arXiv:1905.07289v1 [cs.CL] 17 May 2019

Conversion Prediction Using Multi-task Conditional Attention Networks to Support the Creation of Effective Ad Creatives

Shunsuke Kitada
Hosei University
Tokyo, Japan
shunsuke.kitada.8y@stu.hosei.ac.jp

Hitoshi Iyatomi
Hosei University
Tokyo, Japan
iyatomi@hosei.ac.jp

Yoshifumi Seki
Gunosy Inc
Tokyo, Japan
yoshifumi.seki@gunosy.com

ABSTRACT

Accurately predicting conversions in advertisements is generally a challenging task, because such conversions do not occur frequently. In this paper, we propose a new framework to support creating high-performing ad creatives, including the accurate prediction of ad creative text conversions before delivering to the consumer. The proposed framework includes three key ideas: multi-task learning, conditional attention, and attention highlighting. Multi-task learning is an idea for improving the prediction accuracy of conversion, which predicts clicks and conversions simultaneously, to solve the difficulty of data imbalance. Furthermore, conditional attention focuses attention of each ad creative with the consideration of its genre and target gender, thus improving conversion prediction accuracy. Attention highlighting visualizes important words and/or phrases based on conditional attention. We evaluated the proposed framework with actual delivery history data (14,000 creatives displayed more than a certain number of times from Gunosy Inc.), and confirmed that these ideas improve the prediction performance of conversions, and visualize noteworthy words according to the creatives' attributes.

CCS CONCEPTS

• **Information systems** → **Online advertising**; • **Computing methodologies** → **Multi-task learning**; *Neural networks.*

KEYWORDS

Online Advertising, Supporting Ad Creative Creation, Recurrent Neural Network, Multi-task Learning, Attention Mechanism

ACM Reference Format:
Shunsuke Kitada, Hitoshi Iyatomi, and Yoshifumi Seki. 2019. Conversion Prediction Using Multi-task Conditional Attention Networks to Support the Creation of Effective Ad Creatives. 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, 9 pages. https://doi.org/10.1145/3292500.3330789

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '19, August 4–8, 2019, Anchorage, AK, USA
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6201-6/19/08…$15.00
https://doi.org/10.1145/3292500.3330789

Figure 1: An example of an ad creative in digital advertising: an ad creative is constructed with an image and two short texts. These short texts are called the title and the description.

1 INTRODUCTION

In display advertisements, ad creatives, such as images and texts, play an important role in delivering product information to customers efficiently [6]. Figure 1 shows an example of an ad creative which is constructed by two short texts and an image. The performance of these advertisements is generally defined by *the revenue of conversions* per *the cost of the advertisement*. Conversions are user actions, such as the purchase of an item or the download of an application, and they represent a known metric that advertisers try to maximize through their ad creatives. The costs of advertisements are generally calculated by the cost per click (CPC), where an advertiser pays for the number of times their advertisement has been clicked. Therefore, the high performance of an ad is determined by minimizing the amount paid for the maximum number of conversions. Creating high-performing ad creatives is a difficult but crucial task for advertisers.

The purpose of this study is to supporting the creation of ad creatives with many conversions, and we propose a new framework to support creating high-performing ad creatives, including accurate prediction of ad creative text conversions before delivery to the consumer[1]. If conversions of ad creatives can be predicted before delivery to consumers, advertisers can avoid the losses incurred by the high cost of ineffective advertisements. Moreover, because ad creatives with high click-through rates (CTRs), and low conversions have a tendency to deceive users, we also expect to

[*]This work was conducted while the first author was doing an internship at Gunosy Inc. We thank the ad engineering team who provided useful comments.
[1]We have also improved the CVR prediction using the result of conversion prediction.

improve the user experience on media displaying those ads. As a result, advertisers will be able to focus on improving the CTR of ad creatives.

Some attempts to support the creation of high-performing creatives by predicting ad creative conversions have been reported in the industry[2,3,4], but as far as we know, no academic research has been published in this area. Thomaidou et al. [34, 35] proposed a framework for generating ad creatives automatically. However, this framework focuses on search ads, and generates ad text according to set rules. Thus, this framework cannot be applied for our purpose. Some studies have reported that ad creatives affect the CTR of advertisements [1, 3, 8], but they do not predict the conversions. Prediction of a user's CTR or conversion rate (CVR) is a general task undertaken by many studies in this research area, but there are no studies that have predicted these rates for ad creatives. The prediction of an ad creative's performance is another important issue, but to the best of our knowledge, no study has examined this issue.

Although ad creatives are mainly image and text, we focus on the latter, and predicting its conversions. Because it is difficult to replace ad images, but easy to replace text, in this work, we propose a recurrent neural network (RNN)-based framework that predicts the performance of an ad creative text before delivery. The proposed framework includes three key ideas, namely, multi-task learning, conditional attention, and attention highlighting. Multi-task learning is an idea for improving the prediction accuracy of conversion, which predicts clicks and conversions simultaneously, to solve the difficulty of data imbalance. Conditional attention focuses on the feature representation of each creative based on its genre and target gender, thus improving conversion prediction accuracy. Attention highlighting visualizes important words and/or phrases based on conditional attention. We confirm that the proposed framework outperforms some baselines, and the proposed ideas are valid for conversion prediction. These ideas are expected to be useful for supporting the creation of ad creatives.

This research is motivated to support the creation of high performing creative text. The contributions are summarized as follows:

(1) We propose a new framework that accurately predicts ad creative performance.

    To realize this, we propose two key strategies to improve the prediction performance of advertisement conversion.

    (a) Multi-task learning predicts conversion, together with previous click actions, by learning common feature representations.
    (b) The Conditional attention mechanism focuses attention on the feature representation of each creative text considering the target gender and genre.

(2) We propose attention highlighting that offers important words and/or phrases using conditional attention.

A prototype implementation of the proposed framework with Chainer [36] has been released on GitHub[5].

[2]https://www.facebook.com/business/m/facebook-dynamic-creative-ads
[3]https://www.adobe.com/en/advertising/creative-management.html
[4]https://support.google.com/google-ads/answer/2404190?hl=en
[5]https://github.com/shunk031/Multi-task-Conditional-Attention-Networks

2 RELATED WORK

This study focuses on ad creatives. First, we describe existing studies that analyze high-performing ad creatives, and discuss how to generate them. Many studies on advertising creatives focus on images, and offer few results for texts. Furthermore, these studies focus on the CTR, rather than conversions. Second, we introduce studies on performance prediction for ads. In contrast to this study, which aims to predict the performance of new ads, these studies focus on images. Finally, highlighting studies related to our ideas, we introduce multi-task learning and RNN-based attention mechanisms.

2.1 Analysis and Generation of Effective Advertisements

Because ad creatives play an important role in the performance of ads, some studies analyzed ad creative performance [1, 3, 8]. For example, Azimi et al. [1] tried to predict some features of the CTR using ad creative images, and evaluated the effectiveness of visual features. The motivation of their study is similar to ours, but we focus on text instead of images in ad creatives and predict conversions rather than the CTR. Cheng et al. [8] proposed a model for predicting the CTR of new ads, and reported some knowledge using feature importance, but the text features of that study were based on fixed rules. With the development of deep learning, especially convolutional neural networks (CNNs) [21], visual features can be easily and effectively used for machine learning. Chen et al. [7] proposed Deep CTR, showing that using the features of ad images can significantly improve CTR prediction.

Thomaidou et al. [34] developed GrammaAds, which automatically generates keywords for search ads. In addition, they proposed an integrated framework for the automated development and optimization of search ads [35]. These studies support the creation of text ad creatives, but because these methods are rule-based, focusing only on search ads, the methods cannot be applied to display advertising.

2.2 CTR and Conversion Prediction in Display Advertising

CTR prediction of display advertising is important not only in the industry but also in academia. In [5, 31], a CTR prediction model was proposed using logistic regression (LR), and factorization machines (FMs) have been proposed to predict advertising performance [18, 19, 30]. In industry, LR and FMs are mainly used, because in display advertising, the prediction response time needs to be short to display an advertisement smoothly. In recent years, deep neural networks (DNNs) have been applied for predicting the advertisement CTR [7, 9, 13, 14, 23], and especially, some models combining DNNs with FMs have been proposed, and have improved predictions [9, 14, 23, 26]. The improvements achieved by these models show that explicit interaction between variables is important for advertisement performance prediction, so we adopted explicit interaction in our idea as a conditional attention mechanism.

There are several studies on CVR prediction [27, 29, 39], but there are not as many as the studies on CTR prediction. CVR prediction is difficult, because the number of conversions is imbalanced data that almost ad creative's conversions are zero. Existing studies tackled this difficulty. Yang et al. [40] adopted dynamic transfer learning for

predicting the CVR, and demonstrating feature importance. Punjobi et al. [29] proposed robust FMs for overcoming user response noise. In this study, we tackle this difficulty using multi-task learning.

2.3 Background of the Proposed Strategies

In this paper, we propose two key strategies for improving the prediction performance of advertisement conversion, namely, multi-task learning and a conditional attention mechanism. As the background of these strategies, we describe multi-task learning and the RNN-based attention mechanism.

Multi-task Learning. Multi-task learning [4] is a method that involves learning multiple related tasks. It improves the prediction performance by learning common feature representations. Recently, multi-task learning has been used in various research areas, especially natural language processing (NLP) [12, 28] and computer vision [10, 25, 41], and has achieved significant improvements. Conversions represent extremely imbalanced data, so conversion prediction is difficult. Because ad click actions represent a pre-action of conversion actions, click prediction may be related to conversion prediction. Therefore, we adopt multi-task learning, which predicts clicks and conversions simultaneously.

RNN-based Attention Mechanism. For supporting the creation of ad creative text, we use the knowledge of NLP. RNN-based models, such as long short-term memory (LSTM) [16], gated recurrent unit (GRU) [11], and attention mechanisms [2] have made breakthroughs in various NLP tasks, for example, machine translation [2], document classification [24, 40], and image captioning [38]. An RNN is a deep learning model for learning sequential data, and in NLP, this model can learn word order. Attention mechanisms compute an alignment score between two sources, and make significant improvements in some NLP tasks. Recently, self-attention [24], which computes alignment in a single source, was proposed. In addition, visual analysis using attention can highlight important phrases and/or words using the attention result, so the attention mechanism is also attractive for interpretability. In this study, we adopt a self-attention mechanism for improving conversion prediction performance and visualizing word importance.

3 METHODOLOGY

The outline of the proposed framework for evaluating ad creatives is shown in Figure 2. In the framework, we propose two strategies: multi-task learning, which simultaneously predicts conversions and clicks, and a conditional attention mechanism, which detects important representations in ad creative text according to the text's attributes.

Conversion prediction using ad creatives with an imbalanced number of conversions is a challenging task. Therefore, in multi-task learning, we expect to improve the model accuracy by predicting conversions along with clicks. The conditional attention mechanism makes it possible to dynamically compute attention according to the attributes of the ad creatives, its genre, and the target gender.

3.1 Framework Overview

The input of the proposed framework is ad creative text and ad creative attribute values. Figure 1 shows an example of an ad creative,

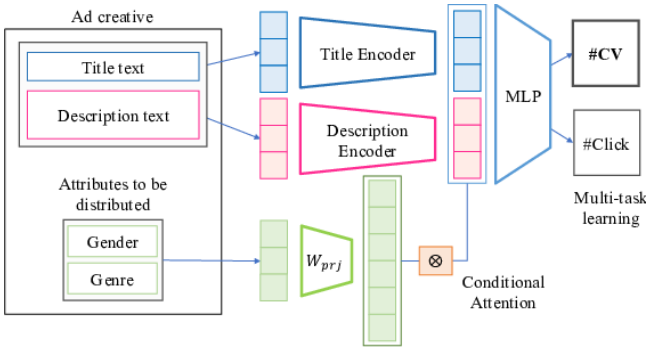


Figure 2: Outline of the proposed framework. In the framework, we propose two strategies: multi-task learning, which simultaneously predicts conversions and clicks, and a conditional attention mechanism, which detects important representations in ad creative text according to the text's attributes.

and these are two short texts which are called titles and descriptions. The ad attribute values are the gender of the delivery target and the genre of the ad creative, and they are related to the ad creatives.

Specifically, the input of the proposed framework is an ad creative text $S = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_n\}$ consisting of n word embeddings, where $\mathbf{w}_i \in \mathbb{R}^{d_w}$ represents the word vector at the i-th position in the ad creative text. Therefore, $S \in \mathbb{R}^{n \times d_w}$ is a two-dimensional matrix of the word sequence.

Incidentally, in the practical situation, a number of ad creative texts that have title and description texts are created for the target product. These texts often have different contexts for maximizing the amount of information empirically. Therefore, the proposed framework uses two *text encoders*, which learn the individual context from the title and the description.

As a *text encoder*, we adopted the GRU, which can extract features from ad creative text considering word order. Specifically, title text $S^{\text{title}} = \{\mathbf{w}_1^{\text{title}}, \mathbf{w}_2^{\text{title}}, \cdots, \mathbf{w}_n^{\text{title}}\}$ and description text $S^{\text{desc}} = \{\mathbf{w}_1^{\text{desc}}, \mathbf{w}_2^{\text{desc}}, \cdots, \mathbf{w}_n^{\text{desc}}\}$ are input from the ad creative into title and description encoders, respectively, and are encoded into feature representations as $\mathbf{h}_t^{\text{title}} \in \mathbb{R}^{u_{\text{title}}}$ and $\mathbf{h}_t^{\text{desc}} \in \mathbb{R}^{u_{\text{desc}}}$; $t = 1, 2, \cdots, n$:

$$\begin{aligned} \mathbf{h}_t^{\text{title}} &= \text{title encoder}(\mathbf{w}_t^{\text{title}}, \mathbf{h}_{t-1}^{\text{title}}), \\ \mathbf{h}_t^{\text{desc}} &= \text{description encoder}(\mathbf{w}_t^{\text{desc}}, \mathbf{h}_{t-1}^{\text{desc}}). \end{aligned} \tag{1}$$

Let u_{title} and u_{desc} be the number of hidden units of the title and description encoders obtained here. The n hidden states can be expressed as $H^{\text{title}} = \{\mathbf{h}_1^{\text{title}}, \cdots, \mathbf{h}_n^{\text{title}}\}$ and $H^{\text{desc}} = \{\mathbf{h}_1^{\text{desc}}, \cdots, \mathbf{h}_n^{\text{desc}}\}$, respectively. Compute a vector $\mathbf{x}_{\text{feats}}$ that concatenates these hidden states, H^{title}, H^{desc}, one-hot vectors of gender features $\mathbf{x}_{\text{gender}} \in \mathbb{R}^{d_{\text{gender}}}$, and genre features $\mathbf{x}_{\text{genre}} \in \mathbb{R}^{d_{\text{genre}}}$:

$$\mathbf{x}_{\text{feats}} = \text{concat}(H^{\text{title}}, H^{\text{desc}}, \mathbf{x}_{\text{genre}}, \mathbf{x}_{\text{gender}}). \tag{2}$$

Note, $\mathbf{x}_{\text{feats}} \in \mathbb{R}^{d_{\text{feats}}}$; $d_{\text{feats}} = n \times (u_{\text{title}} + u_{\text{desc}}) + d_{\text{gender}} + d_{\text{genre}}$. These concatenated vectors are inputted in a multi-layer perceptron (MLP) which is an output layer of the proposed framework. To predict conversions $\hat{y}^{(\text{cv})}$ and clicks $\hat{y}^{(\text{click})}$, multi-task learning

described later predicted $\hat{\mathbf{y}}_{\text{multi}} = \{\hat{y}^{(\text{cv})}, \hat{y}^{(\text{click})}\}$ through the MLP:

$$\hat{\mathbf{y}}_{\text{multi}} = \text{MLP}(\mathbf{x}_{\text{feats}}). \tag{3}$$

To improve the performance of the model robustness, we use wild-card training [32] with dropout [15] for the input word embeddings.

3.2 Multi-task Learning

Conversion prediction is difficult, due to the imbalanced data, so we use the strategy of multi-task learning. Multi-task learning is a method that solves multiple tasks related to each other, and that improves the prediction performance by learning common feature representations. We adapt multi-task learning, and predict clicks and conversions prediction simultaneously. Because click prediction may be related to conversion prediction, we expect to improve the prediction performance by learning common feature representations using multi-task learning.

In multi-task learning, the input is a feature vector of a training sample denoted by $\mathbf{x}$, and the ground truth is y. For training samples $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N\}$, a single model, f, learns to generate predictions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \cdots, \hat{y}_N\}$:

$$\hat{y} = f(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N). \tag{4}$$

We minimize the mean squared error (MSE) over all samples, N, in $l = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$. In K supervised tasks, the multi-task model, $F = \{f_1, f_2, \cdots, f_K\}$, learns to generate predictions $\hat{\mathbf{y}} = \{\hat{y}^{(1)}, \hat{y}^{(2)}, \cdots, \hat{y}^{(K)}\}$:

$$\hat{\mathbf{y}} = F(\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N). \tag{5}$$

The total loss is calculated from the sum of loss in each task,

$$\mathcal{L} = \frac{1}{N}\sum_{k=1}^{K}\sum_{i=1}^{N}\left(y_i^{(k)} - \hat{y}_i^{(k)}\right)^2. \tag{6}$$

In this task, for ground truth of $y^{(\text{cv})}$ and $y^{(\text{click})}$, we minimize losses for predicted conversions $\hat{y}^{(\text{cv})}$ and clicks $\hat{y}^{(\text{click})}$:

$$\mathcal{L}_{\text{multi}} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{(\text{cv})} - \hat{y}_i^{(\text{cv})}\right)^2 + \lambda\frac{1}{N}\sum_{i=1}^{N}\left(y_i^{(\text{click})} - \hat{y}_i^{(\text{click})}\right)^2, \tag{7}$$

where $\lambda > 0$ is the hyper-parameter to control the effect of the click loss.

3.3 Conditional Attention

We propose the strategy of the conditional attention mechanism. Supporting the creation of ad creatives by considering attribute values is useful, but the conventional attention mechanism learns keywords or key phrases, by calculating the alignment score using only the input sentence.

In this paper, we propose a conditional attention mechanism to calculate self-attention, using feature vectors obtained from the attribute values of the ad creative. Figure 3 illustrates the conditional attention mechanism. It can consider ad creative attributes against the conventional attention mechanism.

The conditional attention mechanism is calculated from the attention of the *text encoder* and the feature vector obtained from the attribute values of the ad creative text. Each word in the word sequence S is independent of the others. To capture these word order relations, we apply a *text encoder* to the text, to obtain the hidden

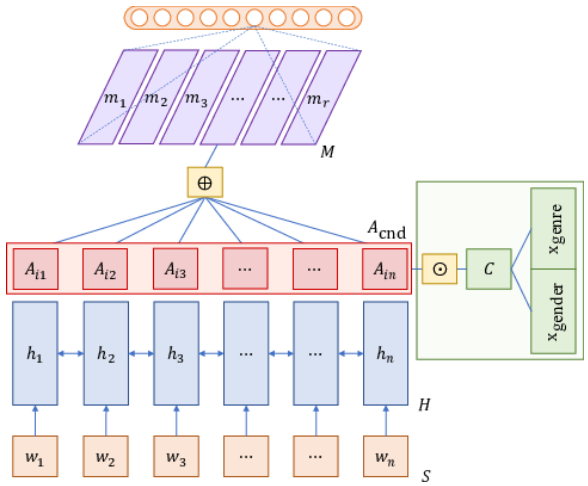


Figure 3: Example of the conditional attention mechanism. Conditional attention is calculated from the element-wise product of the attention matrix A and the feature vector $\mathbf{c}$ consisting of the gender and the genre.

state $\mathbf{h}_t \in \mathbb{R}^u$. The n hidden states of these $u \times n$ dimensions can be expressed as $H = \{\mathbf{h}_1, \mathbf{h}_2, \cdots, \mathbf{h}_n\}$.

To consider ad attribute values, a *conditional* vector, $\mathbf{c} \in \mathbb{R}^n$, is calculated by performing a linear combination of $\mathbf{x}_{\text{feats}} \in \mathbb{R}^{d_{\text{feats}}}$ and trainable parameters $W_{\text{prj}} \in \mathbb{R}^{n \times d_{\text{feats}}}$:

$$\mathbf{c} = W_{\text{prj}}\mathbf{x}_{\text{feats}}. \tag{8}$$

Here, we use *self-attention* [24] for computing the linear combination. The attention mechanism takes the entire hidden state H of the *text encoder* as the input and outputs attention vector $\mathbf{a}$:

$$\mathbf{a} = \text{softmax}(\mathbf{w}_{s2}^T \tanh(W_{s1}H)), \tag{9}$$

where $W_{s1} \in \mathbb{R}^{n \times u}$ and $\mathbf{w}_{s2} \in \mathbb{R}^n$ are trainable parameters. Because H is an $n \times u$ dimension, the size of attention vector $\mathbf{a}$ is n. The softmax$(\cdot)$ is calculated so that the sum of all the weight is 1.

Furthermore, we calculate the *conditional attention vector* using the attributes given to the ad creative. The *conditional attention* vector, $\mathbf{a}_{\text{cnd}}$, is calculated using conditional vector $\mathbf{c}$ and attention vector $\mathbf{a}$:

$$\mathbf{a}_{\text{cnd}} = \mathbf{a} \odot \mathbf{c}. \tag{10}$$

Here, $\odot$ is an element-wise product. We want r different parts to be extracted from the ad creative texts. Thus, the *conditional attention* vector $\mathbf{a}_{\text{cnd}}$ becomes *conditional attention* matrix $A_{\text{cnd}} \in \mathbb{R}^{n \times r}$. Therefore, sentence vector $\mathbf{m}$ with the embedded ad creative text becomes sentence matrix $M \in \mathbb{R}^{u \times r}$. The *conditional attention* matrix, A_{cnd}, is multiplied by hidden state H of the *text encoder*, and the r-weighted sentence matrices are calculated as follows:

$$M = HA_{\text{cnd}}. \tag{11}$$

In the proposed framework, the model makes predictions based on the calculated M and ad creative attributes, such as $\mathbf{x}_{\text{gender}}$ and $\mathbf{x}_{\text{genre}}$.

Table 1: Features included in the ad creative dataset. It contains 1,694 campaigns, some of which were part of campaigns delivered by Gunosy. The average lengths of the title and description texts are about 15 and, 32 characters, respectively. The Campaign ID feature is not directly inputted in the model, because the ID is used for evaluations with cross-validation based on the ID.

	Features	Feature Description	Details
	Campaign ID	Campaign ID in Gunosy Ads	1,694 campaigns
Texts	Title	Title texts	Avg. 15.44±3.16 chars
	Description	Description texts	Avg. 32.69±5.43 chars
Attrs	Genre	Genre of the creatives	20 types
	Gender	Gender of delivery target	3 types

4 EXPERIMENTS

4.1 Dataset

We use real-world data from the Japanese digital advertising program Gunosy Ads[6], provided by Gunosy Inc.[7]. Gunosy Inc. is a provider of several news delivery applications, and Gunosy Ads delivers digital advertisements for these applications. Gunosy is a news delivery application that achieved more than 24 million downloads in January 2019.

For evaluation, we used 14,000 ad creatives delivered by Gunosy Ads from August 2017 to August 2018. In digital advertising, the cost of acquiring a conversion is called the cost per acquisition (CPA). Advertisers set target CPAs for a product, and manage its ad creatives to improve their performance. When the target CPAs for creatives are different, the trend of conversions may also vary, and for this reason, the dataset we selected comprises ad creatives where the target CPA was within a certain range. In addition, we removed creatives with a low number of impressions[8] from the dataset. As shown in Table 1, the title, description, and genre of the ad creative, as well as the gender to which the ad is delivered, are used as input features. Note that the Campaign ID is not a feature directly used as an input in the model, because the ID is used for evaluating with cross-validation based on the ID.

Creative texts written in Japanese are split into words using MeCab [22], a morphological analysis engine for Japanese texts, and mecab-ipadic-neologd [37], which is a customized system dictionary that includes many neologisms for MeCab. The number of clicks and conversions is log-normalized.

Figure 4 shows a histogram of the number of clicks and conversions. The number of conversions is concentrated on zero, and in relation, the number of clicks is a long-tailed distribution. Therefore, the ad creative dataset is definitely imbalanced. Figure 5 shows the distribution between the number of clicks and conversions in the dataset. The correlation coefficient between the number of clicks and conversions is 0.816, which is a strong correlation. As a reminder, we hide the number of clicks and conversions, also their frequencies, for confidentiality reasons.

[6]https://gunosy.co.jp/ad/
[7]https://gunosy.co.jp/en/
[8]An occasion when a particular advertisement is seen by someone using the application.

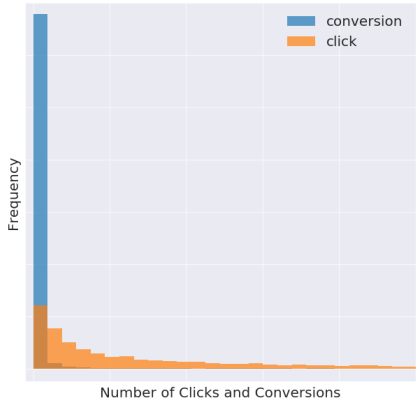


Figure 4: Distribution of clicks and conversions in the dataset. The number of conversions is concentrated on zero. Compared with the number of conversions, the number of clicks indicates a long-tail distribution.

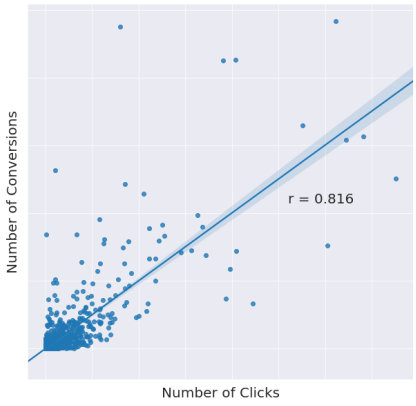


Figure 5: The linear relation between the clicks and conversions in the dataset (correlation coefficient $r = 0.816$).

4.2 Experimental Settings

In these experiments, support vector regression (SVR) and an MLP-based *text encoder* were used as a baseline model. When inputting creative text in the SVR model, we used average-pooled sentence representations computed from word representations, using pre-trained word2vec (w2v) [33]. The same pre-trained w2v was used as word embedding for the proposed model.

We compared and examined the following models: MLP (not considering word order) and GRU (considering word order) as the *text encoder* in the proposed framework. LSTM was also considered as a candidate for the baseline model; however, it showed no improvement in performance, so it was excluded from the experiment. In addition, CNNs are known to be capable of training at high speed, because they can perform parallel calculations, compared with LSTM and GRU, and their performances are also known to be equal. Nevertheless, these methods were excluded in these experiments, because we were targeting an RNN-based model that can apply attention for visualizing the contributions of words to ad creative evaluation.

Table 2: Comparison of the prediction performance of CVs in mean squared error (MSE) criteria. The proposed multi-task learning and conditional attention reduced MSE in almost all the categories, especially estimating cases where the number of conversions (#CV) is one or more (#CV > 0). However, "All predicted as zero" showed sufficiently low MSE in this category, due to too many CV = 0 in this dataset. Therefore, we conclude using MSE as an evaluation metric is not suitable in this study.

Model		MSE			
		All		#CV >0	
		Single-task	**Multi-task**	Single-task	**Multi-task**
MLP		0.01712	0.01698	0.04735	0.03199
GRU	Vanilla	0.01696	0.01695	0.04657	0.04355
	Attention	0.01685	0.01688	0.04695	0.03105
	Conditional attention	**0.01683**	**0.01675**	**0.04641**	**0.02825**
All predicted as zero		**0.02148**		–	

We compared the proposed models used in the proposed framework. The following models were compared and examined, to confirm the effect of multi-task learning in conversion prediction:

Single-task: A commonly known model that predicts conversions only; and

Multi-task: A model that simultaneously predicts the number of clicks and the number of conversions.

To confirm the effect of the conditional attention mechanism, we compared the following models:

Vanilla: A simple *text encoder* without an attention mechanism. It is a baseline in the proposed model;

Attention: A mechanism that introduces self-attention to the *text encoder*. It makes it possible to visualize which word contributed to prediction during creative evaluation; and

Conditional Attention: A mechanism introduced to the *text encoder* of the proposed method. Conditional attention can be computed and visualized considering the attribute values of the ad creative. Different attentions can be visualized by changing the attribute value for the same creative text.

In addition, the hyper-parameter setting is described below. The mini-batch size was set to be 64, and the number of epochs was set to be 50. For multi-task learning, we used a fixed value of $\lambda = 1$. In the *text encoder*, the number of hidden units was set to be 200 for u_{title} and u_{desc}. For all models, we use Adam [20], with a weight decay of $1e^{-4}$, for parameter optimization.

4.3 Evaluation Metrics

First, as evaluation metrics, we adopt not only MSE but also normalized discounted cumulative gain (NDCG) [17], which is evaluation metrics for ranking. MSE measures the average of the squares of the errors, which is the average squared difference between the estimated values and what is estimated. We adopted ranking evaluation metrics because the number of conversions is imbalanced. As shown in Figure 4, most ad creative conversions are zero and imbalanced. A high evaluation score can be achieved by an overfit model that predicts all outputs as zero when such metrics are used. For the creation of high-performing ad creatives, rather than predicting zero conversions, we would like to accurately predict high-conversion creatives as such.

NDCG is mainly used in the experiments. NDCG is the discounted cumulative gain (DCG) normalized score. In DCG, the score decreases as the evaluation of an advertisement declines, so a penalty is imposed if a low effect is predicted for highly effective creatives. At the time of the NDCG calculation, after obtaining the rank of the ground truth, and its predicted value, respectively, evaluation scores are calculated for all the evaluation data, as well as those restricted to the top 1% of conversions.

For ad creative evaluation, the metrics are computed with cross-validation. In most advertising systems, advertisements are delivered in units of campaigns. In a campaign, the target gender and its genre are set, and multiple ad creatives are developed.

In this paper, we predict the number of conversions for ad creative text in unknown campaigns, and confirm the generalization performance of the proposed framework. Therefore, at the time of the evaluation, five-fold cross-validation was performed in such a manner that the delivered campaigns did not overlap.

4.4 Experimental Results

For confirming the accuracy of the proposed framework compared with the baselines, we compared single-task and multi-task learning, and the results of the application of the conditional attention mechanism are described. Through almost all the results, the proposed framework applying multi-task learning and the conditional attention mechanism achieved a better performance than the other methods. Especially, when focusing on ad creatives with many conversions, the proposed framework achieved high prediction accuracy.

Table 2 shows the MSE score with all the evaluations in each model, and with one or more conversions in each model. Almost all the results show that the model applying the multi-task learning and conditional attention mechanism had a smaller MSE score than the other models did. Overall, the RNN-based GRU showed better performance than the baseline models. Therefore, the results suggest that it is important to properly capture word order when evaluating creative texts. Compared with *vanilla* and *attention*, in the proposed model, *conditional attention* showed a better performance.

Although the improvement of all datasets is weak, because as shown in Figure 4, the number of conversions of many ad creatives

Table 3: Comparison of the normalized discounted cumulative gain (NDCG) in the proposed model. When calculating NDCG scores, the results for all data and the scores restricted to the top 1% of conversions (#CV) were calculated.

Model		NDCG [%]			
		All		#CV top 1 %	
		single	**multi-task**	single	**multi-task**
SVM		96.72		83.73	
MLP		96.68	97.18	82.97	84.12
GRU	Vanilla	96.54	97.00	76.39	78.51
	Attention	96.76	97.11	83.00	85.49
	Conditional Attention	**96.77**	**97.20**	**87.11**	**87.14**

Table 4: Comparison of NDCG between the CVR directly predicted by the single-task model and the CVR (#conversions / #clicks) calculated from the multi-task GRU model's predicted conversions and clicks. The threshold value for calculating NDCG is assumed to be a CVR of 0.5 or more.

Model		NDCG [%]
Single-task	Vanilla	80.54
	Attention	82.58
	Conditional attention	83.89
Multi-task	Vanilla	82.63
	Attention	84.27
	Conditional attention	**85.61**

is zero, the MSE is small, even if the conversion of most ad creatives is predicted to be zero. Therefore, we evaluated data with conversions other than zero. As a result, we found that the proposed model exhibits much better performance than the baseline model for data with one or more conversions. The proposed model was able to predict creatives with more conversions than the baseline models.

To evaluate ad creatives with many conversions as such, we used the ranking algorithm NDCG. The NDCG result in the proposed model is shown in Table 3[9]. The NDCG score (regarded as *All* in Table 3) for all the datasets is shown for reference, because as noted above, most samples have zero conversions. The performance of the GRU model that considers word order compared with the baseline model improved by an average of approximately 3-5%, with many conversions.

In the NDCG result (Table 3), the multi-task model realized higher prediction accuracy than the single-task model predicting only conversions did. A score improvement of approximately 1-2% was confirmed when compared with the baselines. Because clicks are highly correlated with target ad conversions, as shown in Figure 5, rather than predicting conversions alone, training the model to multi-task by predicting clicks simultaneously can improve prediction accuracy. By training clicks and conversions, the proposed model seems to implicitly learn features that contribute to conversion prediction.

[9]The same tendency was observed even when mean average precision (MAP) was used as an evaluation metric.

Table 5: Comparison of GRU models for creative texts and their attribute value interactions. Performance is improved using conditional attention rather than giving attribute values directly to word vectors.

Model		NDCG [%]	
		Single-task	**Multi-task**
w2v + attributes	Vanilla	77.84	78.03
	Attention	80.39	83.52
w2v	**Conditional attention**	**87.11**	**87.14**

Because several previous studies predicted the CVR directly, we also calculated it, using the prediction of the multi-task learning model, and compared the accuracy. In a multi-task model, the CVR can be calculated by dividing conversions by clicks. In Table 4, the multi-task model is compared with the single-task model by directly estimating the CVR. The prediction performance of the multi-task model is higher than that of the single-task model. Although the number of clicks and conversions predicted by multi-task learning may not always be close to the ground truth, the ratio of the number of clicks to the conversion number is captured properly.

In Table 3, the conditional attention mechanism achieved better results the NDCG metric. In particular, the conditional attention mechanism showed better results than the conventional attention mechanism did. In the conventional attention mechanism, the training was focused solely on the co-occurrence relation between words in the input text, but the conditional attention mechanism can predict conversion by using the attribute value.

Table 5 shows the result comparing feature interaction between w2v-based embeddings and ad attribute values. In the proposed framework, this interaction is realized with the conditional attention mechanism, explicitly. Because attention is computed by the input variables, this interaction is implicitly expressed by inputting both variables in the *text encoder*. For confirming the effect of this explicit interaction in the conditional attention mechanism, we compared the model that inputted both variables in the *text encoder* with the conditional attention mechanism. The conditional attention mechanism showed the best performance in the single-task and multi-task model. Introducing the vanilla model and the conventional attention model to the word representation with ad attribute values resulted in a poor performance, mainly because the

duplicate interactions were calculated excessively. It is suggested that it is better to introduce the explicit interaction of attribute values.

5 DISCUSSION

5.1 Advantages of the Proposed Framework

The proposed framework aimed to predict not the CVR but conversions. However, in CVR prediction, we also achieved high performance using multi-task learning results. From the business perspective, we assume that predicting conversions can evaluate high-performing ad creatives, rather than predicting the CVR. In the process of advertising management, advertisers stop low-performing creatives and focus cost on high-performing creatives, so there are few conversions of low-performing creatives, and many conversions of high-performing creatives. For that reason, the number of conversions seems to be a good metric for evaluating ad creatives, and conversion prediction may be learn good representation of high-performing ad creatives.

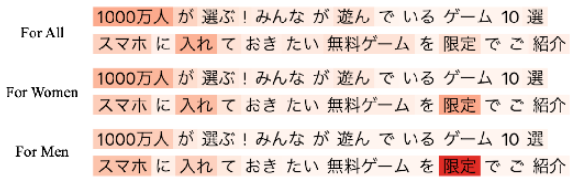
We proposed an RNN-based framework, and achieved high-performance conversion prediction. Normally, when advertisers create the creative text, words are selected in such a way as to change the word order or emphasize the characteristics of the product. We let the model learn feature representation so that it could properly capture the features between words in creative text.

We achieved high-performance conversion prediction by predicting the clicks and conversions simultaneously; this method is called multi-task learning. Many ad creative conversions are zero, which is imbalanced data, so predicting this number correctly is a difficult task. Multi-task learning is a method that learns multiple tasks related to each other, and improves prediction performance. Because ad click actions represent the pre-action of conversion actions, we assumed that click prediction may be related to conversion prediction, and that improved conversion prediction would be obtained using multi-task learning. We expect that this achievement can be applied to various prediction tasks with imbalanced data.

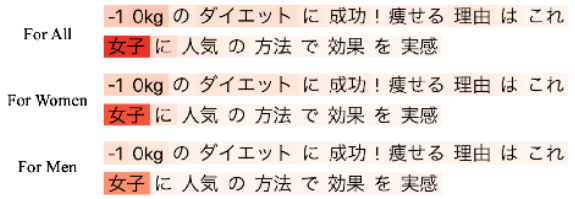
High accuracy was achieved by conditional attention in the experiment. When predicting the CTR or CVR of advertisements, it is important to properly capture the explicit feature interactions [23]. The conditional attention mechanism seems to capture the explicit interactions between the attention gained from creative text and feature representations consisting of the text's attribute values. It is also possible to visualize different forms of attention by controlling different attribute values in the same creative texts. This can greatly support ad creative creation.

5.2 Visualization for High-Performance Ad Creative Creation

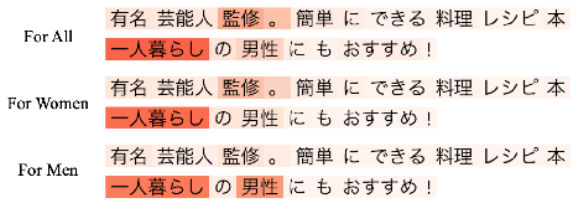
We attempt to highlight important words using attention. If the words contributing to conversions are clarified, advertisers will be able to easily create high-performing ad creatives. Attention is a mechanism that focuses on words contributing to prediction, and the results predicted by these mechanisms are useful for creating ad creatives. The proposed conditional attention mechanism can compute attention based on ad creative attributes, as well as the genre and target gender, so conditional attention highlights important words according to their attribute values.



(a) Title text: "Chosen by 10 million people! The 10 games played by everyone." Description text: "Exclusively introducing free games that you will want to install on mobile phone."



(b) Title text: "Success in -10 kg weight loss! This is the reason for getting slim." Description text: "Realizing the effects popular among girls."



(c) Title text: "Supervised by a famous celebrity; easy cookbook." Description text: "Recommended for men living alone!"

Figure 6: Heatmap showing the change in conditional attention when the distribution target is changed.

Figure 6 shows examples of the visualization of attention when modifying the attributes of gender for three Japanese ad creative texts for different groups (for all audiences, for women, and for men). Different types of attention were gained using conditional attention mechanism.

Figure 6a shows an ad creative for a mobile game. The word "1000万" (*10 million*), a concrete numerical value, and the word "限定" (*exclusively*) contribute to predicting conversion. Especially for men, the word "限定" contributes more to the prediction than it does for women.

Figure 6b is an ad creative in the beauty genre for women. The word "女性" (*girls*) contributes to the conversion prediction. More attention is also given to "ダイエット" (*weight loss*) for women than men. When setting the delivery target to men in this ad creative, the attention score and the number of predicted conversions are smaller than that of all targets or female targets.

Figure 6c is an ad creative in the health food genre for men. The words "一人暮らし" (*living alone*) and "監修" (*supervised by*) are closely highlighted. The word "lived alone" is an expression that narrows down the delivery target. When proposing ad creative text, the term "supervised by" is often used in conjunction with the names of celebrities, and the effect is high. Moreover, it was confirmed that the word "男性" (*men*) is an important factor when the delivery target is male.

Overall, most words that are highlighted by attention are concrete numerical values and expressions focusing on the delivery target. We believe that this knowledge is also empirically correct. In this way, visualization of important words using the conditional attention mechanism of the proposed method can be expected to greatly contribute to supporting the creation of ad creatives. This result is a good example of interpretability.

6 CONCLUSION

In this paper, we propose a new framework to support the creation of high-performing ad creative text. The proposed framework includes three key ideas, multi-task learning and conditional attention improve prediction performance of advertisement conversion, and attention highlighting offers important words and/or phrases in text creatives. We confirmed that the proposed framework realizes an excellent performance thanks to these ideas, through experiments with actual delivery history data.

In the future, we will build a framework that simultaneously uses images attached to ad creatives, and aim to improve the accuracy of conversion prediction.

REFERENCES

[1] Javad Azimi, Ruofei Zhang, Yang Zhou, Vidhya Navalpakkam, Jianchang Mao, and Xiaoli Fern. 2012. Visual Appearance of Display Ads and Its Effect on Click Through Rate. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*. 495–504.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *International Conference on Learning Representations* (2014).

[3] Norris I Bruce, BPS Murthi, and Ram C Rao. 2017. A dynamic model for digital advertising: The effects of creative format, message content, and targeting on engagement. *Journal of marketing research* 54, 2 (2017), 202–218.

[4] Rich Caruana. 1997. Multitask learning. *Machine learning* 28, 1 (1997), 41–75.

[5] Deepayan Chakrabarti, Deepak Agarwal, and Vanja Josifovski. 2008. Contextual advertising by combining relevance with click feedback. In *Proceedings of the 17th international conference on World Wide Web*. 417–426.

[6] Olivier Chapelle, Eren Manavoglu, and Romer Rosales. 2015. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 4 (2015), 61.

[7] Junxuan Chen, Baigui Sun, Hao Li, Hongtao Lu, and Xian-Sheng Hua. 2016. Deep ctr prediction in display advertising. In *Proceedings of the 2016 ACM on Multimedia Conference*. 811–820.

[8] Haibin Cheng, Roelof van Zwol, Javad Azimi, Eren Manavoglu, Ruofei Zhang, Yang Zhou, and Vidhya Navalpakkam. 2012. 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*. 777–785.

[9] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, et al. 2016. Wide & deep learning for recommender systems. In *Proc. of the 1st Workshop on Deep Learning for Recommender Systems*. 7–10.

[10] Xiao Chu, Wanli Ouyang, Wei Yang, and Xiaogang Wang. 2015. Multi-task recurrent neural network for immediacy prediction. In *Proc. of the IEEE international conference on computer vision*. 3352–3360.

[11] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Advances in neural information processing systems Workshop*.

[12] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12, 8 (2011), 2493–2537.

[13] Paul Covington, Jay Adams, and Emre Sargin. 2016. Deep neural networks for youtube recommendations. In *Proc. of the 10th ACM Conference on Recommender Systems*. 191–198.

[14] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2014. Deepfm: a factorization-machine based neural network for ctr prediction. *CoRR arXiv:1703.04247* (2014).

[15] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR arXiv:1207.0580* (2012).

[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[17] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (2002), 422–446.

[18] Yuchin Juan, Damien Lefortier, and Olivier Chapelle. 2017. Field-aware factorization machines in a real-world online advertising system. In *Proc. of the 26th International Conference on World Wide Web Companion*. 680–688.

[19] Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. 2016. Field-aware factorization machines for CTR prediction. In *Proc. of the 10th ACM Conference on Recommender Systems*. 43–50.

[20] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR arXiv:1412.6980* (2014).

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[22] Taku Kudo. 2006. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab.sourceforge.jp*.

[23] Jianxun Lian, Xiaohuan Zhou, Fuzheng Zhang, Zhongxia Chen, Xing Xie, and Guangzhong Sun. 2018. xDeepFM: Combining Explicit and Implicit Feature Interactions for Recommender Systems. In *Proc. of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

[24] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *International Conference on Learning Representations* (2017).

[25] Wu Liu, Tao Mei, Yongdong Zhang, Cherry Che, and Jiebo Luo. 2015. Multi-task deep visual-semantic embedding for video thumbnail selection. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. 3707–3715.

[26] Weiwen Liu, Ruiming Tang, Jiajin Li, Jinkai Yu, Huifeng Guo, Xiuqiang He, and Shengyu Zhang. 2018. Field-aware Probabilistic Embedding Neural Network for CTR Prediction. In *Proc. of the 12th ACM Conference on Recommender Systems*. 412–416.

[27] Quan Lu, Shengjun Pan, Liang Wang, Junwei Pan, Fengdan Wan, and Hongxia Yang. 2017. A Practical Framework of Conversion Rate Prediction for Online Display Advertising. In *Proc. of the ADKDD'17*. 9:1–9:9.

[28] Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task Sequence to Sequence Learning. In *International Conference on Learning Representations*.

[29] Surabhi Punjabi and Priyanka Bhatt. 2018. Robust Factorization Machines for User Response Prediction. In *Proc. of the 2018 World Wide Web Conference*. 669–678.

[30] Steffen Rendle. 2010. Factorization machines. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 995–1000.

[31] Matthew Richardson, Ewa Dominowska, and Robert Ragno. 2007. Predicting Clicks: Estimating the Click-Through Rate for New Ads. In *Proc. of the 16th International World Wide Web Conference(WWW-2007)*.

[32] Daiki Shimada, Ryunosuke Kotani, and Hitoshi Iyatomi. 2016. Document classification through image-based character embedding and wildcard training. In *Big Data (Big Data), 2016 IEEE International Conference on*. 3922–3927.

[33] Masatoshi Suzuki, Koji Matsuda, Satoshi Sekine, Naoaki Okazaki, and Kentaro Inui. 2018. A Joint Neural Model for Fine-Grained Named Entity Classification of Wikipedia Articles. *IEICE Transactions on Information and Systems* E101.D, 1 (2018), 73–81.

[34] Stamatina Thomaidou, Konstantinos Leymonis, and Michalis Vazirgiannis. 2013. GrammAds: Keyword and ad creative generator for online advertising campaigns. In *Digital Enterprise Design and Management 2013*. 33–44.

[35] Stamatina Thomaidou, Kyriakos Liakopoulos, and Michalis Vazirgiannis. 2014. Toward an integrated framework for automated development and optimization of online advertising campaigns. *Intelligent Data Analysis* 18, 6 (2014), 1199–1227.

[36] Seiya Tokui, Kenta Oono, Shohei Hido, and Justin Clayton. 2015. Chainer: a Next-Generation Open Source Framework for Deep Learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*. http://learningsys.org/papers/LearningSys_2015_paper_33.pdf

[37] Sato Toshinori. 2015. Neologism dictionary based on the language resources on the Web for Mecab. https://github.com/neologd/mecab-ipadic-neologd

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. [n. d.]. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[39] Hongxia Yang, Quan Lu, Angus Xianen Qiu, and Chun Han. 2016. Large Scale CVR Prediction through Dynamic Transfer Learning of Global and Local Features. In *Proc. of the 5th International Workshop on Big Data, Streams and Heterogeneous Source Mining: Algorithms, Systems, Programming Models and Applications at KDD 2016*. 103–119.

[40] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1480–1489.

[41] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2014. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*. 94–108.