POPCTR: A MULTI-MODEL ARCHITECTURE FOR CLICK-THROUGH RATE PREDICTION OF BANK POP-UP ADVERTISEMENTS

Anonymous ICME submission

ABSTRACT

Bank pop-up advertisement is one of the most important ways to attract new customers and activate sleeping old customers. Among advertising systems, click-through rate (CTR) prediction is crucial. Different from e-commerce scenario, bank pop-up ads have their special visual characteristics, with some text describing hot buttons (e.g., low loan rate), which makes traditional CTR models lose their roles. In this paper, we propose a multi-model architecture for click-through rate prediction. Specifically, we first extract global and local image features based on ads' layouts. Next, for local image features, we obtain their corresponding local text features. Then, we get final ad representations by combining image features with local text features. That is, we employ an attention aggregation over multi-modal features by considering position information. Finally, we capture users' temporal interests from history click sequences. Experimental results demonstrate our model greatly outperforms previous state-of-the-art in real bank pop-up ad datasets.

***Index Terms*—** pop-up ad, click-through rate, multi-model, attention aggregation

1. INTRODUCTION

With the deepening of banking digitization, online marketing has been moving closer to the business and assuming more responsibilities, in which pop-up advertisements play an important role. Delivering the right ads to the right people, in the right context, at the right time could help banks to efficiently attract new customers and activate sleeping old customers. Therefore, in order to benefit both user experience and business revenue, it is crucial to estimate the Click-Through Rate (CTR) of ads accurately.

Early CTR models usually utilize non-visual features (e.g., ads ID, category [1, 2], and temporal information [3, 4]). However, these methods are insufficient, since most of the ads are displayed with images nowadays. This phenomenon encourage studies on visual aware CTR models [5, 6, 7], which first extract visual features, and then fuse them with non-visual features. To increase the prediction accuracy, more deliberate visual feature extraction methods are proposed [8], and some work also include text features to build multi-model [9] CTR prediction. Although all these methods have achieved excellent performance on public datasets, they could not be applied to bank pop-up ads for its specific characters.

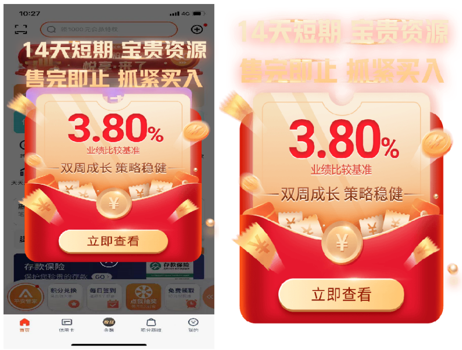

Fig. 1. Home page and pop-up ad

In this paper, we aim at pop-up ads at bank's home page (Fig.1). Different from traditional e-commerce scenarios, bank advertisements aim at financial services and products, most of which are lacking in physical substance. As a result, bank ads are required to include texts to describe the products. According to bank ads designers, ads should describe directly to customers' emotional hot buttons, like a low loan rate or a cash reward. With this goal in mind, most pop-up ads are visually divided into three parts, as shown in the right part of Fig.1. The ad has an eye-catching pattern with hot buttons (better rates) in the middle, ad theme (14-days short-term deposits) at the top of the image, and click link at the bottom.

Based on these observations, we propose a multi-model architecture for click-through rate prediction (PopCTR). Specifically, we first extract global and local image features based on ads' layouts. Next, for local image features, we obtain their corresponding local text features. Then, we aggregate image features and local text features based on their position information to get final ad representation vectors. Finally, we capture users' temporal interests from history click sequences. Experimental results demonstrate PopCTR greatly outperforms previous state-of-the-art in real bank pop-up ad datasets. In summary, we make the following contributions:

- We propose a structure to integrate customers' temporal information and advertisements' multi-model features.
- We design a multi-modal architecture for advertisements to combine image and text features based on their position information.

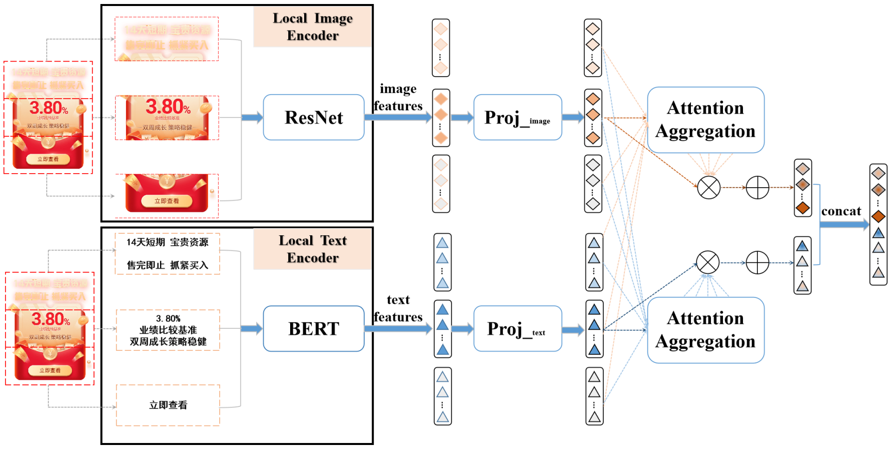

Fig. 2. The main parts of PopCTR. Image feature encoder use ResNet to get global and local image features. Text feature encoder employ BERT to get local text features. Attention aggregation aggregate multi-model data based on position information.

- Our proposed architecture achieves state-of-the-art performance than other baselines on real dataset.

2. OUR PROPOSED MODEL

Let $\mathcal{U} = \{u_1, u_2, \ldots\}$ and $\mathcal{I} = \{i_1, i_2, \ldots\}$ denote the sets of users and ads, respectively. User historical behaviors are presented by a ads sequence in the chronological order $\mathcal{I}^u = \{i^u_1, \ldots, i^u_{|\mathcal{I}^u|}\}$, where i^u_j is the j-th ad user u has clicked. The user-ad interaction matrix $\boldsymbol{Y} = \{y_{ui} | u \in \mathcal{U}, i \in \mathcal{I}\}$ is defined according to users' feedback, where $y_{ui} = 1$ if there is an interaction between u and i, otherwise, $y_{ui} = 0$. In this paper, all ads are displayed by images. Given Y, we aim to predict whether user u would click target ad v. Our goal is to learn a prediction function $\hat{y}_{uv} = f(u, v, \boldsymbol{\Theta})$, where $\hat{y}_{uv}$ denotes CTR probability, and $\boldsymbol{\Theta}$ denotes the model parameters.

Our model contains four parts: image features encoder, text features extractor, attention aggregation, and CTR prediction. Fig.2 shows the main parts of our proposed model. In the following, we will introduce our models in detail.

2.1. Image Features Encoder

We observe that many ads get significantly higher CTR by only switching to more attractive images. This phenomenon indicates the style of advertisement is one of the most important factors for CTR. For a specific ad i, we take the output of ResNet [10] as its style $\boldsymbol{x}^G \in \mathbb{R}^{d_1}$, which is defined as global image feature in this paper. Where d_1 is the embedding dimension. As mentioned before, most pop-up ads are visually divided into three parts. To depict the visual features thoroughly, we divide the ad figure into three parts according to its layout. This step is done by ad designers, as shown in the left part of Fig.2. Equally, we take the three parts as inputs and the outputs of ResNet, $\boldsymbol{x}^L_1$ (the top part), $\boldsymbol{x}^L_2$ (the middle part), $\boldsymbol{x}^L_3$ (the bottom part), as local image features.

2.2. Text Features Encoder

Bank advertisements aim at financial services and products, including texts to describe the products. Under this circumstance, text features are even more significant. For each image part in section 2.1, we ran public text recognizers (e.g., commercial OCR model provided by Google) to get the corresponding texts d^L_1, d^L_2, and d^L_3. Then we use a pre-trained BERT as the context encoder. Take d^L_1 as an example, we take d^L_1 as context input, which is represented as a token sequence in BERT's input format. The output vector $\boldsymbol{d}^L_1 \in \mathbb{R}^{d_2}$ of $[CLS]$ token is treated as local text feature corresponding to the top part, where d_2 is the embedding size. In a similar manner, we get $\boldsymbol{d}^L_2$ and $\boldsymbol{d}^L_3$. $\boldsymbol{d}^L_1$, $\boldsymbol{d}^L_2$ and $\boldsymbol{d}^L_3$ are defined as local text features.

2.3. Attention Aggregation

For each ad, we obtain three local image features ($\boldsymbol{x}^L_1$, $\boldsymbol{x}^L_2$, $\boldsymbol{x}^L_3$) and three local text features ($\boldsymbol{d}^L_1$, $\boldsymbol{d}^L_2$, $\boldsymbol{d}^L_3$). We use $\mathcal{E}^{'} = \{\boldsymbol{x}^L_1, \boldsymbol{x}^L_2, \boldsymbol{x}^L_3, \boldsymbol{d}^L_1, \boldsymbol{d}^L_2, \boldsymbol{d}^L_3\}$ to denote the local feature set. In order to obtain relation-enhanced features, we design an attention aggregation method based on their position information. Specifically, we first project image features and text features into the same space, $\mathcal{E} = \{\boldsymbol{P}_1\boldsymbol{x}^L_1, \boldsymbol{P}_1\boldsymbol{x}^L_2, \boldsymbol{P}_1\boldsymbol{x}^L_3, \boldsymbol{P}_2\boldsymbol{d}^L_1, \boldsymbol{P}_2\boldsymbol{d}^L_2, \boldsymbol{P}_2\boldsymbol{d}^L_3\}$, where $\boldsymbol{P}_1 \in \mathbb{R}^{d \times d_1}$, $\boldsymbol{P}_2 \in \mathbb{R}^{d \times d_2}$ are projection matrix that will be learned. Since the center part contains the most important information, we only update the representations of the center features:

$$\boldsymbol{f}_1 = \boldsymbol{P}_1\boldsymbol{x}^L_2 + \sum_{\boldsymbol{e}_i \in \mathcal{E}, \boldsymbol{e}_i \neq \boldsymbol{P}_1\boldsymbol{x}^L_2} \alpha_i \boldsymbol{e}_i,$$

$$\boldsymbol{f}_2 = \boldsymbol{P}_2\boldsymbol{d}^L_2 + \sum_{\boldsymbol{e}_i \in \mathcal{E}, \boldsymbol{e}_i \neq \boldsymbol{P}_2\boldsymbol{d}^L_2} \beta_i \boldsymbol{e}_i,$$

where

$$\alpha_i = \frac{\exp(f_r(\boldsymbol{P}_1 \boldsymbol{x}_2^L, \boldsymbol{e}_i))}{\sum_{\boldsymbol{e}_j \in \mathcal{E}, e_j \neq \boldsymbol{P}_1 \boldsymbol{x}_2^L} \exp(f_r(\boldsymbol{P}_1 \boldsymbol{x}_2^L, \boldsymbol{e}_i))},$$

$$\beta_i = \frac{\exp(f_r(\boldsymbol{P}_2 \boldsymbol{d}_2^L, \boldsymbol{e}_i))}{\sum_{\boldsymbol{e}_j \in \mathcal{E}, e_j \neq \boldsymbol{P}_2 \boldsymbol{d}_2^L} \exp(f_r(\boldsymbol{P}_2 \boldsymbol{d}_2, \boldsymbol{e}_i))}.$$

After attention aggregation, $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$ can be seen as the relation-enhanced features of image and text.

2.4. CTR Prediction

To get the final representation vector for ad i, we integrate both global image feature and local features.

$$\boldsymbol{x}^i = \boldsymbol{P} \cdot \text{concat}(\boldsymbol{P}_1 \boldsymbol{x}^G, \boldsymbol{f}_1, \boldsymbol{f}_2), \tag{1}$$

where $\boldsymbol{P} \in \mathbb{R}^{d \times 3d}$ is another projection matrix that will be learned during training. Then the final user representation vector is calculate by:

$$\boldsymbol{u} = \text{GRU}\left(\boldsymbol{x}^{i_1^u}, \boldsymbol{x}^{i_2^u}, \ldots, \boldsymbol{x}^{i_{|\mathcal{I}^u|}^u}\right), \tag{2}$$

Then, user representation $\boldsymbol{u}$ (gotten by eq.(2)) and target ad v's representation $\boldsymbol{x}^v$ (gotten by eq.(1)) are combined to predict the CTR probability: $\hat{y}_{uv} = \sigma(f(\boldsymbol{u}, \boldsymbol{x}^v))$, where $\sigma(\cdot)$ is the sigmoid function, and f is a ranking function. The loss function L is:

$$L = \sum_{(u,v) \in \boldsymbol{Y}} -\left(y_{uv} \log \hat{y}_{uv} + (1 - y_{uv}) \log(1 - \hat{y}_{uv})\right).$$

3. EXPERIMENT

In this section, we explore the effectiveness of both our proposed advertisement image modeling module and the whole CTR model PopCTR. The experiments are conducted to answer the following questions:

(i) Does our proposed model PopCTR outperform the state-of-the-art baselines?

(ii) How our proposed model PopCTR is affected by advertisement image modeling module?

3.1. Dataset

The Bank Pop-up Ads Dataset (BPPA, for short) is collected and constructed from one famous bank APP, containing advertisements related to banking products and service content and related users' clicking behavior logs. BPPA will enable us to provide insights on this issue, and train and evaluate our proposed model PopCTR and other CTR models. There are 216,846 users with 1,632,850 iteractions and 5,842 adertisement images in BPPA, which were accumulated from Aug. 1st, 2020 to Sep. 1st , 2020. We discard the click length less than 3 to avoid data sparsity. The average sequence length of BPPA is about 7.53. Assuming there are T behaviors of user u, we use this behavior sequence to predict the user response probability at the target item for the (T+1)-th behavior. Note that 50% target items at the prediction time in each dataset have been replaced with another item from the non-clicked item set for each user, to build the negative samples. The training set is about 70% of the whole dataset, test set is about 20% and the validation set is about 10%.

3.2. Baselines and Metrics

We compare our model PopCTR with the state-of-the-art methods from different types of CTR approaches, including: BPR [1] with no image information and no sequential information, VBPR [5]: with image information from pre-trained and fixed CNN and no sequential information, DVBPR [6] with image information trained with the whole CTR prediction model and no sequential information, GRU [3] with no image information and sequential information.

Given a user, we infer the item that the user would probably buy at a future time. Each candidate method will produce an ordered list of items, we adopt a widely used metrics in sequential recommendation tasks: Recall ratio at rank K(Recall@K, K=1,3,5).

3.3. Parameter Settings

For fair comparison, all algorithms are implemented in the same environment using Tensorflow. To set the parameters in our experiments, we either follow the reported optimal parameter settings or optimize each model separately by Adam using the validation set. We use the same settings as ResNet to capture image features from bank ads and the input size of ads is 600*600*3. The embedding size of global and local features is set as 64. The proposed model PopCTR is trained for 45 epochs and the batch size employed in all our experiments is 128, with a starting learning rate of 5e^{-5} that decays by a factor of 0.2 on the epochs 1000.

3.4. Performance Comparison

In this subsection, we present the performance comparison results and summarize insights. Table 1 shows the results in terms of Recall@K(K=1,3,5). For the BPPA dataset, our proposed PopCTR gives 13.29% improvements on average compared with the best baseline (GRU) and achieves the largest improvement of 16.47% in terms of Recall@5, which demonstrates the benefits of global and local features of bank ads. BPR only utilizing click behavior information perform poorly in all the baselines. Either ads image information or sequential information can improve CTR prediction based on the results of VBPR,DVBPR and GRU.

The difference between our PopCTR and the other baseline models is two-fold. (i) Our model implements ResNet to

Table 1. Performance Comparison on all models.The results of the optimal baseline are expressed in bold.

Methods	Recall@1	Recall@3	Recall@5
BPR-MF	0.162	0.243	0.351
VBPR	0.172	0.263	0.352
DVBPR	0.177	0.289	0.361
GRU	**0.184**	**0.292**	**0.363**
Our model	0.201	0.334	0.422
Our model vs. best	9.24%	14.38%	16.25%

extract the overall figure style from entire ads figure, at the same time , local image and text features are captured from cutted ads figure by ResNet and Bert, respectively. Local image and text features are aggregated by attention mechanism. (ii) Our model utilizes GRU to learn the temporal patterns from user click logs, which can effectively catch the changes of users'interests.

3.5. Ablation Study

As aforementioned in Sec.2, our proposed PopCTR enables us to learn the global and local features of bank pop-up ads. To further analyze the effect of ads' global and local features, we set GRU as base model and design three variations of our proposed PopCTR, base model+global features, base model + local image features and base model +local text features. The results of ablation study is shown in Table 2.

Table 2. Ablation study on Bank-Pop-Up dataset. The percentage in '()' is the relative improvement compared to the base model.

Methods	Recall@1	Recall@3	Recall@5
Base Model(GRU)	0.184	0.292	0.363
Base Model + Global features	0.198 (7.61%)	0.316 (8.22%)	0.405 (11.57%)
Base Model + Local image features	0.194 (5.43%)	0.309 (5.82%)	0.387 (6.61%)
Base Model + Local text features	0.199 (8.15%)	0.323 (10.62%)	0.414 (14.05%)
Our model(Base Model+ Global and Local features)	0.201 (9.24%)	0.334 (14.38%)	0.422 (16.25%)

It can be observed that the complete model can significantly outperform the three weakened variations, indicating that global image, local image and local text features all help to improve the final performance. Base model+local image featues variation achieve lower improvement than base model+global features. Base model only using local image features lacks the information of overall style and magnifies partial style, which leads to more noise than that using global image features. Among the three variations, Base model+local text featues variation gets the largest relative improvement, which attributes to the importance of text in bank ads. The text on bank ads is often the key to attracting customers to click on the advertisement, because bank products like financial product cannot be vividly represented by pictures.

4. CONCLUSION

In this paper, we propose a multi-model architecture for CTR prediction. Specifically, we first extract global and local image features based on ads' layouts. Next, we obtain local text features corresponding to local image feature. Then, we get final ad representations by combining image features with local text features corresponding to their position information. Finally, we capture users' temporal interests from history click sequences. Experimental results demonstrate our model greatly outperforms previous state-of-the-art in real bank pop-up ad datasets. In the feature work, we will design more deliberate method to grasp ads' local image features automatically and deal with multi-model features.

5. REFERENCES

[1] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme, "BPR: bayesian personalized ranking from implicit feedback," in *UAI*, 2009, pp. 452–461.

[2] H.B. McMahan, G. Holt, D. Sculley, M. Young, D. Ebner, and et.al, "Ad click prediction: a view from the trenches," in *SIGKDD*, 2013, pp. 1222–1230.

[3] B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk, "Session-based recommendations with recurrent neural networks," in *ICLR*, 2016.

[4] G. Zhou, N. Mou, Y. Fan, Q. Pi, W. Bian, and et.al, "Deep interest evolution network for click-through rate prediction," in *AAAI*, 2019, pp. 5941–5948.

[5] R. He and J.J. McAuley, "VBPR: visual bayesian personalized ranking from implicit feedback," in *AAAI*, 2016.

[6] W.C. Kang, C. Fang, Z. Wang, and J.J. McAuley, "Visually-aware fashion recommendation and design with generative image models," in *ICDM*, 2017.

[7] X. Yang, T. Deng, W. Tan, X. Tao, J. Zhang, and et.al, "Learning compositional, visual and relational representations for CTR prediction in sponsored search," in *CIKM*, 2019, pp. 2851–2859.

[8] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, and et.al, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," in *SIGIR*, 2020.

[9] H. Liu, J. Lu, H. Yang, X. Zhao, S. Xu, and et.al, "Category-specific CNN for visual-aware CTR prediction at jd.com," in *KDD*, 2020.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.