rate for a given advertiser. The conversion rate for an activity a_i is defined as the ratio of count of users who did a_i, and converted within a time window (*e.g.*, 3 months) over the count of users who did a_i. The choice of k can be such that it is possible to do editorial curation within an ad platform's time constraints, *e.g.*, a few hundred activities can be reviewed by an editor in a matter of hours. The editor can also add a few *obvious* activities like visiting websites or ad clicks of the advertiser to the initial seed list. The conversion rate based method suffers from noise stemming from sparse conversions. The editorial curation is done to remove such noise from the initial seed list. This list is further passed on to a word2vec based expansion method to expand it with contextually similar activity items.

As mentioned in Section 7.1, we consider three real advertisers/brands (anonymized as a mobile phone service provider, an Internet service and cable service provider, and an employer for ride sharing drivers). Similar to the example given for a theme park visit (funnel stages) in Section 1, the domain experts reviewed the seed list for each advertiser, and categorized the activities into different themes. For example, for ADV_3 (driver employer), the observed activity themes (seen in search queries, content views and site visits) included: (i) queries for part time jobs, (ii) reading about life of ADV_3 employees, (iii) clicking on ads from ADV_3 and visiting ADV_3 website, (iv) checking employer requirements, and (v) queries for sign-up page. The domain experts labeled themes (and associated activities) in: (v) as lower funnel (closest to conversion), (ii)-(iv) as mid-funnel, and (i) as upper funnel stages (farthest from conversion). Such an editorial labeling led to a scoring validity of 0.9 (as reported in Table 3 for ADV_3). In this context, it is remarkable to see that an automatic activity-funnel tagging method based on attention weights (LRATT-v1) comes very close (validity=0.8 for ADV_3 as in Table 3) to human labeling performance without employing domain experts and explicit theme identification.

In terms of creative guidance, the funnel tagging process (both editorial and automatic based on attention weights) led to highly interpretable themes in each stage. For example, some search queries in ADV_3 seed list which were automatically identified in the mid-funnel stages (interest and consideration) were about users searching for car leasing options to support their employment as a ride sharing driver. Discovering such interesting activities led to the development of creatives (and landing pages after ad clicks) which could guide interested users towards such leasing option offered by ADV_3. This is one among the numerous insights which were discovered in the funnel tagging process, and which greatly influenced the creatives used in our online experiments.

9.2 Additional details for evaluation datasets

9.2.1 Recsys 2015 dataset. Items that appeared less than 5 times were discarded, and the sessions with length equal to 2 or less were filtered out. These preprocessing steps resulted in 4,428,037 sessions, out of which 377,255 (8.5%) were labeled as positive. We further split sessions into 90% for training and 10% for testing.

9.3 Conversion model configuration and training

In our experiments, we randomly initialized the activity embedding in 128 dimensions. We set both local and global attention size as 50. The GRU hidden unit dimension was also set as 50; so in the case of bidirectional GRU, 100 dimensions for each latent activity representation were generated. For training, we used a mini-batch size of 64. We set the maximum sequence length to be 50, and we padded those sequences that are shorter than 50 with 0's. We used stochastic gradient descent to train all the models with Adam optimizer (in TensorFlow). The best choice of λ was 0.1, and the best learning rate was 0.001 (learnt from grid search). We employed dropout mechanism to regularize the training process with a drop rate of 0.5.

In addition, while tuning the models for conversion prediction AUC, we allowed for an overfit up to 2%, *i.e.*, the maximum allowable difference between test-AUC and train-AUC was 2% of the train-AUC.

9.4 Online campaign setup

Our online ad campaigns for ADV_1, ADV_2, and ADV_3 using funnel specific targeting were setup as independent campaigns within the Yahoo Gemini platform. The creatives for each funnel stage were derived from the editorially tagged seed list for the advertiser (this had to go through a rigorous approval process, limiting the freedom in deciding creatives). After the seed lists were scored, and a funnel stage assigned to each user (with respect to an advertiser), the whole population of scored users was randomly split into test and control buckets (of roughly equal size close to 10 million users). The test population was subjected to funnel specific ads, while no such ads were shown to the control population. Users from both test and control could have been exposed to ads for the concerned advertiser from other advertising platforms outside Yahoo Gemini (such exposure is beyond our control). During our online experiments, both test and control users for ADV_2 were also exposed to a regular campaign for ADV_2 which enabled some comparisons which we have reported in this paper.