Curriculum Modeling the Dependence among Targets with Multi-task Learning for Financial Marketing

Yunpeng Weng*
edwinweng@tencent.com
FiT, Tencent
Shenzhen, Guangdong, China

Xing Tang*
shawntang@tencent.com
FiT, Tencent
Shenzhen, Guangdong, China

Liang Chen†
leoncuhk@gmail.com
FiT, Tencent
Shenzhen, Guangdong, China

Xiuqiang He†
xiuqianghe@tencent.com
FiT, Tencent
Shenzhen, Guangdong, China

ABSTRACT

Multi-task learning for various real-world applications usually involves tasks with logical sequential dependence. For example, in online marketing, the cascade behavior pattern of *impression → click → conversion* is usually modeled as multiple tasks in a multi-task manner, where the sequential dependence between tasks is simply connected with an explicitly defined function or implicitly transferred information in current works. These methods alleviate the data sparsity problem for long-path sequential tasks as the positive feedback becomes sparser along with the task sequence. However, the error accumulation and negative transfer will be a severe problem for downstream tasks. Especially, at the beginning stage of training, the optimization for parameters of former tasks is not converged yet, and thus the information transferred to downstream tasks is negative. In this paper, we propose a prior information merged model (**PIMM**), which explicitly models the logical dependence among tasks with a novel prior information merged (**PIM**) module for multiple sequential dependence task learning in a curriculum manner. Specifically, the PIM randomly selects the true label information or the prior task prediction with a soft sampling strategy to transfer to the downstream task during the training. Following an easy-to-difficult curriculum paradigm, we dynamically adjust the sampling probability to ensure that the downstream task will get the effective information along with the training. The offline experimental results on both public and product datasets verify that PIMM outperforms state-of-the-art baselines. Moreover, we deploy the PIMM in a large-scale FinTech platform, and the online experiments also demonstrate the effectiveness of PIMM.

*Contributed equally
†Corresponding authors

CCS CONCEPTS

• **Computing methodologies → Multi-task learning**; • **Information systems → Computational advertising**.

KEYWORDS

Multi-task learning, sequential dependence, online marketing

1 INTRODUCTION

Multi-task learning (MTL) has been widely introduced in many real-world applications such as product recommendation and online marketing [1, 6, 18, 20, 21]. An MTL model aims to learn multiple related tasks simultaneously and leverages the shared representation to improve performance for all the tasks[3]. In these applications, some tasks are logically dependent on each other, which indicates that downstream task requires information from prior tasks. As a result, we should exploit the knowledge about such a dependency relationship and help the MTL model more effectively share the learned information among tasks. In all, modeling the relationships among tasks properly is one of the critical issues in MTL [19].

For many real-world commercial applications, customer conversion follows a sequential behavior pattern. For instance, when distributing mutual funds to customers online, we usually conduct some marketing campaigns to promote revenue. As Figure 1 illustrates, the user behavior in a campaign follows the path *impression → click → conversion → core conversion*, which indicates that only if the positive feedback in the former task occurs, the positive feedback can be expected in the subsequent task, i.e. the target task *core conversion* is progressive. The data sparsity problem thereby becomes severe along with the behavior path. A common way to handle this problem is to formulate it as a multi-task learning regime. On the one hand, notice that core conversion might occur only after the click and purchase behavior occurred. Hence, the regime requires taking the logical dependencies between tasks into account. On the other hand, the downstream tasks will be
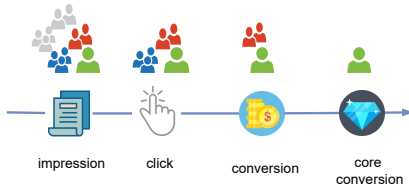
Figure 1: The customer conversion path for a financial marketing campaign. Core conversion (purchase over 1,000 CNY) is the primary target due to the commercial consideration.

affected by the former task at the early stage of training, which leads to a phenomenon called *error accumulation*. Specifically, the downstream tasks will make predictions based on the predicted label of former tasks. However, when the optimization is not converged, the performance of the target task will suffer greatly from the accumulated loss of former tasks.

Previous efforts [9, 12, 13] have been made to share information among tasks, which can alleviate the data sparse problem. Nevertheless, they ignore modeling the sequential logical dependence among tasks[18]. In recent years, the research on MTL with task dependency has received increasing attention. A collection of models capture the task dependency by simply considering the explicit estimated probability of the prior task with a defined function. Ma et al. [10] proposed the Entire Space Multi-task Model (ESMM), which explicitly transfers the predicted probability of the prior task by multiplying it with an estimated conditional probability. Furthermore, wen et al. [15] consider more auxiliary tasks in their proposed ESM^2 for probability transferring. However, the simple function cannot encode enough information among tasks. To enhance the shared information, several methods focus on exploiting the implicit prior task representation. Adaptive Information Transfer Multi-task (AITM) framework [18] transfers implicit information from the former task's hidden layer in the corresponding tower with the self-attention mechanism. Inspired by AITM, Wu [16] proposed a Multi-level Network Cascades Model (MNCM). MNCM replaces the shared-bottom structure of AITM with the experts-bottom structure and uses Expert-Level Information Transfer Module (EITM) to transfer implicit representation among task-specific experts. However, the prior task information is not fully utilized since they ignore the explicit logical dependency information of the previous task. Besides, the aforementioned methods also lead to error accumulation and negative transfer for downstream tasks. Especially, at the beginning stage of training, the optimization for parameters of former tasks is not converged yet, and thus the task-specific information transferred to downstream tasks could be misleading, which could exacerbate the difficulty of target tasks learning as they are usually at the end of the path.

To tackle these challenges, we proposed a prior information merged model (**PIMM**) in this paper, which explicitly models the logical dependence among tasks with a prior information merged (**PIM**) module for multiple sequential dependence task learning in a curriculum manner. PIMM leverages not only the implicit representation but also the explicit premise information. The implicit representation contains rich task-specific knowledge learned in the tower network of the former task. The explicit premise information indicates the probability that the prior positive feedback occurs.

Besides, we introduce a curriculum modeling manner to facilitate information transfer between multiple tasks. Specifically, the PIM randomly selects the true label information or the prior task prediction as explicit premise information. Notice that we introduce a soft sampling strategy following an easy-to-difficult curriculum paradigm to overcome the error accumulation phenomenon. To demonstrate the superiority of our proposed PIMM, we conduct extensive experiments on a publicly available dataset collected from traffic logs of Taobao's recommender system [1] and on a real-world industrial dataset collected from the customer acquisition campaigns in a large-scale FinTech platform. Furthermore, we also conduct online experiments to verify the effectiveness of PIMM in a real-world financial customer acquisition scenario.

2 THE PROPOSED APPROACH

2.1 Problem Formulation

In this paper, we focus on the problem of modeling a collection of binary classification tasks $T = \{T_0, T_1, ..., T_M\}$ with sequential dependence. The corresponding labels of each training sample are $Y = \{y_1, y_2, ..., y_M\}$ where M is the number of tasks. Only if $y_i = 1$, the latter task's label y_{i+1} might be 1, which means only the following three situations are valid for any two adjacent tasks: (1) $y_i = 0 \& y_{i+1} = 0$, (2) $y_i = 1 \& y_{i+1} = 0$, (3) $y_i = 1 \& y_{i+1} = 1$.
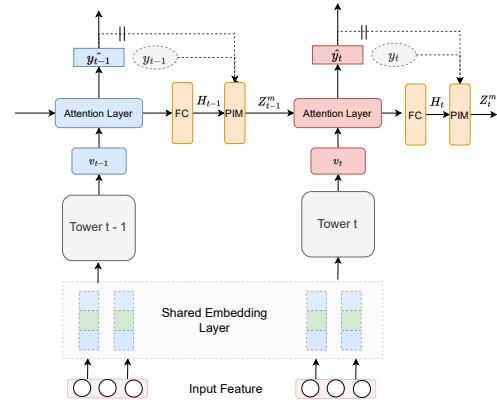
2.2 The Structure of PIMM

Figure 2: The architecture of PIMM. FC represents a fully-connected layer. "∥" represents the gradient that is truncated.

We firstly present the overview structure in Figure 2, each feature field x_i will be embedded as a low-dimension vector in the shared embedding layer. Subsequently, all output field vectors are concatenated and fed to all task-specific towers. Note that the backbone of PIMM could be easily extended to more complex structures such as expert-bottom structures. Besides, the advanced models for representation learning can also be utilized as the bottom or tower networks such as graph neural networks[11, 17], feature interaction models[4, 7, 8], and so on. In this paper, we focus on the problem of representing and utilizing information about logical dependence among tasks. For all tasks except the first one, we use
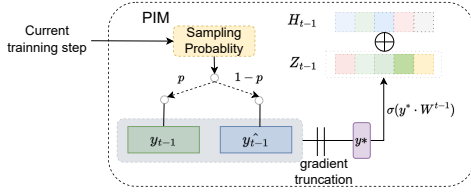
---

[1]https://tianchi.aliyun.com/dataset/408

Figure 3: Prior Information Merged (PIM) Module.

the prior information merged (PIM) module to obtain the transfer information. The PIM first generates an embedding containing explicit premise information from ground truth or the estimated prediction of the former task with a mutable probability. Then PIM merges the explicit information with the transferred hidden vector that contains implicit information with an addition operation. The detail of PIM will be described in Section 2.3. Given the output of t-th task-specific tower network $v_t \in \mathcal{R}^d$ and the output of PIM Z_{t-1}^m which is the prior information merged representation contains the knowledge about the former task, we employ a self-attention layer with the residual connection to generate the representation U_t for the prediction of task t:

$$U_t = v_t + \sum_{\mathbf{a} \in \{v_t, Z_{t-1}^m\}} w_a \cdot V(\mathbf{a}) \tag{1}$$

where V is the linear transformation function that projects v_t to a new d-dimension vector. w_a is the attention weight, which is calculated as:

$$w_a^* = \frac{Q(\mathbf{a}) \cdot K(\mathbf{a})}{\sqrt{d}} \tag{2}$$

$$w_a = softmax(w_a^*) = \frac{\exp(w_a^*)}{\sum_{\mathbf{a}} \exp(w_a^*)} \tag{3}$$

where Q and K are linear transformation that project $\mathbf{a} \in \{v_t, Z_{t-1}^m\}$ to d-dimension vectors as queries and keys respectively.

Besides being fed to the output layer of task t, U_t will be utilized to generate the implicit representation H_t with a fully-connected layer and be transferred to the subsequent PIM module. Specially, there is no attention layer for the first task since it has no former task. Therefore, for the first task, v_0 will be directly fed to the output layer and H_0 is $FC(v_0)$.

2.3 Prior Information Merged Module

To leverage the explicit premise information from the former task, a naive idea is directly transferring the predicted result of $(t-1)$-th task. However, for some applications such as multi-step conversion prediction in the FinTech platform, the path of sequential dependence targets is long and the error of estimated probability will be transferred accumulatively. Especially at the early stage of training, the optimization for parameters of previous tasks is not converged yet and the predictions are not accurate enough. Inspired by the concept of Curriculum learning[14], we implement an easy-to-difficult training strategy in the Prior Information Merged (PIM) Module as shown in Figure 3. The PIM first calculates the sampling probability p according to the current training step. For each training sample in the batch, PIM randomly selects the label y_{t-1} or the predicted probability $\hat{y_{t-1}}$ of the former task with the sampling probability

of p and $1 - p$ respectively. The sampling probability is mutable during the training process. At the beginning of training, p is a large value and gradually decreases as the training process advances until p reaches the set lower limit. The sampling probability p can be formulated as:

$$p = \max\{\alpha - E \times S, \beta\} \tag{4}$$

where α is the initial value of p. $E = 0, 1, 2, ..., K$ is the current training epoch. S is the decrease speed and $\beta \geq 0$ is the set lower limit. At inference, only the predicted probability $\hat{y_{t-1}}$ can be utilized. In this way, the PIM tends to select the ground truth when the model is not well-trained to avoid transferring misleading prior information in the beginning. As the training advances and the predictions of the model becomes accurate, PIM is more likely to choose the predicted results gradually and thus eliminates the gap between training and inference. The selected explicit premise information $y^* \in \{y_{t-1}, \hat{y_{t-1}}\}$ is then transformed into a d-dimension vector Z_{t-1} with the following operation:

$$Z_{t-1} = \sigma(y^* \cdot W^{t-1}) \tag{5}$$

where $W^{t-1} \in \mathbb{R}^{1 \times d}$ is a trainable variable. To avoid affecting the upstream tasks learning, we apply the gradient truncation operation on the selected y^*. Finally, we obtain the prior information merged task-specific knowledge Z_{t-1}^m by adding the transferred implicit representation H_{t-1} from the former task and Z_{t-1}:

$$Z_{t-1}^m = Z_{t-1} + H_{t-1} \tag{6}$$

The curriculum sequential task modeling strategy makes PIMM focus on task-specific parameters learning with the accurate prior information in the beginning and gradually increases the generality as the sampling probability p decreases.

3 EXPERIMENTS

3.1 Datasets and Evaluation Metrics

In the offline experiments, we compare our proposed PIMM model with several baselines including some MTL models on a publicly available dataset and an industrial private dataset:

The Ali-CCP dataset: Alibaba Click and Conversion Prediction dataset[10] consists of over 42 million samples in the training set and over 43 million samples in the testing set with sequential labels of click and purchase. Following the previous research [18], we use the single-values categorical features and randomly retain 10% samples from the training set as a validation set.

The industrial dataset: We collect logs from historical campaigns from a FinTech platform. The training set contains about 3 million records sampled from 2022/09/01 to 2023/01/04 and the testing set contains 5.8 million records collected after the training period. The dataset contains three sequential dependence tasks: click $\rightarrow$ conversion $\rightarrow$ core conversion, where the definition of core conversion is to invest over 1 thousand CNY in mutual funds distributed in the platform.

For all models, we report the average AUC and standard deviation (std) over five runs with different random seeds to reduce randomness.

3.2 Settings

The baseline models are listed as follows:

Table 1: The AUC performance (mean ± std) of different models on the Ali-CCP dataset.

Model	click	purchase
shared-bottom	0.6068 ± 0.0008	0.6388 ± 0.0068
ESMM	0.6060 ± 0.0012	0.6417 ± 0.0064
MMoE	0.6072 ± 0.0006	0.6463 ± 0.0062
PLE	0.6059 ± 0.0018	0.6468 ± 0.0086
AITM	0.6072 ± 0.0013	0.6508 ± 0.0028
MNCM	0.6071 ± 0.0011	0.6511 ± 0.0032
PIMM	**0.6075 ± 0.0010**	**0.6561 ± 0.0030**‡

‡ indicates statistically significant improvement over the best baseline(p-value < 0.05).

Shared-Bottom[2]: The Shared-Bottom is a simple MTL structure. All tasks share the same bottom layers and task-specific towers are employed for each task individually.

MMoE[9]: The MMoE uses multiple expert networks as the bottom structure and applies the gate mechanism to integrate the expert outputs for different tasks.

ESMM[10]: The ESMM transfers the output probability from the previous task p_{t-1} and multiplies it with the estimated conditional probability p_t to model the sequential dependence.

PLE[13]: The PLE introduces task-specific experts besides the shared experts to help mitigate the seesaw phenomenon in MTL.

AITM[18]: The AITM model adaptively transfers implicit representation from the former task with the attention mechanism.

MNCM[16]: The MNCM improves the AITM model by utilizing the task-specific and shared experts pattern.

Following the experimental setting of AITM [18], the embedding size is set to be 5 and 24 for the experiments on the Ali-CCP dataset and industrial dataset respectively. All models are trained with Adam [5] optimizer. We sweep over the learning rate in {0.0005, 0.001, 0.0015,0.002} and select the best-performance parameter for all models. For the models with shared-bottom structures, the hidden layer sizes of each task-specific tower are [128,64,32] for the Ali-CCP dataset and [256,128,64] for the industrial dataset. For MNCM and PLE, the number of shared experts is 2 and each task has two task-specific experts. For the MMoE model, the share experts number are 4 and 6 for the Ali-CCP dataset and the industrial dataset respectively. We set $\alpha = 0.5$, $S = 0.25$, $\beta = 0.25$ for PIMM on the Ali-CCP dataset and $\alpha = \frac{2}{3}$, $S = \frac{1}{3}$, $\beta = 0$ on the industrial dataset respectively.

3.3 Results

Offline Results. Generally, the final task in the sequential task paths is typically the most important for practical applications such as recommendations and marketing. Therefore, we mainly focus on the purchase task and core conversion task for the Ali-CCP dataset and industrial dataset respectively. The experimental results on the public dataset are shown in Table 1. These results illustrate that our proposed PIMM significantly outperforms the various state-of-the-art baseline models in terms of the purchase AUC. In Table 2, we present the experimental results on the industrial dataset with three tasks. The PIMM model also achieves significant improvement over

Table 2: The AUC performance (mean ± std) of different models on the industrial dataset.

Model	click	conversion	core conversion
shared-bottom	0.7397 ± 0.0051	0.7106 ± 0.0082	0.6861 ± 0.0092
ESMM	0.7348 ± 0.0008	0.7224 ± 0.0013	0.6893 ± 0.0033
MMoE	0.7395 ± 0.0010	0.7171 ± 0.0006	0.6907 ± 0.0013
PLE	0.7470 ± 0.0012	**0.7309 ± 0.0008**	0.7045 ± 0.0036
AITM	0.7372 ± 0.0006	0.7104 ± 0.0011	0.7052 ± 0.0033
MNCM	0.7420 ± 0.0011	0.7237 ± 0.0006	0.7042± 0.0023
PIMM	**0.7489 ± 0.0004**‡	0.7296 ± 0.0013	**0.7118 ± 0.0035**‡

‡ indicates statistically significant improvement over the best baseline(p-value < 0.05).

the baselines in terms of core conversion AUC while maintaining competitive performance on other tasks. The results demonstrate the effectiveness of our proposed PIMM in MTL with sequential dependence tasks.

Online Results. We conduct the online experiment on several real financial customer acquisition campaigns in different periods. For each campaign, we randomly take 50% of the traffic as the control group and 50% of the traffic as the experimental group and ensure two groups are homogeneous. Each model predicts the probability of users' core conversion, and users with the highest potential will be impressed with campaign ads. For a fair comparison, the number of impressions is equal for each group, which means the marketing resources consumed are the same. Figure 4 illustrates the online testing results of three different marketing campaigns. Note that each campaign lasts for 5 days. We find that PIMM constantly increases the core conversion rate compared with the baseline in all online campaigns. The online results further verify the superiority of the PIMM in modeling the sequential dependence among tasks.
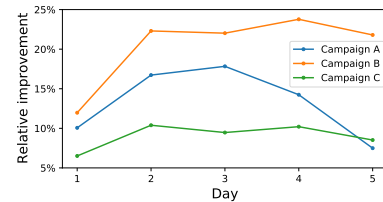
Figure 4: The relative improvement of our PIMM compared to baseline in terms of the core conversion customer number.

4 CONCLUSION

In this paper, we propose a novel Prior Information Merged model for multi-task learning with sequential dependence relationships among tasks. The Prior Information Merged module flexibly generates prior information from the ground truth or predictions of the former task. The dynamic sampling mechanism alleviates the error accumulation phenomenon and also guarantees the generalization ability of PIMM. With the PIM module, PIMM can both leverage the implicit knowledge learned in the task-specific tower of the previous task and the explicit prior knowledge about the dependent task. We conduct extensive offline and online experiments to verify the superiority of PIMM compared with various state-of-the-art multi-task learning models. Besides, the PIM module can be flexibly applied in other advanced model structures.

REFERENCES

[1] Ting Bai, Yudong Xiao, Bin Wu, Guojun Yang, Hongyong Yu, and Jian-Yun Nie. 2022. A Contrastive Sharing Model for Multi-Task Recommendation. In *Proceedings of the ACM Web Conference 2022*. 3239–3247.

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

[3] Michael Crawshaw. 2020. Multi-task learning with deep neural networks: A survey. *arXiv preprint arXiv:2009.09796* (2020).

[4] 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).

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

[6] Hui Li, Yanlin Wang, Ziyu Lyu, and Jieming Shi. 2020. Multi-task learning for recommendation over heterogeneous information network. *IEEE Transactions on Knowledge and Data Engineering* (2020).

[7] 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 *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1754–1763.

[8] Fuyuan Lyu, Xing Tang, Huifeng Guo, Ruiming Tang, Xiuqiang He, Rui Zhang, and Xue Liu. 2022. Memorize, Factorize, or be Naive: Learning Optimal Feature Interaction Methods for CTR Prediction. In *2022 IEEE 38th International Conference on Data Engineering (ICDE)*. 1450–1462. https://doi.org/10.1109/ICDE53745.2022.00113

[9] Jiaqi Ma, Zhe Zhao, Xinyang Yi, Jilin Chen, Lichan Hong, and Ed H Chi. 2018. Modeling task relationships in multi-task learning with multi-gate mixture-of-experts. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*. 1930–1939.

[10] Xiao Ma, Liqin Zhao, Guan Huang, Zhi Wang, Zelin Hu, Xiaoqiang Zhu, and Kun Gai. 2018. Entire space multi-task model: An effective approach for estimating post-click conversion rate. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1137–1140.

[11] Yi Ouyang, Bin Guo, Xing Tang, Xiuqiang He, Jian Xiong, and Zhiwen Yu. 2021. Mobile App Cross-Domain Recommendation with Multi-Graph Neural Network. *ACM Trans. Knowl. Discov. Data* 15, 4, Article 55 (apr 2021), 21 pages. https://doi.org/10.1145/3442201

[12] Zhen Qin, Yicheng Cheng, Zhe Zhao, Zhe Chen, Donald Metzler, and Jingzheng Qin. 2020. Multitask mixture of sequential experts for user activity streams. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3083–3091.

[13] Hongyan Tang, Junning Liu, Ming Zhao, and Xudong Gong. 2020. Progressive layered extraction (ple): A novel multi-task learning (mtl) model for personalized recommendations. In *Fourteenth ACM Conference on Recommender Systems*. 269–278.

[14] Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 9 (2021), 4555–4576.

[15] Hong Wen, Jing Zhang, Yuan Wang, Fuyu Lv, Wentian Bao, Quan Lin, and Keping Yang. 2020. Entire space multi-task modeling via post-click behavior decomposition for conversion rate prediction. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 2377–2386.

[16] Haotian Wu. 2022. MNCM: Multi-level Network Cascades Model for Multi-Task Learning. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4565–4569.

[17] Shiwen Wu, Fei Sun, Wentao Zhang, Xu Xie, and Bin Cui. 2022. Graph neural networks in recommender systems: a survey. *Comput. Surveys* 55, 5 (2022), 1–37.

[18] Dongbo Xi, Zhen Chen, Peng Yan, Yinger Zhang, Yongchun Zhu, Fuzhen Zhuang, and Yu Chen. 2021. Modeling the sequential dependence among audience multi-step conversions with multi-task learning in targeted display advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3745–3755.

[19] Yu Zhang and Qiang Yang. 2021. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering* (2021).

[20] Zhe Zhao, Lichan Hong, Li Wei, Jilin Chen, Aniruddh Nath, Shawn Andrews, Aditee Kumthekar, Maheswaran Sathiamoorthy, Xinyang Yi, and Ed Chi. 2019. Recommending what video to watch next: a multitask ranking system. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 43–51.

[21] Yongchun Zhu, Yudan Liu, Ruobing Xie, Fuzhen Zhuang, Xiaobo Hao, Kaikai Ge, Xu Zhang, Leyu Lin, and Juan Cao. 2021. Learning to Expand Audience via Meta Hybrid Experts and Critics for Recommendation and Advertising. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 4005–4013.