Multi-View Multi-Task Campaign Embedding for Cold-Start Conversion Rate Forecasting

Zijun Yao, Deguang Kong, Miao Lu, Xiao Bai, Jian Yang, and Hui Xiong, *Fellow, IEEE*

Abstract—In online advertising, it is critical for advertisers to forecast conversion rate (CVR) of campaigns. Previous work on campaign forecasting concentrates on the time-series analysis which depend on the availability of a length of history. However, these approaches become inadequate for cold-start campaigns which lack for the observation of past. In this work, we attempt to mitigate this challenge by learning an unsupervised and composite campaign embedding to capture multi-view semantic relationships on campaign information, and consequently forecasting the cold-start campaigns using the nearest neighbor campaigns. Specifically, we propose a novel embedding framework which simultaneously extracts and fuses heterogeneous knowledge from multiple views of campaign data in a multi-task learning fashion, to learn the semantic relationship of ad message, conversion rule, and audience targeting. We develop a hierarchical attention mechanism to refine the embedding model at two levels - an intra-view attention to improve context aggregation, and an inter-task attention to balance task importance. Finally, we adopt the k-NN regression model to predict the CVR based on the neighboring campaigns in the embedding space which encodes the multi-view campaign proximity. We conduct extensive experiments on a real-world advertising campaign dataset. The results demonstrate the effectiveness of the proposed embedding method for CVR forecasting in cold-start scenarios.

Index Terms—Online Advertising; Cold-Start Forecasting; Campaign Embedding; Multi-View Learning

1 Introduction

Online advertisement, aiming to increase user engagement by displaying promotional messages to web visitors, has become one of the major businesses in advertising industry. For an advertiser who is preparing to launch a new campaign, it is crucial to make sure this campaign to be invested has high potential of triggering users to react, called conversion. Therefore, the capacity of forecasting conversion rate (CVR) (i.e., the ratio of conversions to clicks) of a new campaign is considered as a prerequisite for making informed spending decisions.

Ad conversion forecasting has attracted heavy interests in literature [1], [2], [3], [4]. Usually, these studies utilize the availability of history like CVR, clicks, or browsing records for future prediction. However, when we are facing a different scenario - cold-start forecasting for new campaigns, the task becomes challenging as the lack of past makes the history-dependent methods inadequate. Therefore, an alternative way to forecast is to refer to the existing campaign - the idea of collaborative filtering, such as using k-nearest neighborhood (k-NN) to search similar but observed campaigns to guide the new campaign forecasting [5]. To achieve this solution, a key task is to learn a comprehensive and descriptive representation, which can encode the heterogeneous information from different views of a campaign, in order to reflect the similarities that inspire the correlation of conversion performance between campaigns.

Campaigns consist of heterogeneous information presented in multiple views of data. As illustrated in Figure 1, a typical campaign can be characterized from three following views of data. The first example view can be ad message - the sentences showing the contents of campaign such as credit card application bonus or auto insurance discount. The second view is from the massive conversion rules which track various conversion types. For example, some rules may track "browsing a page" conversions while some others would track "submitting an order" conversions. These rules define massive and distinct user actions, and these actions can result very different conversion rates. In many cases, advertisers may place multiple conversion rules to form a funnel for tracking a conversion path from acquaintance to decision, such as "browsing product" $\rightarrow$ "add to cart" $\rightarrow$ "submit order". To learn the contents of conversion rules, we can utilize the sequential structure of funnels, like sentences, to learn rules through their contexts. The last view is the targeting of campaigns. Many online ad platforms can help advertisers to display the ads to a specific group of audience in terms of interests, locations, devices, etc. The logic in considering targeting for CVR forecasting is that different population react differently to the same campaign. Therefore, representing the group of targeting is the third important factor to realize campaign similarities. To represent all the three views of information in a single campaign embedding, we need to learn each view separately due to the information heterogeneity; meanwhile, we want to extract all the views simultaneously that all the knowledge can be fused properly to generate the composite campaign embeddings.

- *Z. Yao is with The University of Kansas. E-mail: zyao@ku.edu*
- *D. Kong is with Google. E-mail: doogkong@gmail.com*
- *M. Lu is with Snap Inc. E-mail: ml4ey@virginia.edu*
- *X. Bai, and J. Yang are with Yahoo Research. E-mail: {xbai,jianyang}@yahooinc.com*
- *H. Xiong is with The Hong Kong University of Science and Technology (Guangzhou). E-mail: xionghui@ust.hk*

To represent these multi-view campaign data, a traditional way is to "hand-crafte" descriptive features such as image colorfulness, word number of ads [6], and numbers of positive users on landing pages [4]. Building these features usually require knowledge-intensive feature engineering, and in many cases, they are not comprehensive enough to capture overall campaign characteristics. Another way is to use one-hot encoding to represent all the raw information of campaigns such as the presence of words, rules, and targetings. However, this will result high dimensional representations in order to accommodate data views with massive information like all the possible vocabulary, rules, and geographic information. More importantly, this representation reflecting the distinction in raw data can hardly express the underlying semantic relationship in the information, therefore they are not capable as well of showing campaign similarities to achieve CVR collaborative filtering.

In recent years, embedding methods which learn the distributed representation of words through the structure of sentences has made great progress in semantic learning, such as word2vec [7] and glove [8]. Without the prior knowledge of the target words, the semantic similarity can be encoded in their representation by analyzing the frequency distribution of their contexts (words appear around it in sentences). Comparing to the disadvantage of above-mentioned representations, this method can be a promising choice to extract and fuse the multi-view information of ads campaign based on three main reasons: (1) campaign information appear in sequential structures, such as the ad messages consisting of words, and conversion funnels consisting of rules. With these "sentence"-like data, we are able to learn the semantics of information pieces like ad words or rules; (2) in addition to learning single information pieces, the whole campaign can be learned as a composite of information like "documents" [9], which serves our goal genuinely to summarize campaigns based on an individual view of data; (3) the way of bag-of-words learning in predicting words through a document provides an extension of fusing multi-view knowledge by making a campaign simultaneously predict heterogeneous information adopted from all the views. In this way, we are expecting to learn a composite space of campaigns, where the campaign proximity can be shown with a unified embedding, and the complex multi-view characteristics are encoded comprehensively.

In this work, we develop a **M**ulti-**V**iew Multi-**T**ask Embedding with Hierarchical **A**ttention (**MVTA**) framework to generate composite campaign representations from heterogeneous ads data. Specifically, by treating the semantic learning within each view as an individual task, we learn the campaign embedding by (1) extracting heterogeneous knowledge from each view separately, and (2) learning all the views at the same time, in order to achieve the optimal fusion of heterogeneous semantics into a composite vector. To refine the embedding process, we apply a hierarchical attention mechanism where an intra-view attention helps to better aggregate the contexts within views, and an inter-task attention aims to balance the importance across tasks in the overall optimization objective. Finally, we validate the quality of campaign embeddings in the scenario of cold-start campaign forecasting using a real-world advertisement campaign dataset. In summary, we highlight the technical

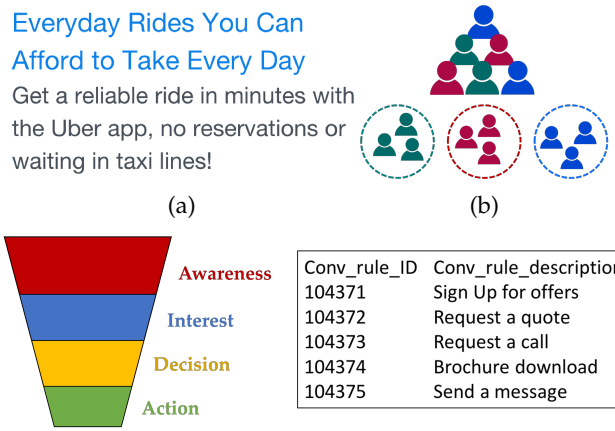


Fig. 1: Illustration of multiple views of campaign information: (a) ad message; (b) audience targeting; (c) conversion funnel and conversion rule.

contributions of this paper as follows:

- We identify a different but important research problem of forecasting cold-start CVR for new campaigns. Under the circumstance where the past of campaigns is unknown, we propose an collaborative filtering way of utilizing similar but existing campaigns based on a sophisticated embedding space.
- We develop a multi-view multi-task embedding framework with a hierarchical attention mechanism to learn a composite campaign representation which is capable of capturing semantic relationship in multiple views of campaign data including ad message, conversion rule and audience targeting.
- Experimental results based on the real-world data demonstrate that the proposed embeddings manage to characterize the multi-view semantic relationship between campaigns and show the effectiveness in adopting k-NN approach to solve the problem of cold-start campaign forecasting.

2 Problem Statement

Definition 1. ***Conversion*** *occurs when an ad click leads to an user engagement (i.e. a valuable action like "place an order"). A campaign* ***conversion rate*** *evaluates the overall conversion-to-click ratio (between 0 and 1) from all the audience in a certain period of time (e.g., a day).*

Definition 2. ***Ad message*** *is the sentences displayed by advertiser to inform the audience of promoted products, services or concepts. It can be viewed as a sequence of words and each unique word is associated with an ID.*

Definition 3. ***Conversion rule*** *tracks the conversion about a specific action (e.g., "input the payment" or "view the white page").* ***Funnel*** *is a sequence of rules for tracking a path of conversions, which usually shows user engagements from shallow to deep levels. Since the rules are designed for tracking fine-grained actions, there are massive number of rules and each one is associated with a unique IDs.*

TABLE 1: Symbols and notations.

Symbol	Definition
w_t	t-th word from the ad message
r_z	z-th rule from the conversion funnel
a_m	m-th audience targeting of the campaign
k	A campaign instance
v	Input vector (embedding), e.g., v_k, v_w
v'	Output vector (parameters), e.g., ${v'}_w$
g, b	Intra-view attention parameters
h	Inter-task attention parameters
δ	Window size of words (hyperparam)
γ	Aggregation factor of rules (hyperparam)

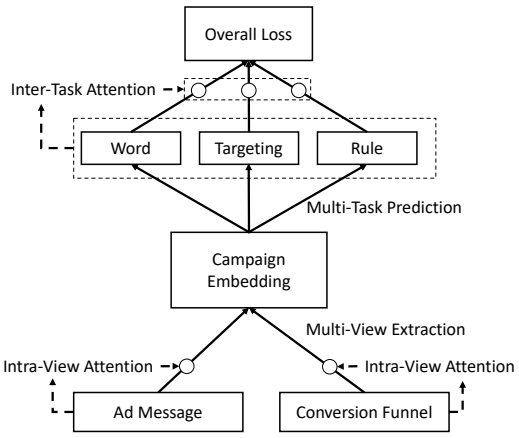


Fig. 2: Overview of campaign embedding framework.

Definition 4. ***Audience targeting*** *is the selection of campaign receivers by interests, device platforms, and so on. Advertisers decide a set of audience targetings based on campaign strategy to exclusively display their ads. In our work, each targeting is a audience group defined by a combination of a browsing interest and a device type, and each unique audience targeting is associated with an ID.*

Definition 5. *A* ***campaign*** *is particular promotion for which advertisers need to prepare the ad messages, the conversion funnel, and the audience targetings before the launch. In this work, a campaign consists of a set of* ***campaign instances*** *where each instance independently indicates the campaign under a unique combination of an ad message, a conversion rule, and an audience targeting. Thus, for each unique campaign instance, there exists a corresponding CVR.*

Suppose we have a set of campaigns categorized into two groups: existing campaigns and cold-start campaigns. Accordingly, for existing campaigns we have N_1 campaign instances with observed CVR denoted by $\{(x_i, y_i)\}_{i=1}^{N_1}$ and for cold-start campaigns we have $N - N_1$ campaign instances $\{x_j\}_{j=N_1+1}^{N}$. For existing campaign instances we have x_i denoting all the campaign information by three views: ad message, conversion funnel, and audience targeting; and the observed CVR label denoted by y_i. For cold-start campaign instances we only have the campaign information x_j but the CVR label is unknown.

Looking into each campaign instance k, the ad message is a sequence of words $\{w_1, w_2, ..., w_T\}$; the conversion rule r is a rule from the funnel $\{r_1, r_2, ..., r_Z\}$ of the campaign; the audience targeting is a particular user group from the targeting set $\{a_1, a_2, ..., a_M\}$ of the campaign. Each unique campaign instance k, word w, rule r and targeting a is associated with a unique ID.

In the embedding phase, our goal is to use the multi-view campaign information x of all the instances k to learn the composite campaign instance embedding $v_k \in \mathbb{R}^D$. In the following prediction phase, we use the embedding and CVR label of existing campaign instances to train a k-nearest neighbors (k-NN) regressor and predict the unknown CVR of cold-start instances using their embeddings. Major notations can be found in Table 1.

3 METHODOLOGY

In this section, we describe the model of **M**ulti-**V**iew Multi-**T**ask Embedding with Hierarchical **A**ttention (**MVTA**) for the goal of learning composite embedding to capture multi-view campaign relationship. The modeling motivation and the overall model diagram are presented first. Then we introduce each module of the model in details.

3.1 Model Overview

Given the multi-view information - ad message, conversion rule and audience targeting, we proposed to learn the three views of data in a **separate-and-fuse** fashion.

Separate: Each view of the campaign data has the heterogeneity thus express a different type of semantics. Therefore, in the information extraction we need to treat the views separately. We illustrate the semantics of each view as follows:

- From the ad message we extract promoted contents. If two campaigns are having the related promoting themes, they are more likely to have similar CVR by being interesting to similar users, for example, "insurance" and "car sale" campaigns are both interesting to auto consumers.
- From the conversion rules we extract user reactions. If two rules track actions in very different aspects of user engagement, they are likely to have different CVR, such as "page view" vs. "order submit".

Fuse: While the heterogeneous views are being extracted, we need to fuse all the views to generate the composite campaign embedding. To achieve optimal information fusion, it is important to integrate the fuse stage with the extract stage as a whole optimization process, because (1) the campaign embedding can be as genuine as the embedding within each view, and (2) can be sophisticated enough to balance each view based on the importance of the information which is being fused.

Model Overview: Figure 2 shows the overview of proposed method. Generally, we develop a multi-task framework by making campaign embeddings (1) separately extract information within multi-views, and (2) simultaneously fuse the target words, rules and targetings included in the campaign.

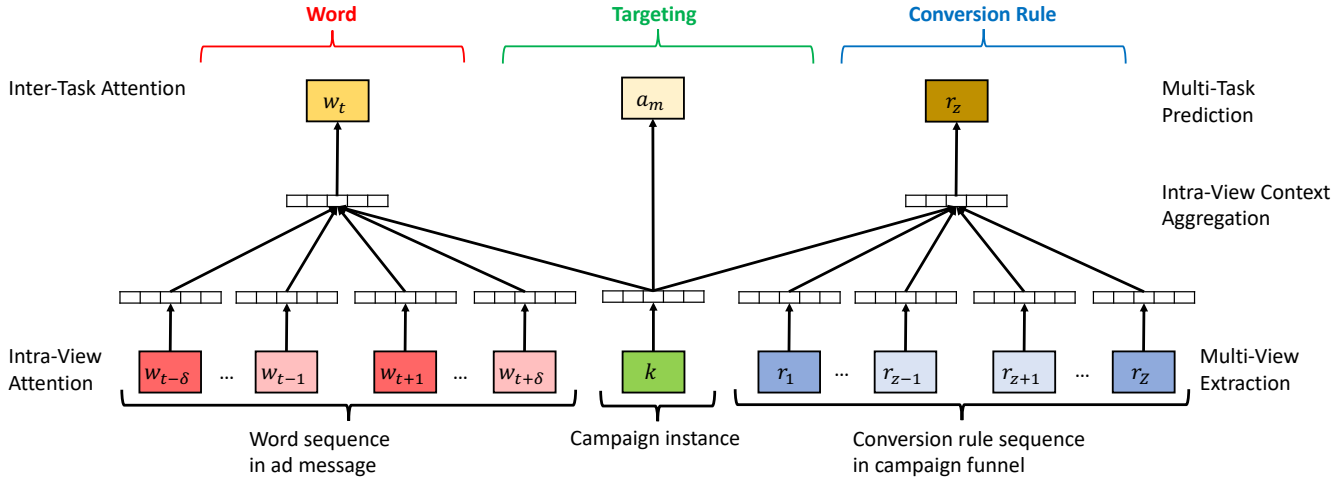


Fig. 3: The model of multi-view multi-task embedding with hierarchical attentions. Three view of information is learned including ad word, targeting and conversion rule. Views of ad message and campaign funnel include context aggregation with the intra-view attentions. All the three views are predicted simultaneously using an inter-task attention to balance task importance. Different categories of color denote different attention functions and the color darkness shows the examples of importance for information on which the attention functions weight.

In detail, the lower level of Figure 2 illustrates the extraction of two sequences of information - ad message and conversion funnel. Within each view, we add an intra-view attention function to better aggregate contexts by distinguishing important ones. In the upper level of Figure 2, we construct a multi-task prediction layer where the campaign embedding is used for fuse its target word, rule and targeting at the same time. Since we need to aggregate the losses of all tasks as an overall objective, we develop an inter-task attention function to better balance the importance of views to achieve descriptive embeddings where more related information gain more attentions in optimization.

3.2 Model Architecture

Figure 3 shows the detail architecture of the proposed model. We depict the model by three parts - each part will cover the knowledge extraction in a single view.

As illustrated in the part of ad message of Figure 3, each word w_t is mapped to a input vector v_{w_t} and an output vector ${v'}_{w_t}$, and each campaign instance k is mapped to a vector v_k. To predict the word w_t that appears in the messages, we form the input using (1) the surrounding words $w_{t-\delta}, ..., w_{t+\delta}$ defined by a context windows size δ and (2) the affiliated campaign instance k. Formally, the objective is to maximize the log likelihood over all words in the campaign instance as

$$\mathcal{L}_{word} = \sum_{w_t \in k} \log \Pr(w_t | w_{t-\delta} : w_{t+\delta}, k). \tag{1}$$

The probability $\Pr(w_t | w_{t-\delta} : w_{t+\delta}, k)$ is defined using a softmax function as in the Continuous Bag of Words (CBOW) model:

$$\Pr(w_t | w_{t-\delta} : w_{t+\delta}, k) = \frac{\exp({v'}_{w_t}^{\top} v_{w_I})}{\sum_{j=1}^{|W|} \exp({v'}_{w_j}^{\top} v_{w_I})}, \tag{2}$$

where W is the entire vocabulary, ${v'}_w$ is output vector of target words, and v_{w_I} is the aggregated input vector of context words and the campaign instance. Here we will continue to the introduction of the next view, and leave the aggregation of v_{w_I} with intra-attention to be specified in the following section 3.3.

As shown in the part of conversion funnel in Figure 3, we aim to learn the semantic of conversion rules by considering the rule co-occurrences in campaign funnels. The goad is to maximize of the log likelihood as

$$\mathcal{L}_{rule} = \log \Pr(r_z | r_1 : r_Z, k), \tag{3}$$

where r_z is the particular conversion rule of campaign instance k. $r_1 : r_Z$ denote all the rules assigned in the campaign except the one of current instance. Since the length of campaign rules are usually short, to make more samples of rule co-occurrence available in the study, we simply consider all the rest of rules in the funnel as contexts instead of applying a context window. The conditional probability $\Pr(r_z | r_1 : r_Z, k)$ is also modeled using a CBOW style softmax function:

$$\Pr(r_z | r_1 : r_Z, k) = \frac{\exp({v'}_{r_z}^{\top} v_{r_I})}{\sum_{j=1}^{|R|} \exp({v'}_{r_j}^{\top} v_{r_I})}, \tag{4}$$

where R is all the unique conversion rules, ${v'}_r$ is output vector of target rules, and v_{r_I} is input vector combined of the context rules and the campaign instance.

The last view is audience targeting as shown in Figure 3. Each campaign has a targeting set for multiple audience populations in terms of interests and device platforms. Unlike the sequences in ad message or funnel where semantic-revealing contexts exist in neighborhoods, the set of audience targeting is more dependent on particular advertising strategy (e.g., budget) rather than targeting semantic dependency. Due to this reason, we did not form the context of targeting following CBOW architecture, instead, we adopt Skip-Gram (SG) structure for learning this view by predicting the current targeting using the campaign instance solely

to form the input vector. The objective in audience targeting learning is formulated by

$$\mathcal{L}_{targeting} = \log \Pr(a_m|k), \tag{5}$$

where a_m is the audience targeting of campaign instance k. The conditional probability $\Pr(a_m|k)$ is computed with a softmax function:

$$\Pr(a_m|k) = \frac{\exp({v'_{a_m}}^\top v_k)}{\sum_{j=1}^{|A|} \exp({v'_{a_j}}^\top v_k)}, \tag{6}$$

where A is all the unique audience targetings, ${v'}_{a_m}$ is the output vector of the current targeting, v_k is the input vector of campaign instance.

3.3 Context Aggregation with Intra-View Attention

For learning semantics of ad message or conversion funnel, the input vector v_{w_I} or v_{r_I} consists of two part: contexts and campaign instance. The contexts in each view need to be aggregated to form an intra-view context vector. To aggregate context words or rules as the input vector, all the context embeddings are averaged with a uniform weight traditionally. However, the influence of different words or rules are not always equally informative in real life. Therefore, we develop an intra-view attention to weight (sum up to 1) each context word or rule to obtain a better aggregation. In addition to weighting contexts, the attention function needs to also discriminate the weighting mechanism on different campaign categories. For example, words like "credit" is informative in financial service or real estate campaigns, but may be not in entertainment campaigns. Therefore, we further make the intra-view attention to have campaign category awareness.

We aggregate context words or rules as non-uniform weighted sum by intra-view attention

$$\begin{aligned} v_{\hat{w_I}} &= \sum_{t-\delta \le j \le t+\delta, j \ne t} \alpha_{w_j,c_k} v_{w_j} \\ v_{\hat{r_I}} &= \sum_{1 \le j \le Z, j \ne z} \alpha_{r_j,c_k} v_{r_j}, \end{aligned} \tag{7}$$

where α_{w_j,c_k} and α_{r_j,c_k} denote the attention weights given to context word w or rule r in the campaign categories c. The attention weights are calculated by the parameters associated with each unique word or rule:

$$\begin{aligned} \alpha_{w_j,c_k} &= \frac{\exp(g_{w_j,c_k} + b_{c_k})}{\sum_{t-\delta \le l \le t+\delta, l \ne t} \exp(g_{w_l,c_k} + b_{c_k})} \\ \alpha_{r_j,c_k} &= \frac{\exp(g_{r_j,c_k} + b_{c_k})}{\sum_{1 \le l \le Z, l \ne z} \exp(g_{r_l,c_k} + b_{c_k})}, \end{aligned} \tag{8}$$

where $g_{w \in W, c \in C} \in \mathbb{R}^{|W| \times |C|}$ and $b_{c \in C} \in \mathbb{R}^{|C|}$ are the parameters for intra-view attention of ad message; $g_{r \in R, c \in C} \in \mathbb{R}^{|R| \times |C|}$ and $b_{c \in C} \in \mathbb{R}^{|C|}$ are the attention parameters for conversion funnel. For simplifying the presentation, here we reuse the notation g and b for intra-view attention parameters of word and rule, which are actually independent from each other.

Once we obtain the aggregation of context words $v_{\hat{w_I}}$ and rules $v_{\hat{r_I}}$, we need to further combine them with campaign instance embeddings to form the final input vectors v_{w_I} and v_{r_I}. [9] suggests that concatenation generally provide good input which also confirms the finding in our study. Therefore, for input vector of ad message v_{w_I}, we concatenate the campaign instance vector and the context words vector as

$$v_{w_I} = concatenate(v_k, v_{\hat{w_I}}). \tag{9}$$

For conversion funnels, some campaigns may only watch one conversion action (e.g., the final purchase). In these cases, unlike ad messages which always include more than one word, conversion funnel can consist of only one rule. Consequently, concatenation can not always be applicable since the rule context can be missing. Therefore, we use an weighted average way to combine the embedding of rule contexts $v_{\hat{r_I}}$ and the embedding of campaign instance v_k. We employ a predefined weight $0 \le \gamma \le 1$ (e.g., $\gamma = 0.5$) to preserve the minimum impact of campaign instance against rule contexts:

$$v_{r_I} = \begin{cases} v_k & \text{if } Z_k = 1 \\ \gamma v_k + (1-\gamma) v_{\hat{r_I}} & \text{otherwise,} \end{cases} \tag{10}$$

where Z_k denotes the number of conversion rule in k's campaign. If only one rule is deployed in current campaign funnel which is the target rule itself, the input will consist of only the campaign instance v_k. Otherwise, we use the weighted average for combination.

3.4 Joint Predictions with Inter-Task Attention

As shown in the output layer of Figure 3, we simultaneously predicting words, conversion rules and audience targetings in a multi-task fashion for fusing all the knowledge in the three views into a single campaign instance embedding. In this joint learning structure, the outputs of three tasks need to be combined as an overall loss for optimization. In real life, the importance of each task is different depends on the informativeness of the predicting word, rule or targeting to be output. For example, when the model is predicting a key word, a high-profit conversion rule, or a major audience group, optimization should put more attention to the loss of this predicting task comparing to others. Based on this motivation, we develop an inter-task attention across the output layer to weight the loss of each view.

In objective function, for each time of outputting a combination of a word w_t, a rule r_z and a targeting a_m, we sum up the predicting loss from each individual task $\mathcal{L}_{word}(w_t)$, $\mathcal{L}_{rule}(r_z)$, $\mathcal{L}_{targeting}(a_m)$ with non-uniform weights:

$$\begin{aligned} \mathcal{L}_{overall} = \sum_{k \in K} \sum_{w_t \in x_k} \big(& \beta_{w_t} \mathcal{L}_{word}(w_t) + \beta_{r_z} \mathcal{L}_{rule}(r_z) \\ & + \beta_{a_m} \mathcal{L}_{targeting}(a_m) \big), \end{aligned} \tag{11}$$

where β_{w_t}, β_{r_z}, β_{a_m} denote the weights to coordinate the output loss of predicting particular word, rule and targeting. It is worth noting that the weights of view here are dynamically changing in each time of output depending on what the combination of three views is. For example, two $\beta(w_t)$ of a same predicting word can be different in two times of output based on different rule or targeting accompanying.

The inter-task attention weights β is parameterized by

$$\beta_{s_i} = \frac{\exp(h_{s_i})}{\sum_{s_i \in \{w,r,a\}_i} \exp(h_{s_i})}, \tag{12}$$

Algorithm 1 The algorithm of MVTA model.

Input: A set of campaign instances k consisting of ad message $\{w_1...w_T\}$, conversion rule r_z, targeting a_m.
Output: Embedding of each campaign instance v_k.

Random initialization for embeddings v_k, network parameters v_w, v_r, ${v'}_w$, ${v'}_r$, ${v'}_a$, and attention parameters g, b, h
while not reaching convergence criteria **do**
  for each campaign instance k **do**
    for each word $w_t \in \{w_1...w_T\}$ **do**
      # Intra-view attention
      Prepare input vector v_{w_I} by Eq. 7, 8, 9
      Prepare input vector v_{r_I} by Eq. 7, 8, 10
      # Inter-task attention
      Predict current word w_t by Eq. 1, 2
      Predict current rule r_z by Eq. 3, 4
      Predict current targeting a_m by Eq. 5, 6
      Obtain the overall loss $\mathcal{L}_i$ of current output by weighting prediction losses for the three views by Eq. 11, 12, 13
    end for
  end for
  Obtain the total loss $\mathcal{L}$ by accumulating losses of all outputs $\mathcal{L} = \sum_{i \in I} \mathcal{L}_i$
  Optimize $\mathcal{L}$ using gradient descent and update all the parameters
end while

where $s_i \in \{w, r, a\}_i$ is a symbol to indicate one of the three predicting views in the i-th multi-task output. For example, when s_i represents the output rule r_i in the three-view combination of i-th output, we calculate β_{r_i} using the softmax function over all the predicting objects including the rest two views - w_i and a_i. The inter-task attention parameters are $h_{w \in W} \in \mathbb{R}^{|W|}$, $h_{r \in R} \in \mathbb{R}^{|R|}$, and $h_{a \in A} \in \mathbb{R}^{|A|}$ which determine the influence of each unique word, rule, and targeting when they are being predicted. Here we reuse the notation h for each view for simplifying the presentation.

Once we explain the inter-task attention module, we can write the detail form of the overall loss to optimize the campaign instance embedding:

$$\mathcal{L}_{overall} = \sum_{k \in K} \sum_{w_t \in x_k} \big(\beta_{w_t} \frac{\exp({v'_{w_t}}^\top v_{w_I})}{\sum_{j=1}^{|W|} \exp({v'_{w_j}}^\top v_{w_I})} + \beta_{r_z} \frac{\exp({v'_{r_z}}^\top v_{r_I})}{\sum_{j=1}^{|R|} \exp({v'_{r_j}}^\top v_{r_I})} + \beta_{a_m} \frac{\exp({v'_{a_m}}^\top v_k)}{\sum_{j=1}^{|A|} \exp({v'_{a_j}}^\top v_k)}\big). \tag{13}$$

To optimize Equation 13, we use mini-batch with shuffling to train the model. We adopt sampled softmax loss to increase efficiency by converting the softmax computation from a large class number to a relatively small class number. We update campaign instance embeddings v_k with network parameters v_w, v_r, ${v'}_w$, ${v'}_r$, ${v'}_a$, and attention parameters g, b, h in each iteration. The detail algorithm is summarized in Algorithm 1. Finally, we obtain the trained embeddings of campaign instances v_k when the optimization converges.

TABLE 2: Statistics of campaign instances across campaign categories.

Category	Number	Category	Number
Auto	137	Professional Services	2978
CPG	388	Real Estate	78
Careers & Education	215	Reference	171
Entertainment	8109	Retail	943
Financial Services & Insurance	7145	Technology	996
Health & Wellness	300	Telecom & Web Services	431
Personals & Social Services	443	Travel & Transportation	319

4 Experimental Study

In this section, we empirically evaluate the performance of the proposed **M**ulti-**V**iew Multi-**T**ask Embedding with Hierarchical **A**ttention (**MVTA**) framework on a real-world ads campaign dataset. [1]

4.1 Data Description

We format a campaign dataset on a major online ads platform in the U.S. market covering a full day in summer 2017. We obtain 1562 unique campaigns after removing inactive ones, which finally result in a total of 22661 campaign instances and span 14 campaign categories. Table 2 shows the number of campaign instances in each campaign category. We can see that most campaigns concentrate on "Entertainment", "Financial Services & Insurance" and "Professional Services". Meanwhile, Figure 4 shows the histogram of CVR value across campaign instances. We can see that most campaign instances receive less than 0.1 CVR.

For multi-view campaign information, we obtain 4850 unique words in ad message excluding stop words and rare words, 485 unique conversion rules in conversion funnel, and 182 unique audience targetings covering 3 device types (i.e., desktop, smartphone, tablet) and 84 browsing interests (e.g., mortgage, dating). We randomly sample 10% campaigns to construct the cold-start campaign set. In total we obtain 157 cold-start campaigns including all their 2884 campaign instances with CVR masked. The CVR prediction of these cold-start instances will be completely dependent on the neighboring instances of existing campaign with CVR observed through k-NN regression using the trained embeddings.

4.2 Baseline Approaches

The experimental study compares our proposed embedding (**MVTA**) framework with the following baselines:

• **Multi-Hot Encoding** [10]: Using one-hot with multiple labels to indicate the presence of all the unique words, rules and audience targetings in campaigns, where 1 indicate existence and 0 for otherwise.

1. Codes are available at: https://github.com/zyao237/mvta

TABLE 3: Performance comparisons of CVR forecasting for cold-start campaign instances using k-NN regression.

Methods	MSE			MAE			R^2			EVS		
Neighbors K	10	20	30	10	20	30	10	20	30	10	20	30
Multi-hot	0.0390	0.0382	0.0385	0.1392	0.1370	0.1378	0.2978	0.3120	0.3075	0.2983	0.3134	0.3098
PCA	0.0391	0.0375	0.0374	0.1408	0.1380	0.1383	0.2957	0.3255	0.3269	0.2958	0.3257	0.3272
LDA	0.0439	0.0434	0.0435	0.1483	0.1472	0.1483	0.2102	0.2185	0.2166	0.2173	0.2265	0.2251
AE	0.0529	0.0507	0.0503	0.1671	0.1644	0.1646	0.0475	0.0883	0.0955	0.0565	0.0942	0.0989
Word2vec	0.0390	0.0378	0.0373	0.1403	0.1378	0.1370	0.2989	0.3199	0.3288	0.3004	0.3215	0.3302
D2V-DM	0.0444	0.0447	0.0452	0.1539	0.1568	0.1591	0.2017	0.1952	0.1864	0.2036	0.1958	0.1867
D2V-DBOW	0.0387	0.0369	0.0363	0.1395	0.1367	0.1364	0.3034	0.3361	0.3468	0.3043	0.3368	0.3470
MVTA	**0.0338**	**0.0336**	**0.0342**	**0.1321**	**0.1319**	**0.1339**	**0.3902**	**0.3944**	**0.3843**	**0.3908**	**0.3949**	**0.3844**
Improvement	12.67%	8.91%	5.77%	5.28%	3.52%	1.84%	28.60%	17.34%	10.82%	28.42%	17.27%	10.79%

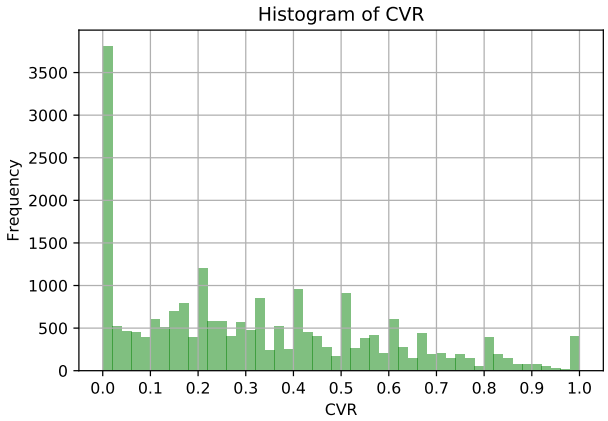


Fig. 4: Histogram of CVR.

• **Principal Component Analysis (PCA)** [11]: A dimension reduction method using orthogonal transformation to convert possibly correlated variables into a set of linearly uncorrelated variables.

• **latent Dirichlet allocation (LDA)** [12]: A generative statistical model that allows data to be explained by latent groups that explain why some parts of the data are similar. We consider each campaign instance as a document which contains the terms from all information views.

• **Autoencoder (AE)** [13]: A type of artificial neural network used to learn efficient data representations in an unsupervised manner. We use this approach to obtain a low-dimensional representation of multi-hot vectors.

• **Word2vec** [7]: A two-layer neural network that is trained to reconstruct linguistic contexts of words. We use CBOW for this method, every campaign instance embedding is the average vector of all contained information.

• **Doc2vec-DM, Doc2vec-DBOW** [9]: Embedding methods of paragraph and document for capturing the overall semantics. We view all the words, the rule and the targeting belonging to an campaign instance as terms of a document. Two implementations are used: the distributed memory (DM) and distributed bag of words models (DBOW).

4.3 Experimental Setting

We use Tensorflow to build the proposed embedding framework. For the hyperparameters, we set window size δ to be 5 for the view of ad message following the general word2vec setting as [7]. Smaller δ results in insufficient training examples, while larger δ is difficult to be accommodated since native ads usually do not have long paragraphs or sentences. For the intra-view learning of conversion funnels, we set aggregation factor $\gamma = 0.5$ to combine the embedding of rule contexts and the embedding of campaign instance to obtain input vector v_{r_I}. The intuition is that as the number of rules increase in the cases of long funnels, the learning of context rules will dominate the learning of campaign instance if using averaging aggregation. To prevent this issue, we use γ to reserve at least half of the input for campaign instance, and let all the context rules to share the rest part. In this way, we give the rule contexts and the campaign instance the same importance within this view. For more details, we use the sampled softmax loss for more efficient computation. The number of negative samples is 200 for words, 100 for rules, and 100 for targetings. The size of mini batches is 30. The optimizor is Adam with learning rate as 0.0001. For all approaches, except multi-hot encoding, we learn the embedding of 50-dimension resulted from a tradeoff between performance and efficiency. For validation, we first learn the embedding for all campaign instances including both cold-start and existing ones. Then we train the CVR regression model on the existing campaign instances with observed CVR. Finally we obtain the predictions of cold-start instances and evaluate them using regression metrics.

4.4 Performance Comparison

We first evaluate the embeddings of campaign instances through our model against the baseline embedding approaches. We use k-NN regression as it is designed to utilize the information relationship: campaign instances with similar semantics are more likely to locate closely. In our method, we attempt to use trained embeddings to find nearest existing campaign instances, and then use their observed CVR to make prediction of cold-start campaign instances. Specifically, k-NN regression identify Top-K existing campaign instances in neighborhood, and use their CVR to make distance weighted interpolation.

We evaluate the error between predicted CVR denoted by $\hat{y}$ and groundtruth CVR denoted by y of cold-start campaign instances as follows:

- **Mean Square Error (MSE)**:

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} (y_i - \hat{y}_i)^2. \quad (14)$$

- **Mean Absolute Error (MAE)** :

$$\text{MAE}(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=1}^{n_{samples}} |y_i - \hat{y}_i|. \quad (15)$$

- **R Squared Score (R^2)** :

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n_{sample}} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n_{sample}} (y_i - \bar{y}_i)^2}. \quad (16)$$

- **Explained Variance Score (EVS)** :

$$\text{EVS}(y, \hat{y}) = 1 - \frac{\text{Var}(y - \hat{y})}{\text{Var}(y)}. \quad (17)$$

Table 3 shows the Mean Square Error (MSE), Mean Absolute Error (MAE), R Squared Score (R^2), and Explained Variance Score (EVS) of cold-start CVR prediction results on all approaches with three different neighbor size K for k-NN regression algorithm: 10, 20, and 30. Generally, we can see that our proposed approach MVTA consistently outperforms baseline methods on all metrics in every K neighbor size settings. We also present the improvement percentage of our method over the second best in the baselines which are mostly D2V-DBOW.

Specifically, first we can see that raw feature approach multi-hot gives a medium-level performance. A possible reason is that although they do not contain semantic relationship as embedding methods do, they do preserve the complete information with the ultra-high dimension. As we discussed, this kind of encoding approach brings vulnerabilities like constraint for new information, computation inefficiency, and overfitting. As a dimension reduced feature of multi-hot, PCA gives average results too. We see that it is similar to multi-hot. A possible reason is that it picks the principle components which select the most informative features. As another common dimension reduction approach, LDA does not provide good results. Since multi-hot features are very sparse, sometimes it is hard for constructing comprehensive latent topic distribution. Then we have the worse approach in our study: autoencoder. A possible reason is that it focuses on encode data for deep neural network representation, which does not provide semantic expressing embeddings for k-NN methods. Another possible reason is also the sparsity problem of multi-hot features, which may make autoencoder network overfitted. Last, we have the three word2vec based methods. For classic word2vec, we can see that it gives nice results because the semantic similarity is captured. Then Doc2vec-DM does not perform well. A possible reason is that the information from all the views are mixed together for extraction, and attention mechanism is not performed in context aggregation. Doc2vec-DBOW perform relatively well, make the second place generally. Across all the K setting, we find that K=20 gives the best overall results for almost all approaches.

TABLE 4: Ablation study on attention mechanism.

Methods	MSE			R^2		
Nbr. K	10	20	30	10	20	30
No attention	0.0367	0.0359	0.0362	0.3404	0.3545	0.3482
Intra-view	0.0355	0.0348	0.0349	0.3609	0.3743	0.3715
Inter-task	0.0357	0.0357	0.0359	0.3572	0.3580	0.3540
MVTA	0.0338	0.0336	0.0342	0.3902	0.3944	0.3843

To summarize, the experimental results demonstrate that (i) the proposed embedding framework is able to capture semantic relationship of campaign instances based on the three campaign information views; (ii) the proposed design which extracts multi-view information separately and fuses knowledge with multi-task objectives is helpful for better mining ads campaign data.

4.5 Ablation Analysis

4.5.1 Validation of Hierarchical Attention

To validate the effectiveness of the hierarchical attention mechanism, we conduct an ablation analysis as shown in Table 4 to demonstrate the importance each attention module contributes to the MVTA model. In the first method "No attention" we utilize the multi-view multi-task architecture but without any attention mechanism. In the second and third methods we simply add the intra-view or the inter-task attention function respectively. In the last row we show the performance of complete MVTA model with both attention functions.

From the results shown in Table 4, we can observe that the model with neither of attention functions (i.e., "No attention") shows the lowest performance, but still it performs better than the baselines. It suggests that when we are facing heterogeneous multi-view data, this "separate-and-fuse" framework is able to effectively extract composite embeddings. Next, we compare the two partial attention models (i.e., "intra-view" and "inter-task") against the full attention model (i.e., "MVTA"). From the perspective of scores, we can see that (1) single intra-view attention has improved the performance by around 3% and 6% for MSE and R^2 while single inter-task attention has less improvement - 2% on MSE and 3% on R^2; (2) by having both attention components working together (MVTA), we can achieve 6% improvement in average of MSE, and 13% improvement in average of R^2. These results suggest a collaboration effect of two attentions which exceeds the sum of their individual performances. From the perspective of utility, we believe that attention mechanism is necessary for the proposed model to work on other potential applications or datasets. By flexibly coordinating the intra-view and the inter-task aggregations, the model can be sophisticated and robust to handle different data situations.

4.5.2 Validation of Multi-View Fusion

To compare the contribution of different views and to validate the effectiveness of view fusion, we conduct an ablation study of using individual views or two views instead of using them all. All the methods use the same MVTA model

TABLE 5: Ablation study on multi-view fusion.

Views	MSE			R^2		
Nbr. K	10	20	30	10	20	30
Ad msg	0.0399	0.0395	0.0401	0.2824	0.2895	0.2786
Conv. rule	0.0398	0.0385	0.0378	0.2829	0.3062	0.3205
Target	0.0460	0.0442	0.0438	0.1717	0.2043	0.2119
Msg+rule	0.0361	0.0352	0.0352	0.3498	0.3661	0.3671
Msg+target	0.0378	0.0370	0.0370	0.3202	0.3335	0.3332
Rule+target	0.0361	0.0358	0.0357	0.3502	0.3560	0.3571
All views	0.0338	0.0336	0.0342	0.3902	0.3944	0.3843

where one or two of the views are removed. As shown in Table 5, we have the first 3 rows showing the performance of using each individual view, and the next 3 rows showing the performance of using combination of any two views. Starting from the single views, we can see that generally individual views have the lowest performances in terms of MSE and R^2. Among them, the view of conversion rules has the highest scores while the view of ad message has a close second place. When we see the combinations of any two views, we can see that first, combination of two views generally provide better performance than single views; and second, the combinations including the view of rules have the highest performance. Based on the best score cross all the neighbor K, the views of msg+rule rank the first place and the views of rule+target rank the second. The last row shows the combination of all views which is the full MVTA model. We can see that by fusing all the views of campaign information, we achieve the best performance. Through this analysis, we first rank the contribution of the views in our study as conversion rules $>$ ad message $>$ targeting. Secondly, we validate that the proposed model is capable of fusing multi-view information and using more views of data is helpful to improve forecasting results.

4.6 Sensitivity on Dimensions

Figure 5 shows the sensitivity of the model performance on embedding size. Three different dimensions are tested - 25, 50, and 100 and two metrics are shown - MSE and R^2. Generally, we see that the performance increases as the embedding size go large (except dim-100 vs. dim-50 at K=10). But the computational cost also goes high when we increase the embedding size. Embedding at 50 dimension shows a relatively good tradeoff between performance and efficiency. For example on K=20 which shows the best overall performance, we can see that the improvement from dim-25 to dim-50 is significant, but the growth from dim-50 to dim-100 slows down notably. In this case, it is not worth enough using expensive double size of embedding to buy this limited improvement.

4.7 Visualization of Campaign Embeddings

Based on trained embeddings of campaign instances, we apply t-SNE [14] to generate 2-D representation of them for visualization. We aim to see if the embeddings of campaign instances are able to capture important character of each

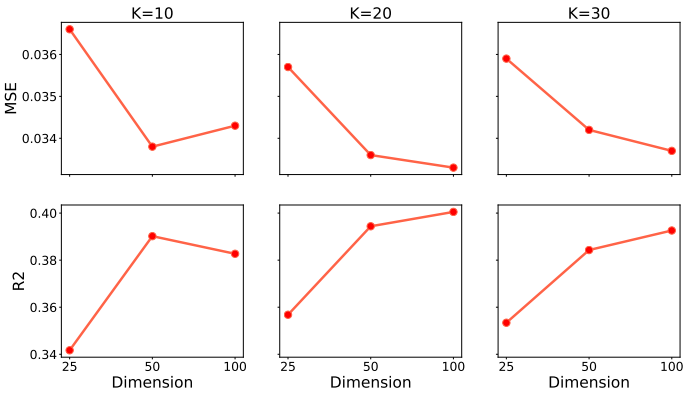


Fig. 5: Performance comparison on embedding sizes.

information view, which can be expressed by proximity in t-SNE space.

Figure 6a shows the campaign instances colored based on the words in ad messages. We aim to see if the instances expressing similar promotions in their messages are close from each other. We use topics to represent the words in contents. As shown in Table 6, we collect the top frequent words for each campaign category, and try to label instances with topics based on the words in their messages. In detail, some words may belong to more than one topics, we count the frequency for each word on each topic in each instance. Then we choose the highest frequent topic as the instance's label. We color the instance with its ad message topic label. We can see that for "Financial Services" and "Entertainment" topics, instances are quite separated from other topics. But for "Professional Service", instances are easier to mix with other topics including "Technology" and "CPG". Generally, we can see that instance with similar ad messages are more likely to locate nearby.

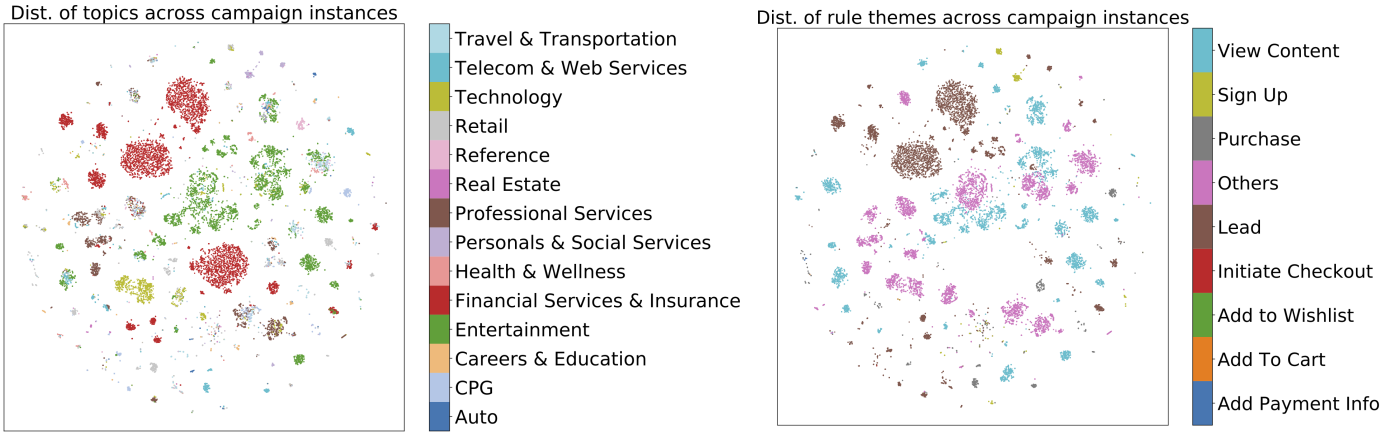
Figure 6b colors the campaign instances based on their conversion rules. Although advertisers customize a large number of conversion rules for tracking various actions, each rule can be categorized by engagement themes generally. In our system, 83% conversion rules have categorical themes, and we visualize those available instances in Figure 6b. We find that the similar rules locate closely in embedding space.

Figure 7a shows the distribution of top-20 frequent audience targetings (ranked from bottom to top labels in color bar). Each targeting is defined by an fine-grained browsing interest and a device type. The most frequent interest targetings are credit cards, content providers and local services. Both of their desktop and smartphone targetings are on the top list. Meanwhile, we can see that even for the same interest, they are not embedded closely by different device types. It suggests that audiences from different device platforms also have different conversion behaviors.

Figure 7b shows the correlation between the proximity on embedding and the difference on CVR value. The CVR of each campaign instance is indicated by color gradient between 0 and 1 where Red means high and blue means low. We can see that low CVR instances and high CVR instances tend to locate separately. This shows that two campaign instances are more likely to receive different CVR if they

TABLE 6: Top frequent words of each campaign category.

Categorical topic	Top words
Auto	deals, browse, today, price, toyota, hyundai, sport, honda, chevrolet, ford
CPG	skin, chuck, fun, wings, love, home, food, wrinkles, treatment, tightening
Careers & Education	free, research, job, post, hiring, earn, learn, data, skills, site, student
Entertainment	famous, insane, celebrities, weight, loss, people, transformations, woman, years, gorgeous
Financial Serv. & Insu.	apply, terms, card, earn, delta, bonus, platinum, offer, credit, miles, hilton, points, rewards
Health & Wellness	surprise, foods, cancer, eat, weight, loss, para, headaches, immunotherapy, headache
Personals & Social Serv.	women, single, meet, free, men, car, pics, view, click, inventory, unsold, online
Professional Services	computer, voice, internet, touchscreen, pride, review, service, free, improvements, touring
Real Estate	find, listings, apartment, view, zip, search, price, homes, foreclosed, sale, site
Reference	search, degree, info, early, childhood, title, loans, requirements, master, education, models
Retail	shop, sale, save, macy, summer, independence, season, shipping, july, mattress, prices
Technology	screen, galaxy, amazing, edge, world, infinity, larger, expansive, display, vr
Telecom & Web Serv.	dental, affordable, teeth, fixing, internet, implant, price, convenient, local, pricing
Travel & Transport.	book, car, rapid, credit, uber, paid, travel, fare, earn, delta, save, points



(a) Ad message

(b) Conversion rule

Fig. 6: T-SNE visualization of campaign instance by the views of ad message and conversion rule.

have small semantic similarity in terms of ad message, conversion rule, and audience targeting.

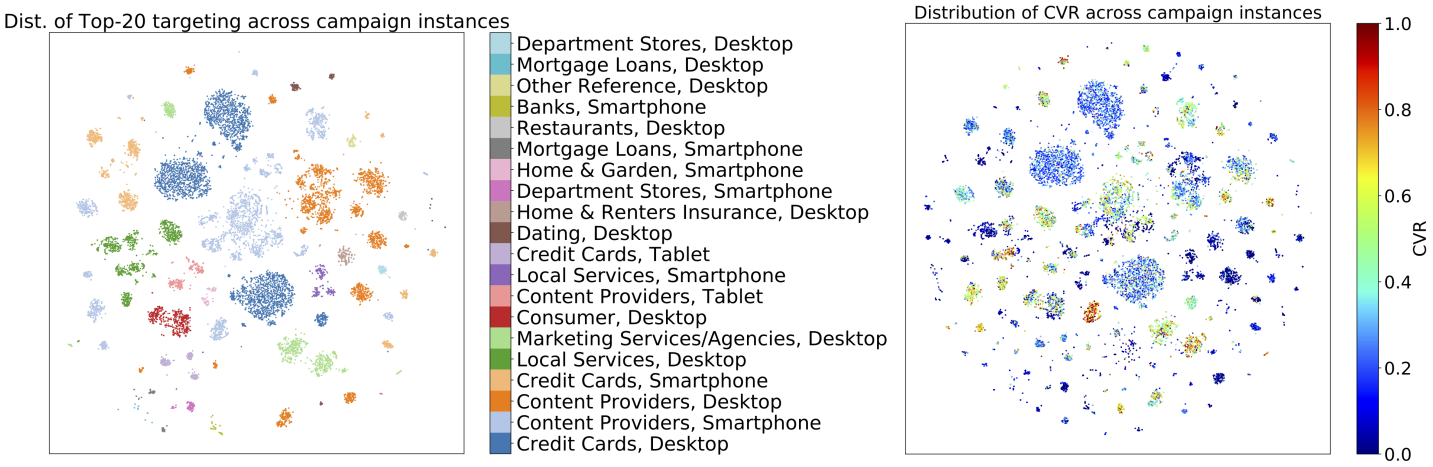
4.8 Campaign Instance Relevance Analysis

In Table 7, we show a case study of k-NN search results in the embedding space learned from the MVTA model. The left panel of Table 7 shows the cold-start campaign instance while the right panel shows the relevant existing campaign instance search by embedding distance. Each ad consists of a title and a description. In the first row, we can see two campaigns for different credit cards: airline card and hotel card. However, these two ads both aim for travel purpose. Therefore, using the hotel card campaign instance we can forecast the cold-start airline card campaign instance. In the second row, existing campaign instance for internet connection service is found to be one of the nearest neighbors for the cold-start campaign instance of TV cable service. The two campaign instances in the third row are both about mortgage loans. The forth row is about financial management. The cold-start campaign instance is related to the medicare expense while the existing campaign case is about financing. In the fifth row, cold-start ad is for car sales while the existing instance is for auto insurance. In the last row, both ads are for IT training. All these searching cases further validate the MVTA embedding in terms of capturing the semantic similarity between campaign instances, and mitigating the challenge on cold-start campaign forecasting.

5 Related Work

5.1 Conversion Modeling

In computational ads, conversion rate models the ratio of number of conversions from all clicks, which is widely used for personalized targeting and cost-per-acquisition (CPA) optimization. For example, CVR for an e-commerce website models the likelihood of this visitor becoming a buyer after the visitor clicks the website, i.e., the probability the visitor may convert such as purchase, order, etc. In particular, accurately predicting the conversion rate



(a) Audience targeting

(b) Conversion rate

Fig. 7: T-SNE visualization of campaign instance by the view of targeting and CVR value.

TABLE 7: The neighboring campaign instances searched for the cold-start campaign instances through embeddings. Left panel: cold-start campaign instances; right panel: the existing campaign instances which are relevant.

Cold-start campaign instances	Relevant campaign instances
Amex Gold Delta SkyMiles Card, "Official Site. Earn More, Go Further with the New 60K Bonus Miles & $50 Statement Credit Offer from Gold Delta Skymiles. Offer Ends 7/5. Terms Apply."	Special Offer Ends 5/31, "Earn 80,000 Hilton Honors Points & Enjoy Travel Rewards! Ends 5/31. Terms Apply."
Premium TV Experience Has People Rushing Home, "See how others are enjoying more control, more storage and more choice, now including over 150,000 Video On Demand titles all with FiOS by Frontier."	Get an Internet adrenaline rush., "Capture the excitement of FiOS with blazing fast upload and download speeds, 2 years of HBO included and over 150,000 Video On Demand titles."
Congress Gives Florida Homeowners Huge Bailout, "If you own a home in Florida and owe less than $625,000 on your home, you better read this now."	Forget Social Security if you Own a FL Home, "If you own a home, you should read this. Thousands of homeowners did this yesterday, and banks are furious! Do this now before it's..."
2017 Lowest Medicare Supplement Rates, "Get a medicare supplement plan that will cover up to 100% of your yearly health expenses."	How Wealthy is Megyn Kelly?, "Some stars are the news; others deliver the news ... and make a lot of money while they're at it. Here's the top 18."
Best Small SUV, "Looking for a Small SUV? Find the Right Information and Save Time"	Boston Grads Disrupted A $19 Billion Industry, "This small team of data scientists has made an algorithm that is turning the giant 19 billion dollar auto insurance industry upside down"
Explore Big Data Careers, "We need more people to create positive change through data. Are you up to the challenge?"	Today's IT Jobs, "There's more than one career path in IT. Check out some common IT jobs."

for user post-click conversions (such as purchasing, filling forms, etc.) is very important for bid optimization. Almost all the existing works focus on user-level CVR prediction by leveraging various ad features to develop sophisticated algorithms, i.e., predicting CVR for a particular user. For example, [15] introduced an enhanced collaborative filtering (CF) based approach for ad conversion event prediction. [1] proposed an approach of conversion rate estimation by leveraging observations from user, publisher and advertiser data hierarchies for ad targeting. [16] estimated the post-click engagement on native ads by predicting the dwell time on ad landing pages. [17] proposed a novel probabilistic generative model by tightly integrating natural language processing and dynamic transfer learning for CVR prediction. [18] proposed an additional multi-touch attribution model by considering the ad exposure time and ad exposure browsing path in order to enhance conversion prediction.

5.2 Ads related Cold-Start Problem

Cold-start concerns the issue that the computation system is unable to perform prediction (e.g., CTR prediction, item recommendation, etc.) due to the lack of sufficient history information, which in fact appears frequently in ad serving platforms or recommender systems. To address these issues, [6] extracted multimedia features such as brightness, colorfulness from display ads to model click prediction. [19] proposed a new CNN type feature learning pipeline to learn distinctive features from images for cold-start CTR prediction in image display ads. [20] addressed the cold-start conversation recommendation problem in the online learning setting by developing a preference elicitation framework to identify the questions that should be asked to the new users. [21] presented a context-aware semi-supervised co-training method using factorization model to tackle the cold-start problem in recommendation systems. [22] designed a framework to predict the pre-click quality of native ads based

on a crowd-sourcing study to identify important ad criteria in native advertisement. [23], [24] developed multi-armed bandit algorithms for cold-start ads recommendation.

5.3 Embedding Learning

Word representation has been largely improved with the rise of neural networks [25], [26]. Recent years, the work in GloVE [8] and word2vec [7], [27] which utilize word dependency in documents to infer word semantics, have greatly increased performances in key NLP tasks, such as document clustering [28], and word similarity [29]. Doc2vec model [9] has shown very promising performance in learning the representations of paragraph and document. Later, [30] introduced attention mechanism to differentiate word contexts in CBOW model. Because of the advantage in semantic learning, word2vec was then extended to broader application. For example, [31] presented another query embedding model to capture similarity between queries for helping advertisers identify relevant queries in sponsored search advertising. [32] learned word embeddings to capture similarity between queries and ads to improve ads display in sponsored search advertising. [33] developed a time-aware attention mechanism for medical concept embedding from time-series health records. [34] learned the embeddings of online listing of Airbnb for large-scale personalized searching.

5.4 Multi-View and Multi-Task Learning

Multi-view representation learning has grown rapidly as information in multiple forms has been pervasively collected of in many applications [35], [36], [37], [38]. Generally, multi-view representation learning can be categorized into two groups: (1) multi-view representation alignment and (2) multi-view representation fusion. Our work falls under the second category which aims to integrate multi-view knowledge into a single composite representation. For example, multi-view CNN [39] was developed for integrating multiple 2D views of objects for 3D object recognition in image processing. RNN encoder-decoder [40] can also be incorporated for connecting multi-modal sequences. Multi-modal deep autoencoder is proposed to learn shared representation for knowledge fusion [41].

More recent efforts in this thread have been analyzing the unique properties of individual views in order to achieve an adaptive integration over multi-view information. [42] proposed a model CPM-Nets to jointly consider the completeness and versatility to conduct flexible and generalizable partial multi-view learning, and the unified representation is able to improve the prediction performance because of its enhanced separability. [43] developed a latent multi-view subspace clustering algorithm which exploits the complementary information from different views to study the underlying clustering structures comprehensively, and eventually achieves a self-expressive multi-view integration. [44] presented a classification model which dynamically assesses the quality of views in terms of uncertainty, therefore a trusted and interpretable multi-view fusion can be assured at the sample level.

Multi-task learning (MTL) especially in deep neural networks [45] has successfully improved performance in numerous applications by sharing representations over multiple related tasks for better generalization. For example, [46] proposed a deep relationship network which use matrix priors to learn the relationship between tasks. Recently, self-paced MTL [47] has attracted increasing interest by automatically assigning the weights to each task given that tasks have different complexity, importance, or uncertainty. [48] proposed to adjust single tasks relative weight in overall loss function by maximizing the Gaussian likelihood with task uncertainty. [49] presented a gradient normalization algorithm which dynamically tunes the gradient magnitudes of each task for balancing the training of multiple tasks automatically.

6 Conclusion

In this paper, we developed a multi-view multi-task embedding framework (MVTA) with hierarchical attentions to learn an unsupervised and composite campaign embedding for solving cold-start forecasting of conversion rates (CVR) by searching for similar existing campaigns. Specifically, we proposed a multi-view multi-task model to learn campaign instance embedding by simultaneously extracting and fusing three views of campaign information including ad message, conversion rule, and audience targeting. We developed a hierarchical attention mechanism to better aggregate information during optimization at two hierarchies - intra-view attention for individual view context aggregation and inter-task attention for across-views output loss aggregation. In experimental study, we used k-nearest neighbor (k-NN) regression method to forecast cold-start campaign instances based on similar existing campaign instances which have observed CVR. The results based on the real-world ads campaign dataset demonstrated the effectiveness of the proposed method with a consistent performance improvement over baselines. We also conducted ablation analysis on attention mechanism, individual views, and sensitivity analysis on embedding size. In addition, we provided qualitative analysis including t-SNE visualization of campaign instance regarding different views, and case studies on search results for relevant ad campaigns.

References

[1] K. Lee, B. Orten, A. Dasdan, and W. Li, "Estimating conversion rate in display advertising from past erformance data," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 768–776.

[2] Y. Zhang, Y. Wei, and J. Ren, "Multi-touch attribution in online advertising with survival theory," in *IEEE International Conference on Data Mining*. IEEE, 2014, pp. 687–696.

[3] X. Shao and L. Li, "Data-driven multi-touch attribution models," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 258–264.

[4] Y. Liu, S. Pandey, D. Agarwal, and V. Josifovski, "Finding the right consumer: optimizing for conversion in display advertising campaigns," in *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012, pp. 473–482.

[5] G. Marvin, "Google rolls out similar audiences for search and shopping," https://searchengineland.com/google-rolls-similar-audiences-search-shopping-274210.

[6] H. Cheng, R. v. Zwol, J. Azimi, E. Manavoglu, R. Zhang, Y. Zhou, and V. Navalpakkam, "Multimedia features for click prediction of new ads in display advertising," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 777–785.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, 2013, pp. 3111–3119.

[8] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[9] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International Conference on Machine Learning*, 2014, pp. 1188–1196.

[10] D. Harris and S. Harris, *Digital design and computer architecture*. Morgan Kaufmann, 2010.

[11] N. Halko, P.-G. Martinsson, and J. A. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM Review*, vol. 53, no. 2, pp. 217–288, 2011.

[12] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[13] M. A. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[14] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[15] M. Aharon, A. Kagian, and O. Somekh, "Adaptive online hyperparameters tuning for ad event-prediction models," in *Proceedings of the 26th International Conference on World Wide Web Companion*, 2017, pp. 672–679.

[16] N. Barbieri, F. Silvestri, and M. Lalmas, "Improving post-click user engagement on native ads via survival analysis," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 761–770.

[17] H. Yang, Q. Lu, A. X. Qiu, and C. Han, "Large scale cvr prediction through dynamic transfer learning of global and local features," in *Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications*, 2016, pp. 103–119.

[18] W. Ji and X. Wang, "Additional multi-touch attribution for online advertising," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 1360–1366.

[19] K. Mo, B. Liu, L. Xiao, Y. Li, and J. Jiang, "Image feature learning for cold start problem in display advertising," in *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[20] K. Christakopoulou, F. Radlinski, and K. Hofmann, "Towards conversational recommender systems," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 815–824.

[21] M. Zhang, J. Tang, X. Zhang, and X. Xue, "Addressing cold start in recommender systems: A semi-supervised co-training algorithm," in *Proceedings of the 37th international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2014, pp. 73–82.

[22] K. Zhou, M. Redi, A. Haines, and M. Lalmas, "Predicting preclick quality for native advertisements," in *Proceedings of the 25th International Conference on World Wide Web*, 2016, pp. 299–310.

[23] S. Li, A. Karatzoglou, and C. Gentile, "Collaborative filtering bandits," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 539–548.

[24] C. Gentile, S. Li, P. Kar, A. Karatzoglou, G. Zappella, and E. Etrue, "On context-dependent clustering of bandits," in *Proceedings of the 34th International Conference on Machine Learning*, 2017, pp. 1253–1262.

[25] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin, "A neural probabilistic language model," *Journal of Machine Learning Research*, vol. 3, no. Feb, pp. 1137–1155, 2003.

[26] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, 2008, pp. 160–167.

[27] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[28] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International conference on machine learning*, 2015, pp. 957–966.

[29] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.

[30] W. Ling, Y. Tsvetkov, S. Amir, R. Fermandez, C. Dyer, A. W. Black, I. Trancoso, and C.-C. Lin, "Not all contexts are created equal: Better word representations with variable attention," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1367–1372.

[31] M. Grbovic, N. Djuric, V. Radosavljevic, F. Silvestri, R. Baeza-Yates, A. Feng, E. Ordentlich, L. Yang, and G. Owens, "Scalable semantic matching of queries to ads in sponsored search advertising," in *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016, pp. 375–384.

[32] S. Lee and Y. Hu, "Joint embedding of query and ad by leveraging implicit feedback," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 482–491.

[33] X. Cai, J. Gao, K. Y. Ngiam, B. C. Ooi, Y. Zhang, and X. Yuan, "Medical concept embedding with time-aware attention," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*, 2018, pp. 3984–3990.

[34] M. Grbovic and H. Cheng, "Real-time personalization using embeddings for search ranking at airbnb," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018, pp. 311–320.

[35] Y. Li, M. Yang, and Z. Zhang, "A survey of multi-view representation learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 10, pp. 1863–1883, 2018.

[36] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv preprint arXiv:1304.5634*, 2013.

[37] W. Liu, D. Tao, J. Cheng, and Y. Tang, "Multiview hessian discriminative sparse coding for image annotation," *Computer Vision and Image Understanding*, vol. 118, pp. 50–60, 2014.

[38] S. Kiani, S. Awan, C. Lan, F. Li, and B. Luo, "Two souls in an adversarial image: Towards universal adversarial example detection using multi-view inconsistency," in *Annual Computer Security Applications Conference*, 2021, pp. 31–44.

[39] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 945–953.

[40] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[41] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 689–696.

[42] C. Zhang, Y. Cui, Z. Han, J. T. Zhou, H. Fu, and Q. Hu, "Deep partial multi-view learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[43] C. Zhang, H. Fu, Q. Hu, X. Cao, Y. Xie, D. Tao, and D. Xu, "Generalized latent multi-view subspace clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 86–99, 2018.

[44] Z. Han, C. Zhang, H. Fu, and J. T. Zhou, "Trusted multi-view classification," in *The 9th International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021.

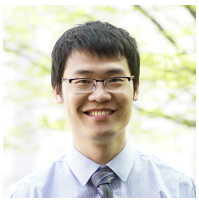
[45] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

[46] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," in *The 2015 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2015, pp. 912–921.

[47] C. Li, J. Yan, F. Wei, W. Dong, Q. Liu, and H. Zha, "Self-paced multi-task learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[48] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.

[49] Z. Chen, V. Badrinarayanan, C. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in *Proceedings of the 35th International Conference on Machine Learning*, 2018, pp. 793–802.



Zijun Yao received his Ph.D. in Information Technology from Rutgers, the State University of New Jersey. He is an Assistant Professor in EECS Department at The University of Kansas. Prior to that, he was a Research Staff Member in Center of Computational Health at IBM TJ Watson Research Center. His general interests lie in data mining with applications in health informatics, mobile intelligence, and natural language processing. He has published in peer-reviewed journals and conferences such as IEEE TKDE, ACM TKDD, KDD, IJCAI, WSDM, ICDM, SDM, and ECML. He served as journal reviewer and conference program committee for TKDE, TKDD, KAIS, TMIS, KDD, AAAI and IJCAI.

Deguang Kong has been leading many ad quality projects (e.g., response prediction, bidding and forecasting, marketplace optimization, etc) and machine learning projects (e.g., sparse learning, model compression, dialogues and natural language understanding) at Google, Yahoo Research, etc. He published over 40 papers at major venues in NIPS, ICML, WWW, AAAI, KDD, WSDM, etc, and served as a PC for these top-tier conferences as well.

Miao Lu is a Machine Learning Engineer at Snap Inc. He has strong interdisciplinary background in statistics, machine learning and data mining, with wide applications in biomedical science and internet technology. He published peer-reviewed papers in journals/conferences like WWW, NIPS TSW, SDM MLREC, PLOS NTDs, EbioMedicine, Biomarker Research, mBio, Respiratory Medicine, etc, with global impact, particularly on child health and online advertisement.

Xiao Bai received her PhD degree in computer science from INRIA, Rennes, France. She is a Principle Research Scientist with Yahoo Research, Sunnyvale, USA. Prior to that, she was a Research Scientist with Yahoo Labs, Barcelona, Spain. Her research areas include information retrieval, online advertising, and distributed data management. Her contributions to various domains of research have been presented in top venues where she regularly serves as PC member, such as SIGIR, CIKM, WSDM, and WWW.

Jian Yang is a Senior Director at Yahoo Research. He obtained Ph.D. in Electrical and Computer Engineering from University of California, Davis. His research interests include optimization, forecasting, and machine learning, with applications in online advertising, pricing and revenue management, and supply chain management. He has published in various venues of computer science, engineering, and business, such as IEEE Transactions, Operations Research, Journal of Machine Learning Research, Journal of Marketing, TheWebConf, WSDM, and NeurIPS.

Hui Xiong (F20) is currently a Chair Professor at The Hong Kong University of Science and Technology (Guangzhou). Dr. Xiongs research interests include data mining, mobile computing, and their applications in business. Dr. Xiong received his PhD in Computer Science from University of Minnesota, USA. He has served regularly on the organization and program committees of numerous conferences, including as a Program Co-Chair of the Industrial and Government Track for the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), a Program Co-Chair for the IEEE 2013 International Conference on Data Mining (ICDM), a General Co-Chair for the 2015 IEEE International Conference on Data Mining (ICDM), and a Program Co-Chair of the Research Track for the 2018 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. He received the 2021 AAAI Best Paper Award and the 2011 IEEE ICDM Best Research Paper award. For his outstanding contributions to data mining and mobile computing, he was elected an AAAS Fellow and an IEEE Fellow in 2020.