

Towards a Better Tradeoff between Effectiveness and Efficiency in Pre-Ranking: A Learnable Feature Selection based Approach

Xu Ma[†], Pengjie Wang[†], Hui Zhao, Shaoguo Liu, Chuhan Zhao, Wei Lin, Kuang-Chih Lee,
Jian Xu[‡] and Bo Zheng^{*}

Alibaba Group

{maxu.mx,pengjie.wpj,shuqian.zh,shaoguo.lsg,hz_chuhan.zch,kuang-chih.lee,xiyu.xj,bozheng}@alibaba-inc.com,lwsaviola@163.com

ABSTRACT

In real-world search, recommendation, and advertising systems, the multi-stage ranking architecture is commonly adopted. Such architecture usually consists of matching, pre-ranking, ranking, and re-ranking stages. In the pre-ranking stage, vector-product based models with representation-focused architecture are commonly adopted to account for system efficiency. However, it brings a significant loss to the effectiveness of the system. In this paper, a novel pre-ranking approach is proposed which supports complicated models with interaction-focused architecture. It achieves a better tradeoff between effectiveness and efficiency by utilizing the proposed learnable Feature Selection method based on feature Complexity and variational Dropout (FSCD). Evaluations in a real-world e-commerce sponsored search system for a search engine demonstrate that utilizing the proposed pre-ranking, the effectiveness of the system is significantly improved. Moreover, compared to the systems with conventional pre-ranking models, an identical amount of computational resource is consumed.

CCS CONCEPTS

• Information systems → Learning to rank.

KEYWORDS

pre-ranking, effectiveness, efficiency, feature selection

ACM Reference Format:

Xu Ma[†], Pengjie Wang[†], Hui Zhao, Shaoguo Liu, Chuhan Zhao, Wei Lin, Kuang-Chih Lee., Jian Xu[‡] and Bo Zheng^{*}. 2021. Towards a Better Tradeoff between Effectiveness and Efficiency in Pre-Ranking: A Learnable Feature Selection based Approach. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*, July 11–15, 2021, Virtual Event, Canada. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3404835.3462979>

[†] Joint first authors.

[‡] This author gave a lot of guidance in the work.

^{*} Corresponding author.

Project funded by China Postdoctoral Science Foundation (2019TQ0290).

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.

SIGIR '21, July 11–15, 2021, Virtual Event, Canada.

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-8037-9/21/07...\$15.00

<https://doi.org/10.1145/3404835.3462979>

1 INTRODUCTION

As important internet services, large-scale search engine and recommendation systems play important roles in information retrieval and item recommendation, where a ranking system selects only a few items from tens of millions of candidates. Under the constraint of extremely low system latency, a single complicated ranking model cannot rank the entire candidate set. Therefore, multi-stage ranking architecture is commonly adopted [2, 7, 9] (shown in Figure 1 (a)). Large-scale deep neural networks (DNNs) with interaction-focused architecture [3] are usually employed for the ranking model to maintain good system performance, while only less complicated models with representation-focused architecture [3] are adopted in pre-ranking to ensure efficiency.

However, simple pre-ranking models with representation-focused architecture will inevitably diminish the model expression ability. Vector-product based model [4, 13], which is classified as representation-focused architecture, is often employed in pre-ranking (see Figure 1 (b)). The limitation is that neither explicit interactive features nor implicit interactive semantics, which are less efficient for computation but very effective for the expression ability [11], can be used. The pre-ranking models with representation-focused architecture focus excessively on the efficiency optimization and remain a large effectiveness gap to models with interaction-focused architecture. Moreover, as the online feature generation process consume as much resource as that needed for the online inference of the model, the utilization for the pre-ranking with interaction-focused architecture is possible by using feature selection considering both effectiveness and efficiency. In this context, the pre-ranking with an interaction-focused architecture is studied in this paper.

In this paper, a pre-ranking model with tradeoff of both effectiveness and efficiency is proposed for a real-world e-commerce application, where the effectiveness improves significantly with slight decrease of efficiency shown in Figure 1 (b). The contributions of this paper are summarized as follows. 1) A pre-ranking model with interaction-focused architecture is proposed by inheriting the architecture of ranking model to solve the problem of performance loss of the model with representation-focused architecture. 2) A learnable Feature Selection method based on feature Complexity and variational Dropout (FSCD) is proposed to search for a set of effective and efficient feature fields for the pre-ranking model. 3) Extensive experiments are carried out which show the proposed approach achieves great improvement in effectiveness of the system compared to the conventional baselines, while the efficiency of the system is also maintained.

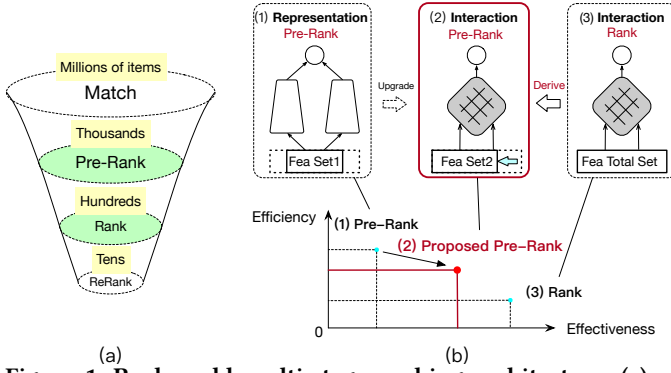


Figure 1: Real-world multi-stage ranking architecture. (a) Multi-stage architecture including matching, pre-ranking, ranking, and re-ranking with scoring item numbers. (b) Intuitive view for effectiveness and efficiency of pre-ranking and ranking, where the pre-ranking is derived from ranking and optimized to an interaction-focused architecture.

2 THE PROPOSED APPROACH

The proposed pre-ranking is derived from the ranking model. Both models utilize interaction-focused architectures, which shares an identical feature set $S = \{f_1, f_2, \dots, f_M\}$, where f_j is the j -th feature field in S , and M is the number of feature fields. In the system, the ranking model employs all feature fields in S to maintain effectiveness, while the pre-ranking model utilizes only a subset to reduce computational complexity and scores more items. In this context, offline processes such as sample generation can be reused for both models, which saves offline computational resources. The interaction-focused architecture of pre-ranking is inherited from that of ranking model, where both explicit interactive features and implicit interactive semantics can be utilized, which reduces the gap in the models' optimization objectives and achieves significant improvement of the model effectiveness for pre-ranking compared to the model with representation-focused architecture.

For introducing interaction-focused architecture into pre-ranking, a great challenge is the model efficiency. The overall efficiency of the pre-ranking model, whose learnable variables can be divided into feature embeddings \mathbf{v} and dense weights \mathbf{w} , is strongly influenced by the feature embeddings \mathbf{v} . For example, in real-world scenarios, storage for feature embeddings exceeds over 95% of that for the whole model, while feature generation process for embeddings consumes as much resources as that needed for the online inference of the model. Therefore, the feature selection for features with both high effectiveness and efficiency is important for the utilization of pre-ranking with interaction-focused architecture.

2.1 FSCD for Pre-Ranking Model

Inspired by the Dropout FR method based on the variational dropout layer [1] where the efficiency of the model is ignored, FSCD is proposed that considers both effectiveness and efficiency in a learnable process as illustrated in Figure 2. In the proposed FSCD, the effectiveness is optimized by the cross entropy based loss function, while the efficiency is optimized by the feature-wise regularization term in Eq. (3). Both effective and efficient features can be selected by FSCD in one single training process, while the expression ability of pre-ranking model is improved utilizing these features compared to the vector-product based model. The details are as follows.

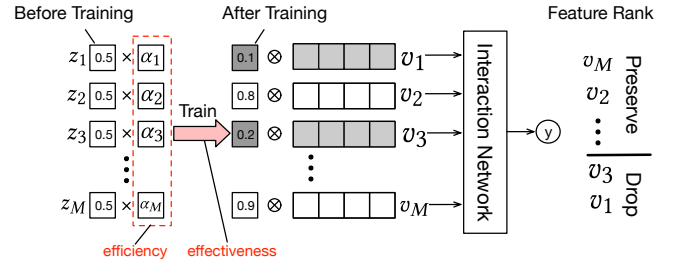


Figure 2: The proposed FSCD for pre-ranking model.

To select feature fields with both high effectiveness and high efficiency, each feature field f_j is expected to learn a dropout parameter $z_j \in \{0, 1\}$ to indicate whether the feature field is dropped ($z_j = 0$) or preserved ($z_j = 1$). f_j 's embeddings v_j are multiplied by the dropout z_j to form the embedding layer, and z_j is subject to a Bernoulli distribution parameterized by θ_j , i.e.,

$$z_j \sim \text{Bern}(\theta_j), \quad (1)$$

where the hyperparameter θ_j is the priori probability for the preservation of feature field f_j and is configured as function of feature complexity c_j , i.e.,

$$\theta_j = \mathcal{H}(c_j) = 1 - \sigma(c_j), c_j = \mathcal{G}(o_j, e_j, n_j) \quad (2)$$

where $\sigma(\cdot)$ is the sigmoid function. $\theta_j = 1 - \sigma(c_j)$ is one of alternatives to relate θ_j and c_j which works well in practice. The feature complexity c_j measures the computational and storage complexity of the j -th feature field including but not limited to the online computational complexity o_j , the embedding dimension e_j , and the number of keys n_j for one feature field, where o_j is configured according to the feature type specified in Section 3.1. Linear combination $c_j = \gamma_1 o_j + \gamma_2 e_j + \gamma_3 n_j$ is one of the alternatives with hyperparameter $\gamma_{1,2,3}$. According to Eq. (2), a feature with large complexity has a small value of θ_j , and vice versa.

Given the training samples $\mathcal{D} = \{(x_i, y_i) | i = 1, 2, \dots, N\}$, where N is the number of samples, the overall loss function for the learnable feature selection is derived by Bayesian rule [10] as,

$$L(\mathbf{v}, \mathbf{w}, \mathbf{z}) = -\frac{1}{N} \log P(\mathcal{D} | \mathbf{v}, \mathbf{w}, \mathbf{z}) + \frac{\lambda}{N} (\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2) + \sum_{j=1}^M \frac{\alpha_j z_j}{N}, \quad (3)$$

where α_j is the regularization weight for z_j and can be derived as (See derivation details in Appendix A)

$$\alpha_j = \log(1 - \theta_j) - \log(\theta_j). \quad (4)$$

α_j is a function that decreases with θ_j and increases with c_j . Therefore, a feature with larger complexity c_j is penalized with a larger value of α_j , and more likely to be dropped. In this way, the feature complexity is included in the proposed FSCD, which previous works do not address to the best of the authors' knowledge.

Subject to Bernoulli distribution, z_j is discrete and not differentiable, it is relaxed to a differentiable function as [5]

$$z_j = \mathcal{F}(\delta_j) = \sigma\left(\frac{1}{t}(\log(\delta_j) - \log(1 - \delta_j) + \log(u_j) - \log(1 - u_j))\right), \quad (5)$$

where $u_j \sim \text{Uniform}(0, 1)$ is subject to a uniform distribution and changes during the training process, while $t = 0.1$ is a constant that works well in the experiments. $\mathcal{F}(\delta_j)$ is close to 0 or 1 for most δ_j

values, which approximates the discrete Bernoulli distribution. In contrast to the dropout z_j , $\delta_j \in (0, 1)$ is a differentiable parameter. Moreover, it acts as the posterior probability for feature preservation, which is influenced by the priori probability for feature preservation θ_j , and can be learned as the feature importance.

In this context, the entire process of training for feature selection is built in Eq. (3) and (5). Note that, the learnable variables are \mathbf{v} , \mathbf{w} and δ , which are trained simultaneously. Through a fast convergence of δ , the feature set of the pre-ranking model can then be obtained by selecting the feature fields with top- K of δ values.

2.2 Fine-Tuning the Pre-Ranking Model

After feature selection, the feature fields not in the pre-ranking set acquired by FSCD are masked, and the models are fine-tuned using weights \mathbf{v} and \mathbf{w} as initialization parameters. Concretely, the model can be trained with the following loss function

$$L(\mathbf{v}', \mathbf{w}') = - \sum_{i=1}^N \log p(y_i | f(x'_i, \mathbf{v}', \mathbf{w}')), \quad (6)$$

where x'_i , \mathbf{v}' , and \mathbf{w}' are the samples with the selected feature fields, the remaining feature embeddings, and dense weights for the pre-ranking model, respectively. \mathbf{v}' and \mathbf{w}' are initialized by \mathbf{v} and \mathbf{w} , which accelerates the training. The Bernoulli dropout \mathbf{z} is omitted in the fine-tuning process.

In this way, the pre-ranking model is obtained. As the model is in an interaction-focused architecture and adopts both effective and efficient feature fields, its expression ability significantly improves. Therefore, the effectiveness of the system can be improved with high efficiency. The entire method is illustrated in Algorithm 1.

Algorithm 1 The Proposed Pre-ranking based on FSCD

Input: Feature set $S = \{f_1, f_2, \dots, f_M\}$, feature complexity metrics $\{o_j, e_j, n_j\}$, training samples \mathcal{D} , hyperparameters $\gamma_{1,2,3}$.

Output: Pre-ranking model \mathcal{M} .

- 1: Calculate c_j , θ_j and α_j for each feature field using Eq. (2) and (4).
- 2: Build the model with interaction-focused architecture where the feature embeddings are multiplied by z_j parameterized by δ_j in Eq. (5).
- 3: Train using Eq. (3), and then select the feature fields with top- K of δ_j .
- 4: Fine-tune according to Eq. (6) and obtain \mathcal{M} .

3 EXPERIMENTS

3.1 Experiment Configurations

Evaluations are mainly based on the online sponsored search system for a real-world e-commerce app named Taobao, where both the click-through rate (CTR) and response per mile (RPM), which evaluates platform revenue, are optimized. The proposed effective and efficient pre-ranking are compared with the baselines of the vector-product based model [4] and COLD [12] based pre-ranking.

Both the pre-ranking and ranking models are based on an interaction-focused architecture. After the feature fields are transformed into embeddings, hidden layers with sizes of 1024, 512, and 256 are adopted for the ranking, while only 2 hidden layers with sizes of 1024 and 256 are used for the pre-ranking to ensure high efficiency. Finally, the sigmoid function is utilized to predict the final CTRs for both models.

The feature set for the pre-ranking model is a subset of that for the ranking model, which is selected by the proposed FSCD. All feature sets, consisting of 246 feature fields in their entirety,

feature type index	feature type	o_j
I	Simple query/user features that directly look up embeddings	0.4
II	Simple item correlated features that require complicated computations	1.5
III	Complex query/user features that directly look up embeddings	1.0
IV	Complex item correlated features that require complicated computations	3.0

Table 1: Different types of features for the entire feature set with the values of o_j .

feature type index	number of feature fields	minimum ranking	median ranking	maximum ranking
I	141	1/2/1	129/137/85	244/246/156
II	5	15/31/18	36/44/42	99/52/159
III	68	3/1/4	137/131/185	246/243/231
IV	32	7/3/8	108/58/229	178/239/246

Table 2: Feature ranking and distributions for different types of features, where the ranking results for different methods are provided and divided by slashes in order, including the Dropout FR method [1], the DFS method [6], and the proposed FSCD.

include different types, e.g., simple features that directly look up embeddings and complex features that require complicated computations for online embeddings. Each type can be divided into either user/query features or item correlated features which include item features and interactive features between item and user/query. The user/query features are computed only once for online inference regardless of the number of candidate advertisements and require less computational consumption, while the item correlated features should be computed as many times as the number of advertisements to be scored, thus consuming much more computational resources. The o_j values are determined by the above feature types and the detailed configurations are listed in Table 1. For other hyperparameters, $\gamma_1 = 1$, $\gamma_2 = 10^{-2}$, and $\gamma_3 = 10^{-7}$ are configured.

3.2 Analysis of FSCD

After the proposed FSCD process, the feature fields are ranked using the output parameter δ . Table 2 lists the minimum, median, and maximum ranking indices for different types of features using Dropout FR [1] and deep feature selection (DFS) method [6]. For the proposed FSCD in this paper, the ranking indices for the simple features are much smaller than those for the complex feature types (see median and maximum). The simple features tend to have a more forward ranking, while the complicated features are backward, which emphasizes the efficiency of the features. However, the minimum ranking index of type IV features is only 8, which means that a feature with large complexity can still rank forward if it is truly effective for the training task. For the conventional methods, the rankings of the features are less dependent on the feature type.

The pre-ranking models with different feature field number K are well trained based on Eq. (6) with 2 billion samples. The area under the curve (AUC), latency, and CPU consumption for online inference of the different models are illustrated in Table 3. The results of the conventional feature selection methods [1, 6] are also provided for comparison. Specifically, when $K = 246$, the feature set of pre-ranking model becomes exactly the same as that of the ranking model. The results show that the proposed FSCD has slightly smaller AUC than the other methods when identical K values are considered. However, the complexity is extremely

K	method	AUC	latency(ms)	CPU(%)
30	Dropout FR	0.6910	4.32	13.26
	DFS	0.6765	4.64	15.47
	FSCD	0.6903(-0.007)	2.41(-44.2%)	9.31(-29.7%)
50	Dropout FR	0.6946	5.75	20.55
	DFS	0.6788	5.19	20.57
	FSCD	0.6927(-0.0019)	3.54(-31.7%)	12.10(-41.1%)
100	Dropout FR	0.6975	6.48	30.82
	DFS	0.6962	6.34	26.11
	FSCD	0.6949(-0.0026)	4.37(-31.0%)	19.18(-26.5%)
150	Dropout FR	0.6985	7.98	36.35
	DFS	0.6972	7.85	35.41
	FSCD	0.6958(-0.0027)	4.74(-39.6%)	21.47(-39.3%)
200	Dropout FR	0.6983	8.10	37.26
	DFS	0.6981	8.04	36.94
	FSCD	0.6971(-0.0010)	6.91(-14.0%)	30.09(-18.5%)
whole (246)	/	0.6990	9.02	40.60

Table 3: AUC results, latency, and CPU consumption for the first K feature fields of the proposed FSCD and conventional methods. Each model is trained with 2 billion samples.

Methods	recall rate	offline AUC
Vector-product based model	88%	0.695
COLD model	96%	0.738
The proposed model	95%	0.737

Table 4: Offline results of the proposed pre-ranking model and the conventional models. Each model is trained with more than 200 billion samples.

Methods	input item number	CTR	RPM	response time	CPU consumption
vector-product based model	6000	/	/	58.4 ms	79%
COLD model	600	+1.11%	+1.04%	62.3 ms	85%
The proposed model	800	+1.54%	+2.76%	59.9 ms	79%

Table 5: Online effectiveness and efficiency of the proposed pre-ranking model and the conventional models on real-world mobile online queries.

reduced. When $K = 100$, the AUC difference relative to that of the model for the other methods is only 0.0026, while the complexity is approximately 30% lower than that of the other methods for both latency and CPU cost. When $K > 100$, the AUC increases slowly, while the complexity increases significantly. Therefore, the pre-ranking utilizes the feature fields that have top-100 of δ values as the final feature set.

3.3 Performance Comparisons

Table 4 and Table 5 show the offline and online experimental results using the proposed effectiveness and efficiency based pre-ranking, while the original vector-product and COLD based pre-rankings are configured as benchmarks. Each model is trained with more than 200 billion real-world e-commerce samples from a log system to achieve the best online performance. Therefore, the offline AUCs are much larger than those shown in Table 3. The recall rate is defined as the preserved probability of pre-ranking model for the top-5 items ranked by the ranking model. The offline results in Table 4 show that the proposed model is much more effective than the vector-product based model, while it is slightly less effective than COLD due to efficiency considerations. For the online effect, 30 continuous days of mobile real-world requests are evaluated and the results are shown in Table 5. It shows that the proposed pre-ranking model obtains a good balance for online CTR and RPM performance and gains significant improvement in both CTR and RPM compared with its counterparts.

Finally, the efficiency is analyzed. Although a complicated pre-ranking with an interaction-focused architecture is adopted in the proposed approach, the feature selection method based on effectiveness and efficiency optimizes the computational complexity to a low degree. Moreover, the number of input item is reduced from 6000 to 800, which further reduces the overall complexity of the system. In this context, the online CPU consumption of the system is almost the same as that of the vector-product based model, while the response time is slightly increased. In regard to the COLD model, the response time and CPU consumption are both greater than those of the proposed approach, which means the proposed pre-ranking model is more efficient. The key metrics for efficiency are listed in Table 5 at a peak number of queries per second (QPS).

4 CONCLUSIONS

In this paper, to solve the problem of performance loss caused by the pre-ranking model with representation-focused architecture, a pre-ranking model based on feature selection with joint optimization for both effectiveness and efficiency is proposed in an interaction-focused architecture. The pre-ranking model with interaction-focused architecture, which is inherited from that of ranking model, utilizes the feature subset selected by the proposed learnable FSCD, which includes not only simple features but also interactive features with significant effectiveness. The experiments on offline training demonstrate the validity of the proposed FSCD method in both effectiveness and efficiency. Moreover, the offline and online effects in the real-world sponsored search system illustrate the performance improvements in the proposed pre-ranking model compared to the conventional benchmarks. The proposed pre-ranking has been utilized as an online running model for a real-world sponsored search system and has generated substantial revenue for the company.

A DERIVATION FOR EQ. (3) AND (4)

Assuming that \mathbf{v} and \mathbf{w} are subject to a joint Gaussian distribution [8], the joint distribution for \mathbf{v} , \mathbf{w} and \mathbf{z} is

$$P(\mathbf{v}, \mathbf{w}, \mathbf{z}) = \mathcal{N}(\mathbf{v}, \mathbf{w} | \mathbf{0}, \Sigma) \prod_{j=1}^M \text{Bern}(z_j | \theta_j), \quad (7)$$

where \mathbf{v} , \mathbf{w} and \mathbf{z} are all learnable variables.

To optimize these variables, the loss function can be concluded by maximizing the posterior probability of \mathbf{v} , \mathbf{w} and \mathbf{z} given the training samples \mathcal{D} . In this context, by applying Bayesian rule [10], maximizing $P(\mathbf{v}, \mathbf{w}, \mathbf{z} | \mathcal{D}) \propto P(\mathcal{D} | \mathbf{v}, \mathbf{w}, \mathbf{z}) P(\mathbf{v}, \mathbf{w}, \mathbf{z})$ is equivalent to:

$$\mathbf{v}^*, \mathbf{w}^*, \mathbf{z}^* = \arg \min_{\mathbf{v}, \mathbf{w}, \mathbf{z}} -\log P(\mathcal{D} | \mathbf{v}, \mathbf{w}, \mathbf{z}) - \log P(\mathbf{v}, \mathbf{w}, \mathbf{z}). \quad (8)$$

The first part is the cross entropy, while the second part $-\log P(\mathbf{v}, \mathbf{w}, \mathbf{z})$ can be rewritten according to Eq. (7) as

$$-\log P(\mathbf{v}, \mathbf{w}, \mathbf{z}) = \lambda(\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2) + \sum_{j=1}^M \alpha_j z_j + C, \quad (9)$$

where λ , α_j and C are all constants derived by Eq. (7). $\lambda(\|\mathbf{w}\|^2 + \|\mathbf{v}\|^2)$ is the common l_2 regularization term with weight λ . The nontrivial term $\sum_{j=1}^M \alpha_j z_j$ is a new regularization term derived by the Bernoulli distribution, where the regularization factor α_j is derived by Eq. (1) and (7) as

$$\alpha_j = \log(1 - \theta_j) - \log(\theta_j). \quad (10)$$

Then the total loss function can be written as Eq. (3).

REFERENCES

- [1] Chun-Hao Chang, Ladislav Rampasek, and Anna Goldenberg. Dropout feature ranking for deep learning models. *arXiv preprint arXiv:1712.08645*, 2017.
- [2] Luke Gallagher, Ruey-Cheng Chen, Roi Blanco, and J Shane Culpepper. Joint optimization of cascade ranking models. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 15–23, 2019.
- [3] Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W Bruce Croft, and Xueqi Cheng. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067, 2020.
- [4] Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338, 2013.
- [5] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.
- [6] Yifeng Li, Chih-Yu Chen, and Wyeth W Wasserman. Deep feature selection: theory and application to identify enhancers and promoters. *Journal of Computational Biology*, 23(5):322–336, 2016.
- [7] Shichen Liu, Fei Xiao, Wenwu Ou, and Luo Si. Cascade ranking for operational e-commerce search. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1557–1565, 2017.
- [8] Xun Liu, Wei Xue, Lei Xiao, and Bo Zhang. Pbodl: Parallel bayesian online deep learning for click-through rate prediction in tencent advertising system. *arXiv preprint arXiv:1707.00802*, 2017.
- [9] Vikas C Raykar, Balaji Krishnapuram, and Shipeng Yu. Designing efficient cascaded classifiers: tradeoff between accuracy and cost. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 853–860, 2010.
- [10] Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. Bpr: Bayesian personalized ranking from implicit feedback. *arXiv preprint arXiv:1205.2618*, 2012.
- [11] Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee. Billion-scale commodity embedding for e-commerce recommendation in alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 839–848. ACM, 2018.
- [12] Zhe Wang, Liqin Zhao, Biye Jiang, Guorui Zhou, Xiaoqiang Zhu, and Kun Gai. Cold: Towards the next generation of pre-ranking system. *arXiv preprint arXiv:2007.16122*, 2020.
- [13] Wenjin Wu, Guojun Liu, Hui Ye, Chenshuang Zhang, Tianshu Wu, Daorui Xiao, Wei Lin, and Xiaoyu Zhu. Eenmf: An end-to-end neural matching framework for e-commerce sponsored search. *arXiv preprint arXiv:1812.01190*, 2018.