Joint Optimization of Ad Ranking and Creative Selection

Kaiyi Lin*
Alibaba Group
Beijing, China
linkaiyi.lky@alibaba-inc.com

Xiang Zhang*
Alibaba Group
Beijing, China
genshen.zx@alibaba-inc.com

Feng Li†
Alibaba Group
Beijing, China
adam.lf@alibaba-inc.com

Pengjie Wang‡
Alibaba Group
Beijing, China
pengjie.wpj@alibaba-inc.com

Qingqing Long
Alibaba Group
Beijing, China
lantu.lqq@alibaba-inc.com

Hongbo Deng
Alibaba Group
Beijing, China
dhb167148@alibaba-inc.com

Jian Xu
Alibaba Group
Beijing, China
xiyu.xj@alibaba-inc.com

Bo Zheng‡
Alibaba Group
Beijing, China
bozheng@alibaba-inc.com

ABSTRACT

In e-commerce, ad creatives play an important role in effectively delivering product information to users. The purpose of online creative selection is to learn users' preferences for ad creatives, and to select the most appealing design for users to maximize Click-Through Rate (CTR). However, the existing common practices in the industry usually place the creative selection after the ad ranking stage, and thus the optimal creative fails to reflect the influence on the ad ranking stage. To address these issues, we propose a novel Cascade Architecture of Creative Selection (CACS), which is built before the ranking stage to joint optimization of intra-ad creative selection and inter-ad ranking. To improve the efficiency, we design a classic two-tower structure and allow creative embeddings of the creative selection stage to share with the ranking stage. To boost the effectiveness, on the one hand, we propose a soft label list-wise ranking distillation method to distill the ranking knowledge from the ranking stage to guide CACS learning; and on the other hand, we also design an adaptive dropout network to encourage the model to probabilistically ignore ID features in favor of content features to learn multi-modal representations of the creative. Most of all, the ranking model obtains the optimal creative information of each ad from our CACS, and uses all available features to improve the performance of the ranking model. We have launched our solution in Taobao advertising platform and have obtained significant improvements both in offline and online evaluations.

CCS CONCEPTS

• **Information systems** → **Computational advertising**.

KEYWORDS

Creative Selection, Online Advertising, Adaptive Dropout Network

ACM Reference Format:
Kaiyi Lin, Xiang Zhang, Feng Li, Pengjie Wang, Qingqing Long, Hongbo Deng, Jian Xu, and Bo Zheng. 2022. Joint Optimization of Ad Ranking and Creative Selection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3477495.3531855

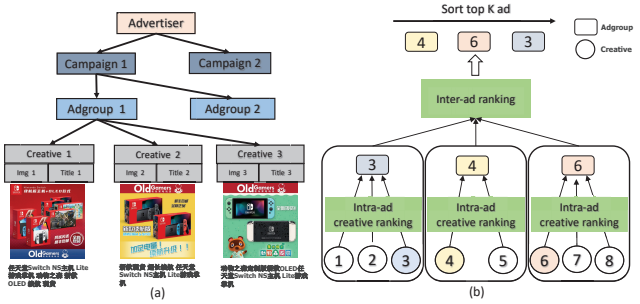


Figure 1: (a) Illustration of an ad campaign setup and (b) our cascade architecture of intra-ad creative selection and inter-ad ranking.

1 INTRODUCTION

As an important advertising carrier, ad creatives(i.e., ad image and title) can quickly deliver rich product information to users in a visual manner. Although the creatives of an ad represent the same product, their CTRs are largely different because different users have different concerns. For instance, for the same ad, some people like the images with promotional information, while others like succinct product images. Thus it is crucial to personalize to display

*Co-first authorship.
†This author is the one who gives a lot of guidance in the work.
‡Corresponding author.

SIGIR '22, July 11–15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-8732-3/22/07…$15.00
https://doi.org/10.1145/3477495.3531855

ad creative based on users' preferences to maximize CTR. In Taobao sponsored search system, as shown in Fig 1(a), the advertiser can create multiple campaigns and each campaign can have multiple ad-groups. An ad-group has multiple creatives (e.g., 10) which are composed of the combination of images and titles materials. The purpose of the creative selection is to select the most attractive ad creative online according to the real-time user request.

Traditional sponsored search systems generally adopt a multi-stage cascade architecture, i.e., ad matching stage, ad ranking stage and creative selection stage, to trade-off between the effectiveness and efficiency. The first stage matching is to reduce the ad candidates from tens of millions to thousands. And the next stage of ranking is to rank ads to generate the final top-k ads. Finally, given the top-k ads from the ranking stage, the creative selection is to select the best visualization for each ad. Existing creative selection methods [5, 16, 19, 21] follow this architecture and mainly focus on improving the performance of the creative selection stage. Though promising results have been reported, the effect is limited because the creative selection stage is at the end of the system. Since the ad ranking is a fine-grained task, its model is larger and more complex. Under the computation constraint, it is unacceptable to simultaneously perform creative selection and ad ranking in the ad ranking stage, in that the number of candidates will explode several times. The current practice in the industry ranking stage usually includes not using creative, or using random creative, or using the top hot creative selected offline, but none of them can obtain the actual display of creative. However, actual creative information is crucial to the performance of the ranking model, so there is still a lot of room for improvement in the CTR prediction [3, 6, 7, 12, 13, 22, 23].

The ideal multi-stage architecture should be ad matching stage, creative selection stage, and ad ranking stage, that is, creative selection should be carried out first, and then ad ranking can obtain the optimal creative, so as to improve the performance of the ranking model and maximize the benefits of the system. However, there are dual challenges in this architecture: 1) *Efficiency*: The number of candidates for creative selection increases by an order of magnitude, and the computing cost increases significantly; 2) *Effectiveness*: A large number of creatives cannot be fully displayed, exacerbating the problem of data sparsity. Therefore, traditional learning strategies (e.g., CTR prediction) based on historical feedback behaviors (e.g., clicks or purchases) becomes a huge challenge.

In this paper, as shown in Fig 1(b), we propose a novel Cascade Architecture of Creative Selection (CACS), which is built before the ranking stage to correlate intra-ad creative ranking and inter-ad ranking. To improve the system efficiency, we make the following efforts: 1) design a classic two-tower structure [9], which can save a significant amount of time because its score can be predicted directly by calculating the simple vector inner product between the two towers; 2) allow the creative embeddings generated by the creative selection model to share with the downstream ad ranking model. To boost the system effectiveness, on the one hand, inspired by the ranking distillation [11, 14, 17], we propose a soft label list-wise ranking distillation method to learn the relative order of creatives and distill the ranking knowledge to guide CACS learning. The soft label of creatives order in each ad is produced by the powerful ranking model to alleviate the dilemma that a large number of creatives not fully displayed cannot obtain real labels. On the other hand, fusing creative ID features and content (i.e., image and title) features is an effective method to alleviate the problem of data sparsity. Therefore, we propose an adaptive dropout network by selecting appropriate dropout [15, 18] ratios based on the impressions to encourage the model to ignore ID features in favor of content features to learn the multi-modal creative representation.

We highlight our contributions in this paper as follows:

- To our best knowledge, we are the first to place the creative selection module before the ranking stage. Simultaneously, intra-ad creative selection and inter-ad ranking are optimized through an efficient and effective cascade structure.
- Considering the efficiency and effectiveness, we make the following efforts: 1) design a classic two-tower structure to alleviate the computational cost and share creative embeddings between creative selection and ad ranking to avoid duplicated computation; 2) propose a soft label list-wise ranking distillation method to distill the powerful ranking knowledge from the ranking stage to guide CACS learning and design an adaptive dropout network to balance the memorability of ID features and the generalization of content features.
- Extensive experimental results demonstrate the effectiveness and superiority of our CACS approach in both offline and online evaluations.

2 METHODOLOGY

2.1 Problem Formulation

We consider a search ads system with a query $q \in Q$ and an ad set $O = \{a_j\}_{j=1}^{n}$. Each ad consists of m creatives $a_j = \{c_i\}_{i=1}^{m}$, where the i-th creative $c_i = (v_i, t_i, id_i, y_i)$. Here v_i, t_i, id_i and y_i denote its image feature, text feature, creative ID features (each creative has a unique ID) and the soft label that is predicted by ranking model, respectively. Our goal is to select optimal creative for each ad.

2.2 Architecture of CACS

The overall framework of our CACS is shown on the right hand side of Fig 2(a). Compared with traditional creative selection methods, our CACS places the creative selection module before the ad ranking stage to correlate intra-ad creative selection and inter-ad ranking. Specifically, we propose a list-wise loss based on a two-tower model to predict the relative order of the creatives in an ad instead of predicting the absolute CTR. For the ad tower, considering a creative contains multiple heterogeneous modalities features, we design an adaptive dropout network to learn the multi-modal creative representation. Significantly, the creative representation was allowed to share with the ranking model. We will elaborately depict the two key components in our CACS as follows.

2.3 List-wise Ranking Distillation

To train the creative selection model, the straightforward way is to predict CTR. However, there are several problems: 1) Creative selection only needs to learn the relative order of ad creatives instead of the CTR score. 2) Large proportion of creatives cannot be fully displayed, making a simple two-tower model difficult to optimize CTR. In this paper, we cast this task as a learning-to-rank problem to order the creatives. Inspired by the ranking distillation,

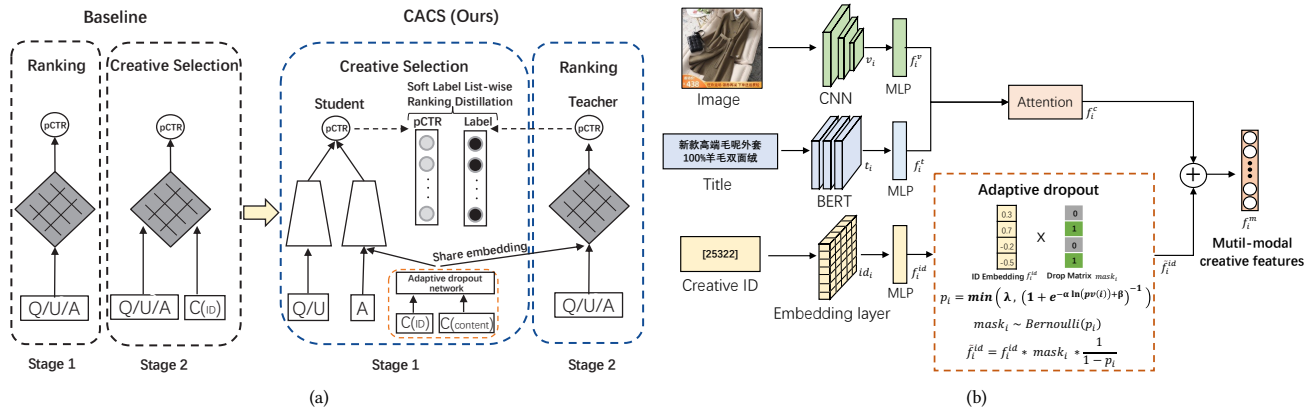


Figure 2: (a) The comparison between the existing method and the proposed CACS, where Q, U, A, C, ID and Content represent query, user, ad, creative, creative ID and creative content features, respectively. (b) Framework of the proposed adaptive dropout network, including two subnetworks: the content network learns the content representations of image and title by attention and the creative ID network uses a mask controlled by p_i to probabilistically set the ID embeddings to 0. And then we element-wise sum the content embeddings and ID embeddings to generate the creative embeddings.

we propose a soft-label list-wise ranking distillation method to distill the ranking knowledge from the teacher ranking model to guide the student CACS learning. Since the ranking model is so powerful that it can predict CTR for creatives fairly accurately, we utilize the ranking model to predict the CTR of creatives in each ad and take the CTR score as the soft label of the creatives' order during the offline training phase.

Similar to other two-tower models, our model is comprised of two major components. For query tower, we use query, user ,and other contexts features to go through an encoder (DNNs) to generate query embeddings. For ad tower, we also use ad and creative(i.e., ad image, title ,and creative ID) features to go through an encoder to generate ad embeddings. Then, we choose cosine similarity to produce a score s_i between query embeddings and ad embeddings. In an ad, there are multiple creative scores $\{s_1, s_2, ..., s_i, ..., s_m\}$ and their corresponding soft labels $\{y_1, y_2, ..., y_i, ..., y_m\}$. Inspired by [2, 19], we first map a list of predicted scores and the soft labels CTRs of creatives to a permutation probability distribution, respectively, and then adopt Cross Entropy as a metric to measure the difference between the two distributions. We consider using the top one probability, which is consistent with our goal, since there is only one creative that will be displayed for each impression. The probability of a creative being ranked top one is defined as:

$$\mathcal{P}_i = \frac{\exp(s_i)}{\sum_{k=1}^{m} \exp(s_k)}, \tag{1}$$

where $\exp(\cdot)$ is an exponential function. And the probability of soft labels are:

$$\hat{\mathcal{P}}_i = \frac{\exp(y_i, T)}{\sum_{k=1}^{m} \exp(y_k, T)}, \tag{2}$$

where T is a temperature coefficient. Since y_m is small, the scale of the values needs to be adjusted by T so that the probability of top1 is close to 1. We calculate the distance between the two distributions with Cross Entropy. The list-wise loss function becomes:

$$\mathcal{L} = -\sum_{i} \hat{\mathcal{P}}_i \log(\mathcal{P}_i). \tag{3}$$

Such objective function is optimized, where the the relative ordering of creatives is captured and the knowledge of the ranking model can be smoothly transferred to our CACS.

2.4 Adaptive Dropout Network

The details of the modeling of the multi-modal creative feature are shown in Fig 2(b). We first consider the content network. For the input features v_i of an image and t_i of a title, we first devise two encoder E_v and E_t to map both image and title features into the common space as $f_i^v = E_v(v_i)$ and $f_i^t = E_t(t_i)$. For different categories of ads, users pay different attention to images and titles. Therefore, we use a simple attention mechanism to learn dynamic weights of images and titles for creatives. Besides, it is important to model user preference jointly with creative ID features and content features. Existing approaches [12, 20] use attention to learn dynamic weights of different modalities. We argue that under the strong influence of ID features, content features will be weakened and it is difficult to learn appropriate weights. In order to reduce the model heavily rely on ID features, we use dropout [15] to regularize it. During the training phase, we set the ID feature to 0 with a probability by controlling the dropout rate, forcing the model to only use content features, so that the content-side model can learn the appropriate parameters. Furthermore, we propose an adaptive dropout strategy based on impressions to control the dropout rate instead of a fixed rate, which can be derived as:

$$p_i = min(\ \lambda\ ,\ (1 + e^{-\alpha \ln(pv(i)+\varphi)+\beta})^{-1}\ ). \tag{4}$$

where λ is adopted to control the max dropout rate and $pv(i)$ denotes the impression. We use $\ln(\cdot)$ to normalize the impression and use small number φ in case $pv(i) = 0$. The right of Eq.4 is a sigmoid function with rescale α and offset β. We found that the greater the amount of impressions, the more the model depends on the ID features in the training process, so we set a smoothing function to control the dropout rate. As the number of impressions increases, the dropout rate also increases. But the rate cannot exceed the threshold λ to ensure ID features are not all discarded. Similarly,

we consider the input ID features and devise an encoder E_{id} to generate ID embeddings as $f_i^{id} = E_{id}(id_i)$. During training, we use a mask matrix $mask_i \sim Bernoulli(p_i)$ to probabilistically set the ID embeddings to 0 controlled by an adaptive dropout strategy as $\tilde{f}_i^{id} = f_i^{id} * mask_i * \frac{1}{1-p_i}$, where $*$ denotes element-wise product and $\frac{1}{1-p_i}$ is adopted to rescale the embeddings to ensure the same distribution in the training phase and the test phase [15]. And then we element-wise sum the $\tilde{f}_i^{id}$ and f_i^c to generate the creative embeddings $f_i^m = \tilde{f}_i^{id} + f_i^c$.

Our adaptive dropout network avoids using complex models to fuse multi-modal features, and is able to improve the performance of creative selection with a much simpler network. Significantly, in each user request, the ranking model can directly use the creative embeddings generated by the adaptive dropout network, and work with other features to predict CTR.

3 EXPERIMENT

3.1 Experimental Setup

Datasets. We collected a period of user click history from Taobao search ad system, and there were about 4 billion samples in training set, 500 million samples in testing set. We evaluate in both offline and online to justify the rationality of our CACS's design.

Metrics. In offline experiments, we use Simulated CTR (sCTR) [19] as an evaluation metric. sCTR is an metric used to simulate creative selection online performance. It replays the impression records of all ads. For one impression, we predict all the creatives under the ad, and select the best creative based on the predicted scores. If the selected creative is the same as this recorded impression creative, we consider this to be an effective impression. Then $impression = impression + 1$, $clicks = clicks + y$, where y represents the actually clicked label, and sCTR defined as $sCTR = \frac{clicks}{impression}$. In online experiments, we use the Click-Through Rate (CTR), Conversion Rate (CVR) and Revenue Per Mile (RPM) that evaluates platform revenue as the metrics for online experiments. Besides, we use Response Time (RT) to observe the timeliness of the system. Note that we show the relative increment.

Details. For query and ad tower, there are 25 and 40 feature slots respectively, and contain three hidden layers with dimensions [512,128,64]. For the creative content feature representations, we use 2,048-dim features extracted by the resnet-50 [8] network to represent each image and use BERT [4] model to extract 512-dim features for representing each title. Adam optimizer with a learning rate of $1.5e^{-6}$ is used for training, with the mini-batch size as 1024.

3.2 Online Evaluation

Effect of the location of the creative selection. To investigate where the creative selection module should be placed in the entire ranking system, we conducted 2 baseline experiments, *no-CR* and *post-CR* represents no creative selection and creative selection after the ranking stage, respectively. From the comparison results in Table 1, we can observe that: 1) The baseline *no-CR* obtains the worst performance, showing that the creative selection can select the most appealing creative for users to improve platform revenue. 2) Our CACS outperforms the counterpart *post-CR* respectively by relative 3.12%, 2.08% and 2.87% in terms of CTR, CVR and RPM.

Table 1: Performance of Online evaluations.

Method	CTR	CVR	RPM	RT (ms)
no-CR	-	-	-	110 (-)
post-CR	+2.21%	+1.04%	+2.29%	114 (+3.71%)
CACS(Ours)	**+5.33%**	**+3.12%**	**+5.16%**	**120 (+9.22%)**

The results show that our CACS correlates the process of intra-ad creative selection and inter-ad ranking, so that the result of creative selection in each ad can affect the ranking ads to improve the performance. 3) Although putting the creative selection module forward will increase the candidate ads, RT does not increase much due to the efficiency of the two-tower model.

Figure 3: Exemplars obtained by CACS and VAM-HBM.

3.3 Offline Evaluation

Comparison with State-of-the-art method. Our CACS obtains the best accuracy of sCTR(%) compared with the VAM-HBM [19] (6.151 vs. 6.032). In addition, we also provide typical exemplars obtained by our CACS in Fig 3. For each user request, we use CACS to rank each creative in the ad candidates. It can be seen that the creative order rank by CACS is better than VAM-HBM, where the number in the upper right corner of the figure denotes the ground truth order of the CTR which is predicted by the ranking model.

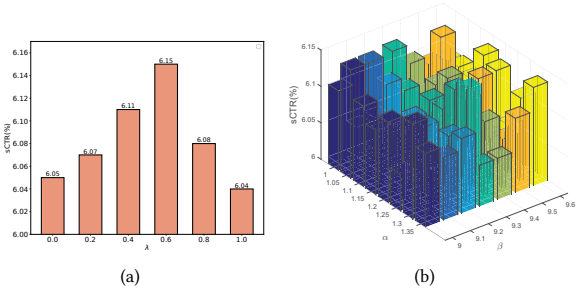
Effect of the multi-modal features. We evaluate the performance of our CACS of the multi-modal and the result are shown in the top panel of Table 2, where the *ID* indicates that only creative ID features are used in the creative selection model, and *Content* indicates that only content-based (i.e., image and title) features are used. We can observe that: 1) When in the low impressions range of 0 to 1000, the performance of utilizing content-based features are obviously better than ID features,it is because content-based features have better generalization, but if the impressions are adequate, ID features have a strong memory ability, and can better understand user interests from historical behaviors. 2) Our CACS approach consistently achieves the best accuracy, showing content features can provide intrinsic visual descriptions, and thus bring better generalization for the model to improve performance.

Effect of the multi-modal learning schemes. To investigate the impact of learning schemes of multi-modal, we design 3 baselines. The comparison results in the bottom panel of Table 2 shows that: 1) The performance obtained by using the attention [1] and gate [10] is better than directly concatenating ID features and content features, because it can adaptively learn the weights of multi-modal

Table 2: The sCTR scores(%) of only ID feature & multi-modal features(top panel) and the multi-modal learning(bottom panel) schemes in different range of impressions.

Baselines	sCTR (%)						
	[0,100]	(100,500]	(500,1000]	(1000,2000]	(2000,5000]	>5000	All
ID	4.737(-)	4.957(-)	5.059(-)	5.238(-)	5.532(-)	6.191(-)	6.045(-)
Content	4.929(+4.05%)	5.120(+3.28%)	5.156(+1.93%)	5.284(+0.88%)	5.531(-0.04%)	6.175(-0.24%)	6.032(-0.20%)
ID + Content (Concat)	4.841(+2.20%)	5.078(+2.43%)	5.196(+2.71%)	5.315(+1.47%)	5.529(-0.06%)	6.183(-0.13%)	6.038(-0.12%)
ID + Content (Attention)	4.850(+2.39%)	5.035(+1.57%)	5.117(+1.63%)	5.324(+0.58%)	5.565(+0.58%)	6.195(+0.08%)	6.073(+0.47%)
ID + Content (Gate)	4.845(+2.28%)	5.122(+3.31%)	5.172(+2.23%)	5.295(+1.08%)	5.570(+0.69%)	6.195(+0.07%)	6.074(+0.49%)
CACS(Ours)	**4.953(+4.56%)**	**5.132(+3.52%)**	**5.188(+2.56%)**	**5.363(+2.40%)**	**5.637(+1.89%)**	**6.292(+1.64%)**	**6.151(+1.75%)**

features. 2) Nevertheless, regardless of high or low impressions, our adaptive dropout network consistently obtains the best sCTR score compared to other multi-modal learning schemes. Under the strong influence of ID features, multi-modal features will be weakened and it is difficult to learn appropriate weight. Therefore, we propose an adaptive dropout strategy based on the impressions to randomly discard the ID features, so that the content-based features can be learned, and finally get the best performance.



(a) (b)

Figure 4: Experiments of our CACS method with (a) different λ and (b) different α, β in Eq.4.

Parameter Sensitivity. Furthermore, we investigate the effect of λ in Eq.4, which is adopted to control the max dropout rate(probability to drop). As shown in Fig 4(a), with the λ gradually increases, the performance gradually improves, and it starts to decline after reaching the peak, and the highest sCTR score is achieved when the rate = 0.6. Besides, for α, β that control the rescale and offset in Eq.4, we set the α in the range of [1,1.35], β in the range of [9,9.6] and increase it by step. From the results shown in Fig 4(b), we can see that performance varies with different value of the parameters, the optimal values for α, β are 1.05 and 9.1.

4 CONCLUSION

In this paper, we propose a novel cascade architecture of creative selection, which is built before the ranking stage to correlate intra-ad creative selection and inter-ad ranking. Besides, we designed several key strategies to address efficiency and effectiveness issues and ultimately improve the performance of both creative selection and ad ranking. We have deployed our CACS in the Taobao ad platform, and have obtained the significant improvements in both offline and online experiments.

REFERENCES

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

[2] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*. 129–136.

[3] 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 *Proceedings of the 1st workshop on deep learning for recommender systems*. 7–10.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Hongliang Fei, Jingyuan Zhang, Xingxuan Zhou, Junhao Zhao, Xinyang Qi, and Ping Li. 2021. GemNN: gating-enhanced multi-task neural networks with feature interaction learning for CTR prediction. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2166–2171.

[6] Tiezheng Ge, Liqin Zhao, Guorui Zhou, Keyu Chen, Shuying Liu, Huimin Yi, Zelin Hu, Bochao Liu, Peng Sun, Haoyu Liu, et al. 2018. Image matters: Visually modeling user behaviors using advanced model server. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 2087–2095.

[7] Huifeng Guo, Ruiming Tang, Yunming Ye, Zhenguo Li, and Xiuqiang He. 2017. DeepFM: a factorization-machine based neural network for CTR prediction. *arXiv preprint arXiv:1703.04247* (2017).

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[9] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. 2333–2338.

[10] Tongwen Huang, Qingyun She, Zhiqiang Wang, and Junlin Zhang. 2020. GateNet: Gating-Enhanced Deep Network for Click-Through Rate Prediction. *arXiv preprint arXiv:2007.03519* (2020).

[11] SeongKu Kang, Junyoung Hwang, Wonbin Kweon, and Hwanjo Yu. 2020. DE-RRD: A knowledge distillation framework for recommender system. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 605–614.

[12] Xiang Li, Chao Wang, Jiwei Tan, Xiaoyi Zeng, Dan Ou, Dan Ou, and Bo Zheng. 2020. Adversarial multimodal representation learning for click-through rate prediction. In *Proceedings of The Web Conference 2020*. 827–836.

[13] H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. 2013. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1222–1230.

[14] Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. 2021. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2368–2376.

[15] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research* 15, 1 (2014), 1929–1958.

[16] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE Transactions on Image Processing* 27, 8 (2018), 3998–4011.

[17] Jiaxi Tang and Ke Wang. 2018. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*.

2289–2298.
[18] Maksims Volkovs, Guang Wei Yu, and Tomi Poutanen. 2017. DropoutNet: Addressing Cold Start in Recommender Systems.. In *NIPS*. 4957–4966.
[19] Shiyao Wang, Qi Liu, Tiezheng Ge, Defu Lian, and Zhiqiang Zhang. 2021. A Hybrid Bandit Model with Visual Priors for Creative Ranking in Display Advertising. In *Proceedings of the Web Conference 2021*. 2324–2334.
[20] Chuhan Wu, Fangzhao Wu, Mingxiao An, Jianqiang Huang, Yongfeng Huang, and Xing Xie. 2019. Neural news recommendation with attentive multi-view learning. *arXiv preprint arXiv:1907.05576* (2019).
[21] Zhichen Zhao, Lei Li, Bowen Zhang, Meng Wang, Yuning Jiang, Li Xu, Fengkun Wang, and Weiying Ma. 2019. What You Look Matters? Offline Evaluation of Advertising Creatives for Cold-start Problem. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*. 2605–2613.
[22] Guorui Zhou, Na Mou, Ying Fan, Qi Pi, Weijie Bian, Chang Zhou, Xiaoqiang Zhu, and Kun Gai. 2019. Deep interest evolution network for click-through rate prediction. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5941–5948.
[23] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. 2018. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1059–1068.